MeLos: Analysis and Modelling of Speech Prosody and Speaking Style

Nicolas Obin

MeLos:
Analysis and Modelling of Speech Prosody and Speaking Style

Nicolas Obin

This thesis has been defended on 2011, June 23^{th} to obtain the grade of Docteur de l'Université Paris VI - Pierre et Marie Curie (UPMC) with major in signal processing from the Ecole Doctorale Informatique, Télécommunications et Electronique (EDITE)

Reviewers	Nick Campbell	Professor	CLCS, University of Dublin
	Simon King	Professor	CSTR, University of Edinburgh
Examiners	Jean-François Bonastre	Professor	LIA, University of Avignon
	Eric de la Clergerie	Researcher	INRIA
	David Wessel	Researcher	CNMAT, UC Berkeley
	Jean-Luc Zarader	Professor	UPMC
	Anne Lacheret	Professor	MoDyCo, University of Paris X
	Xavier Rodet	Researcher	IRCAM

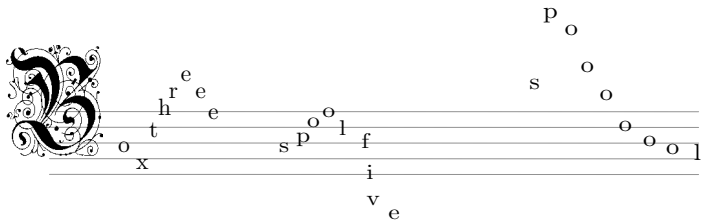
This work has been supervised by Xavier Rodet and Anne Lacheret, conducted at the Institut de Recherche et Coordination Acoustique/Musique (Ircam) , and Modyco Lab.

Ircam - CNRS-UMR9912-STMS
Sound Analysis/Synthesis Team
1, place Igor Stravinsky
75004 Paris, FRANCE

Monday 2nd January, 2012

Submitted version

For any comment, suggestion, or correction please contact the author.



music of speech...
Samuel Beckett, Krapp's Last Tape.

Je ne vis que de-ci de-là à l'intérieur d'un mot dans l'inflexion duquel je perds pour un instant ma tête inutile.

Franz Kafka, Journal.

Regarde(z)-moi ça, en voilà du propre !

Jean Genet, Les Bonnes.

Abstract

This thesis addresses the issue of modelling speech prosody for speech synthesis, and presents *MeLos*: a complete system for the analysis and modelling of speech prosody - "*the music of speech*".

Research into the analysis and modelling of speech prosody has increased dramatically in recent decades, and speech prosody has emerged as a crucial concern for speech synthesis. The issue of speech prosody modelling is to model speech prosody variations depending on the context - linguistic (e.g. linguistic structure), para-linguistic (e.g., emotion), or extra-linguistic (e.g., socio-geographical origins, situation of a communication). Modelling the variability of speech prosody is required to provide natural, expressive, and varied speech in many applications of high-quality speech synthesis such as multi-media (avatars, video games, story telling, dialogue systems) and artistic (cinema, theatre, music, digital arts) applications. The objective of the present study on the analysis and the modelling of speech prosody is to vary and adapt the strategy, alternatives, and speaking style of a speaker for natural, expressive, and varied speech synthesis.

The objective of this thesis is to model strategies, alternatives, and speaking style of a speaker for natural, expressive, and varied speech synthesis. The present study presents original contributions that correspond to a special attention paid to the combination of theoretical linguistics and statistical modelling to provide a complete speech prosody system that can be used for speech synthesis. In particular, speech prosody characteristics are described in three linguistic levels from signal variations to abstract representations. A unified discrete/continuous context-dependent HMM is presented to model the symbolic and the acoustic characteristics of speech prosody. A rich description of the text characteristics based on a linguistic processing chain that includes surface and deep syntactic parsing is proposed to refine the modelling of the speech prosody in context. Segmental HMMs and Dempster-Shafer fusion are used to balance linguistic and metric constrains in the production of a pause. A context-dependent HMM is proposed to model the f_0 variations based on the stylization and the trajectory modelling of short and long-term variations simultaneously over various temporal domains. The proposed system is used to model strategies and alternatives of a speaker, and is extended to the modelling of speaking style shared among speakers using shared-context-dependent modelling and speaker normalization techniques.

Keywords: speech prosody, speaking style, speech synthesis, discrete/continuous HMMs, stylization, trajectory modelling, linguistic analysis.

Résumé

Cette thèse s'intéresse au problème de la modélisation de la prosodie dans le cadre de la synthèse de la parole, et présente *Me$\mathcal{L}$os* : un système complet d'analyse et de modélisation de la prosodie - "*la musique de la parole*".

Les études sur l'analyse et la modélisation de la prosodie ont littéralement explosé au cours de la dernière décennie, et la prosodie s'est graduellement imposée comme un enjeu majeur en synthèse de la parole. La modélisation de la prosodie vise à modéliser les variations de la prosodie en fonction du contexte - linguistique (structure linguistique), para-linguistique (émotion), ou extra-linguistique (origines socio-régionales, situation de la communication). En particulier, la modélisation de la variété prosodique est nécessaire à la synthèse d'une parole naturelle, expressive, et variée dans de nombreuses applications de la synthèse de parole de haute qualité, aussi bien dans les domaines multi-média (avatars, jeux vidéo, livre audio, systèmes de dialogue) qu'artistique (cinéma, théâtre, musique, arts numériques). L'objectif de cette thèse est de modéliser les stratégies, les alternatives, et le style de parole d'un locuteur pour permettre une synthèse naturelle, expressive, et variée.

La présente thèse propose de nouvelles directions pour répondre aux multiples enjeux de la modélisation de la prosodie : une description explicite des différents niveaux et domaines de variation de la prosodie (modèle de signal), l'intégration d'une description riche des caractéristiques du texte (description du contexte), ainsi qu'une modélisation statistique des caractéristiques de la prosodie (modèle statistique). En particulier, la présente étude a été le lieu d'une synergie poussée de la théorie linguistique et des méthodes de modélisation statistique dans l'élaboration d'un système complet de modélisation de la prosodie. Un modèle unifié fondé sur des modèles de Markov cachés (HMMs) à observation discrète/continue est présenté afin de modéliser les caractéristiques symbolique et acoustique de la prosodie. Une description riche des caractéristiques du texte fondée sur une chaîne de traitement linguistique de surface et profonde est introduite pour enrichir la modélisation des variations prosodiques en contexte. Une méthode pour combiner les contraintes linguistiques et métrique dans la production des pauses est proposée, basée sur un modèle segmental et la fusion de Dempster-Shafer. Un modèle de trajectoire basé sur la stylisation des contours prosodiques sur différents domaines temporels est présenté pour modéliser simultanément les variations à court et long terme de la f_0. Le système proposé est utilisé pour modéliser les stratégies et le style d'un locuteur, et est étendu à la modélisation du style de parole par des méthodes de modélisation en contexte partagé et de normalisation du locuteur.

Mots-clefs: prosodie, style de parole, synthèse de la parole, modèle de Markov caché (HMM) à observation discrète/continue, stylisation, modèle de trajectoire, analyse linguistique.

Notations

General

x	=	scalar
$\mathbf{x}$	=	scalar / vector sequence
$\mathbf{X}$	=	matrix
$x^\top$	=	transpose
$\widehat{x}$	=	estimate of variable x
$\bar{x}$	=	mean of variable x
$x^{(i)}$	=	iteration on variable x
$x^{(d)}$	=	dimension of variable x with total dimension D
x_k	=	index of variable x with total dimension K
$[\,.\,]$	=	sequence
$\{\,.\,\}$	=	indexed set
$(\,.\,)$	=	set

Probabilities

$\mathbf{p}(x)$	=	probability of x
$\mathbf{p}(x\|y)$	=	conditional probability of x given y
$\mathbf{p}(x,y)$	=	joint probability of x and y

HMM

$\boldsymbol{\lambda}$	=	model
$(\mathbf{\Pi},\mathbf{A},\mathbf{B})$	=	parameters of a HMM
$(\mathbf{\Pi},\mathbf{A},\mathbf{B},\mathbf{D})$	=	parameters of a segmental HMM
$\mathbf{\Pi}=\{\pi_i\}_{i=1}^N$	=	a-priori probabilities
$\mathbf{A}=\{a_{i,j}\}_{i,j=1}^N$	=	transition probabilities
$\mathbf{B}=\{b_i\}_{i=1}^N$	=	observation probabilities
$\mathbf{D}=\{d_i\}_{i=1}^N$	=	state-duration probabilities
$\mathbf{o}$	=	observation sequence
$\mathbf{q}$	=	state sequence
$\mathbf{s}$	=	segment sequence
$\mathbf{d}$	=	state-duration sequence
α,β	=	forward/backward probabilities
γ	=	state-occupancy probabilities
$\mathcal{N}$	=	normal distribution
α,$\boldsymbol{\mu}$,$\mathbf{\Sigma}$	=	weight, mean, and covariance of a normal distribution

Context-dependent HMM

T	=	context-dependent tree
S_m	=	node of a context-dependent tree
$\boldsymbol{\lambda}_{S_m}$	=	context-dependent HMM associated with a node of the context-dependent tree

Trajectory model

$\mathbf{c}$	=	static observation sequence
$\Delta\mathbf{c}$	=	dynamic observation sequence
$\mathbf{o}$	=	static/dynamic augmented observation sequence
$\mathbf{W}$	=	static/dynamic transformation matrix

Speaker-Independent Model

$\mathbf{W}$	=	transformation matrix
$\boldsymbol{\xi}$	=	augmented mean vector

Speech Prosody

$\boldsymbol{\lambda}^{(symbolic)}$	=	discrete model of symbolic characteristics
$\boldsymbol{\lambda}^{(acoustic)}$	=	continuous model of acoustic characteristics
$\mathbf{o}$	=	acoustic sequence of speech prosody characteristics
$\mathbf{l}$	=	symbolic sequence of speech prosody characteristics
$\mathbf{q}$	=	sequence of linguistic contexts
D	=	dimension of acoustic vector
L	=	dimension of linguistic vector
K	=	number of long-term trajectories
M	=	number of context-dependent HMMs
R	=	number of speakers
$\mathbf{Q}$	=	set of linguistic contexts
t	=	continuous time (frame) of total length T
n	=	discrete time (linguistic unit, e.g. phoneme, syllable) of total length N

Contents

Chapter 1

Introduction

Contents

"*Par son pouvoir expressif, par sa pérennité vis-à-vis de l'univers instrumental, par son pouvoir d'amalgame avec un texte, par la capacité qu'elle a de reproduire des sons inclassables par rapport aux grammaires - la grammaire du langage comme la grammaire musicale - , la voix peut se soumettre à la hiérarchie, s'y intégrer ou s'en dégager totalement. Moyen immédiat, qui n'est pas soumis inéluctablement à la contrainte culturelle pour communiquer, pour exprimer, la voix peut être, autant qu'un instrument cultivé, un outil "sauvage", irréductible*".

Pierre Boulez, Automatisme et décision, Points de repère III, Leçons de musique.

1.1 IRCAM, Music, and ... Speech (!?)

The IRCAM Institute (Institute for Research and Coordination into Acoustics/Music), created in 1997 on the initiative of the French composer Pierre Boulez and the French Ministry of Culture, is the world's largest public center for research into music creation. The fundamental principle of IRCAM is to encourage productive interaction among scientific research, technological developments, and contemporary music production. The research pole covers all the cross-disciplinary fields that relate to sound and music, from perception, acoustics, analysis and synthesis, representation, real-time applications, to the analysis of musical practices.

Following the original intuition of the French composer Pierre Boulez and the impulsion of the emeritus researcher and former head of the analysis-synthesis team Xavier Rodet, voice and speech have been historically considered as a primary research domain in interaction with major concerns about music and artistic creation [Rodet, 1977]. The research on voice originally focused on the analysis and synthesis of the singing voice, leading to the development of the singing voice synthesizer CHANT [Rodet et al., 1984, Bennett and Rodet, 1989] and culminating in the reconstruction of the singing voice of the castrato *Farinelli* [Depalle et al., 1994]. Subsequently, research on speech gradually progressed [Peeters, 2001, Peeters, 2002, Schwarz, 2003] in response to the increasing demand of composers and artists with the development of high-quality speech technologies (ircamAlign, ircamTTS, ircamHTS, and SuperVPTrax) and numerous implications for artistic creation [Fineberg, 2006, Rohmer, 2007, Gervasoni, 2008, Parra, 2009, Lanza and Pasquet, 2009]. The study of speech prosody - *the music of speech* - recently arose from a simultaneous need in speech technologies and in artistic applications, as the cross-disciplinary dimension that unifies speech and music.

1.2 General Background

1.2.1 Speech Synthesis

1.2.1.1 Origins

Speech synthesis is the artificial production of speech[1]. Speech synthesis originated with the "*speaking machine*" in 1791 [von Kempelen, 1791] which consisted of a mechanical system that reproduced the physical production of speech. Many refinements of the original speech synthesizer were developed over the nineteenth century, culminating in the presentation in 1846 of "*The Euphonia, or Speaking Automaton*" [The Euphonia, 1846]. The early-age of modern speech synthesis took place in the first half of the twentieth century with the advent of electrical articulatory and formant speech synthesis systems. [Dudley et al., 1939, Fant, 1953]. Speech synthesis required manual intervention to control the production and the variation of speech parameters. Modern speech synthesis corresponds to the automatic processing of information (informatics) and the development of Text-To-Speech (TTS) synthesis systems. The principle of a Text-To-Speech synthesis system is to automatically synthesize the acoustic parameters of a speech utterance that corresponds to a given text. Text-to-Speech originated with diphone synthesis, and modern Text-to-Speech synthesis systems can be classified into *formant*, *articulatory*, *unit-selection* and *parametric* synthesis systems. In particular, Text-to-Speech Synthesis explodes over the last decades with the emergence of corpus-based systems that provide *intelligible* and *natural* speech [Acapela Group, 2010, AT&T Labs Natural Voices, 2010, Cepstral, 2010, HTS, 2010, Festival, 2010, Loquendo, 2010, Nuance, 2010, Orange Labs, 2010, SoftVoice, 2010, SVOX, 2010]. Modern speech synthesis systems cover a wide range of applications in telecommunications, multimedia, medicine, and artistic domains - from the design of artificial avatars, interactive dialogue systems, speech recovering of pathological individuals, to the ability to reconstruct the speech of deceased personalities, and the capacity to create artificial languages.

[1] see [Klatt, 1987] for an exhaustive historical review of speech synthesis.

1.2.1.2 Unit Selection

The principle of unit-selection speech synthesis is to select and concatenate speech units from a large single-speaker speech database [Hunt and Black, 1996]. Unit-selection speech synthesis historically derives from diphone synthesis [Hamon et al., 1989], subsequently generalized to larger and non-uniform speech units. A speech database of a single speaker is first segmented into linguistically-motivated speech units (e.g., phoneme, syllable). During the training, speech units are clustered into acoustically similar units depending on their linguistic characteristics. During the synthesis, the text is first converted into a sequence of linguistic contexts. Then, the sequence of speech units to be concatenated is selected so as to minimize the concatenation cost of the units sequence given the sequence of linguistic contexts. The concatenation cost divides into *state occupancy cost* and *state transition cost*. The state occupancy cost denotes the acoustic distance of a speech unit candidate to a target unit. The transition cost denotes the acoustic distance for the concatenation of consecutive speech units. During the concatenation, overlap-add methods are used to locally interpolate the selected units so as to minimize the acoustic distance at the juncture of consecutive speech units.

Unit-selection currently remains the most popular method in speech synthesis. Unit-selection speech synthesis benefits from the naturalness and the variety of real speech units that compose the speech database. However, unit-selection speech synthesis is limited to the linguistic and acoustic content of the speech database. In particular, conventional unit-selection speech synthesis is generally limited to a single speaker, a single speaking style, and a single emotional content. Additionally, unit-selection speech synthesis requires a large speech database and optimal recording conditions (e.g., no variation in the acoustic quality, no reverberation, no background noise).

The implementation of the IRCAMTTS speech synthesis system for French is partially based on the FESTVOX Toolkit [Festival, 2010] and the IRCAMCORPUSTOOLS system [Beller et al., 2009][2].

1.2.1.3 HMM-based

The principle of parametric speech synthesis is to model the statistical characteristics of speech based on parametric statistical methods. Parametric speech synthesis historically derives from statistical methods used in speech recognition based on Hidden Markov Models (HMMs) [Ljolje and Fallside, 1986, Farges and Clements, 1988, Giustiniani and Pierucci, 1991, Fukada et al., 1994, Donovan and Woodland, 1995, Tokuda et al., 1995]. The first full parametric speech synthesis was the HMM-based speech synthesis system (HTS) [Yoshimura et al., 1999, Tokuda et al., 2000] which provided a unified statistical framework used for the analysis and synthesis of speech.

The HMM-based speech synthesis system simultaneously models the spectrum, f_0, and duration with a context-dependent HMM [Yoshimura et al., 1999]. Various refinements were proposed to the original context-dependent HMM model to improve the modelling of speech variations in context (Maximum-Likelihood Minimum-Description-Length Context-Clustering (ML-MDL) [Yoshimura et al., 1999], and the f_0 (Multi-Space-Distributions HMM (MSD-HMM) [Tokuda et al., 1999]), the temporal structure (Hidden Semi Markov Model (HSMM) [Zen et al., 2004]), and the dynamic (Trajectory Model [Tokuda et al., 2003], Global Variance (GV) [Toda and Tokuda, 2007], Minimum Generation Error (MGE) [Qian et al., 2009], and rich linguistic context [Yan et al., 2009]) of speech.

During the training, both the spectrum and excitation parameters are extracted from a speech database [Imai, 1983, Kawahara et al., 1999a] and used to estimate context-dependent HMM models. Due to the large amount of linguistic contexts, context-dependent mod-

[2]the author thanks Xavier Rodet, Diemo Schwarz, Grégory Beller, Thomas Huebert, Christophe Veaux, and the Analysis/Synthesis of Speech Team for their implication in the development of the IRCAMTTS speech synthesis system.

els are clustered into acoustically similar models using decision-tree-based context-clustering (ML-MDL [Yoshimura et al., 1999]). Multi-space probability distributions (MSD) are used to model continuous/discrete parameter f_0 sequence to manage voiced/unvoiced regions properly [Tokuda et al., 1999]. Each context-dependent HMM is modelled with state duration probability density functions (PDFs) to account for the temporal structure of speech [Zen et al., 2004]. Finally, speech dynamics is modelled according to the trajectory model and the global variance (GV) that model local and global speech variations over time [Tokuda et al., 2003, Toda and Tokuda, 2007]. During the synthesis, the text to be synthesized is first converted into a sequence of linguistic contexts, and then an utterance HMM is constructed by concatenating the most appropriate context-dependent models according to the sequence of linguistic contexts and the context-dependent model. State durations of the utterance HMM are then determined based on the state duration PDFs. Then, the speech parameters are synthesized so as to maximize the likelihood of the state sequence and the spectral and excitation parameters sequence conditionally to the context-dependent model. Finally, a speech waveform is synthesized using a speech synthesis filter [Kawahara et al., 1999b].

Parametric speech synthesis is generally considered as more robust, expressive and flexible, while unit-selection speech synthesis remains more natural in adequate optimal recording conditions. Firstly, HMM-based speech synthesis system is not restricted to the content of a particular speech database, and several speech databases can be used to model various speech characteristics separately, which then can be combined during speech synthesis. In particular, the HMM-based speech synthesis can be used to model the speech characteristics that are associated either with a single or an arbitrary set of speakers [Shichiri et al., 2002, Yamagishi, 2006], and to modify these characteristics to interpolate and to adapt to those of a target speaker, speaking style, or emotion [Yoshimura et al., 1997, Tachibana et al., 2005, Yamagishi et al., 2004].

The implementation of the IrcamHTS speech synthesis system for French[3] is based on the HTS Toolkit[HTS, 2010].

[3]the author thanks Pierre Lanchantin for the implementation of the IrcamHTS speech synthesis system.

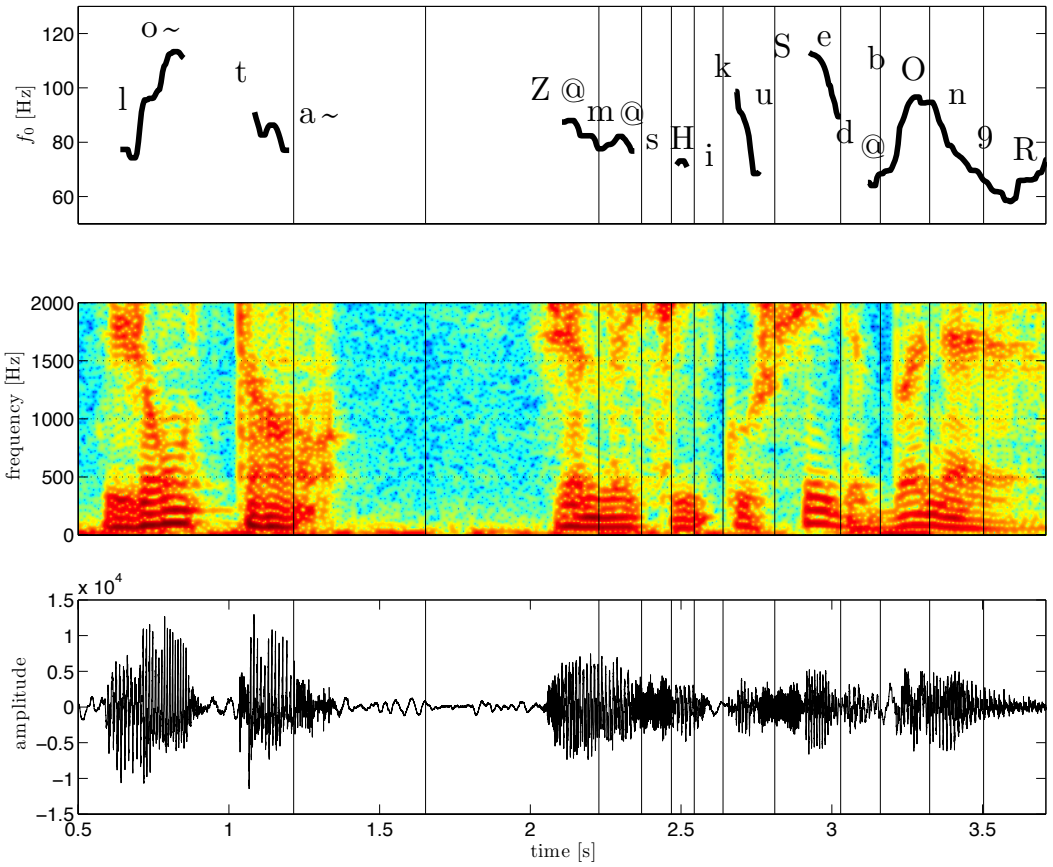


Figure 1.1: Illustration of unit-selection speech synthesis for the sentence: "*Longtemps, je me suis couché de bonne heure.*" ("*For a long time I used to go to bed early.*"). From top to bottom: f_0 variations, spectrogram, and speech waveform. Plain lines represent boundaries of the selected units.

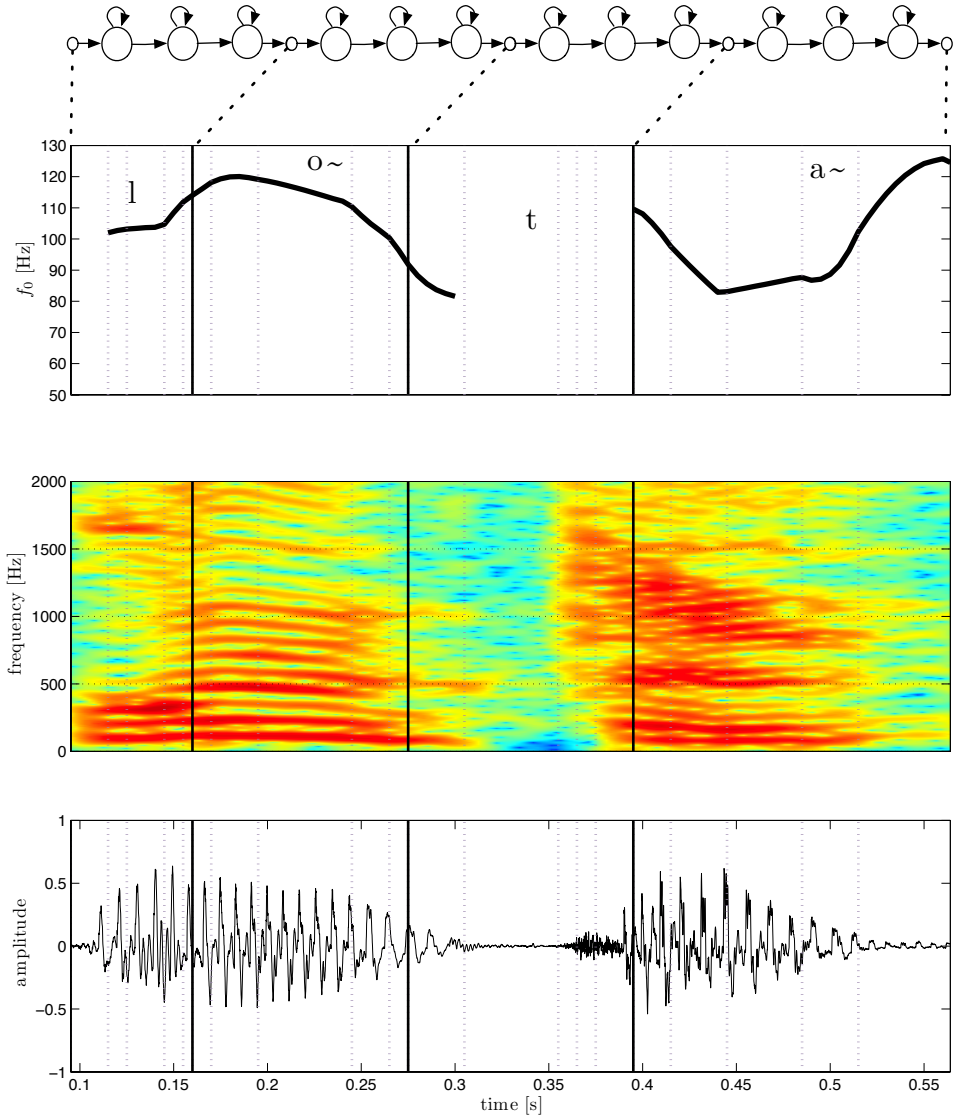


Figure 1.2: Illustration of HMM-based speech synthesis for an extract of the sentence: "*Longtemps, je me suis couché de bonne heure.*" ("*For a long time I used to go to bed early.*"). From top to bottom: f_0 variations, spectrogram, and speech waveform. Plain lines represent boundaries of the utterance context-dependent models (phoneme state). Dashed-lines represent boundaries of each context-dependent model (phoneme sub-states).

1.2.2 Current Issues in Speech Prosody Modelling

While conventional speech synthesis systems provide *intelligible* and *acoustically natural* synthetic speech, most of them remain *prosodically* poor. Actually, the development of current speech synthesis systems focused on the physiological speech production and short-term variations of speech parameters such as articulation and co-articulation, while the long-term variations that relate to higher levels of speech communication used to be ignored. Thus, a major issue for speech synthesis systems is the monotony of the synthesized speech, and the control of speech prosody. Consequently, research into the analysis and modelling of speech prosody has increased dramatically in recent decades and speech prosody has emerged as a crucial concern in speech synthesis. In particular, modelling the variability of speech prosody is required to provide natural and expressive speech in many applications of high-quality speech synthesis such as multi-media (avatars, video games, story telling) and artistic (cinema, theatre, music) applications.

The monotony of the synthesized speech prosody is due to the *poor dynamics* and *poor variability* of the generated prosodic parameters. Poor dynamic is due to the averaging problem that is inherent to characteristics of the statistical modelling. Poor variability is due to the generation of a stereotypical speech prosody, i.e. a speech prosody that does not vary with the context - linguistic, para-linguistic, or extra-linguistic.

The speech prosody dimension raises additional issues for the evaluation of a speech synthesis system: *correctness*, *variety*, and *liveliness*, which all participate in the naturalness of a synthesized speech utterance.

correctness : the speech prosody adequately reproduces that which can be expected from a native speaker. Correctness directly relates to the intelligibility of speech, since speech prosody is used by a speaker and a listener to organize the acoustic content so as to clarify the meaning that the speaker intends to convey. A correct speech prosody facilitates the intelligibility of a speech utterance, while incorrect speech prosody can degrade the intelligibility of a speech utterance. The correctness of a speech prosody mostly relates to the linguistic structure of the speech utterance since speech prosody is primarily used to clarify the meaning and the structure of the speech utterance.

variety : the variety of speech prosody refers to the variation in speech prosody depending on the context. The variety of speech prosody relates both to the intra and inter speech prosody variations that occur over utterances. Intra-variations denote the variations occurring within a speech utterance, while inter-variations denote the variations occurring across speech utterances. The variety of speech prosody relates either to linguistic (semantic, syntactic, discursive), para-linguistic (emotion, pragmatic,...), or extra-linguistic (individual strategy) characteristics. Semantic and syntactic contexts explain intra-utterance variations exclusively, while the other contexts explain either intra or inter utterance variations.

liveliness : The liveliness of a speech prosody includes the variety and the dynamic of speech prosody variations. Variety refers to the variety of prosodic contours that are observed within an utterance and across utterances depending on the context. Dynamic refers to the actual dynamic in the realization of a particular prosodic contour, the variations in speech prosody over and across utterances.

These criteria combine incrementally in the perception of the naturalness of a speech prosody: the perception of correctness in speech prosody mostly depends on the linguistic adequacy. Variety depends on the correctness and the variety of prosodic contours within and over utterances. Liveliness depends on the adequacy, the variety of the prosodic contours, and the actual dynamic of the local and global prosodic variations within and over utterances.

Speech prosody modelling decomposes into three issues:

the signal model that is used to represent the speech prosody variations;

The issue in the representation of speech prosody variations relates to the number of levels and domains on which relevant speech prosody variations occur. Firstly, the symbolic and acoustic characteristics of speech prosody are to be described and modelled. Secondly, speech prosody comprises variations that occur over various temporal domains and are associated with various communicative functions. One of the major issues in speech prosody modelling is to identify the temporal domains over which relevant speech prosody variations occur, to describe the variations over the different temporal domains with an appropriate representation, and to model the statistical characteristics of the speech prosody variations.

the context that is used to describe the linguistic characteristics of a text (syntactic, semantic, discursive), and additional para-linguistic and extra-linguistic information;

The issue in the description of the context is to identify the characteristics of a text and the environment of a communication that relate to relevant variations of speech prosody, and to model the speech prosody characteristics in context. Due to the variety of contexts that may be relevant, and in the absence of comprehensive studies on para and extra-linguistic characteristics, the description of the context generally remains limited to the linguistic context, with the exception of specific para or extra-linguistic information that can be easily described and modelled in specific speech databases (attitudes, emotions, geographical). In particular, the description of the morpho-syntactic structure of a text is the most commonly used linguistic description due to the availability of automatic linguistic chains and syntactic parsers, and the large number of studies on the syntactic/prosodic interface. However, the syntactic description generally remains limited to a surface description and is confined to the description over the sentence - considered as the maximal syntactic unit. In particular, description of the text structure over large text domains (discursive structure) and description of the semantic structure of a text are not considered due to the absence of available methods for their automatic extraction.

the statistical model that is used to model the speech variations in context.

The issue in the modelling of speech prosody variations is to model variations of speech prosody in context accurately, and in particular the long-term speech prosody variations. Modelling of speech prosody in context requires the identification of the contexts that relate to relevant speech prosody variations. One of the major issues in speech prosody modelling is the very large number of contexts of various nature that may interact in the production of speech prosody variations. In particular, modelling the variations in context increases exponentially when the description of the context is enriched. Modelling speech prosody variations in context becomes even more complex since different contexts may actually explain variations on different temporal domains. Modelling long-term variations requires the formulation of a statistical model that can adequately account for long-term variations, either by using an appropriate representation of the speech prosody variations or by reformulating conventional statistical methods. Finally, hybrid statistical methods are required to model adequately the temporal structure and the variations in the speech prosody parameters over time.

The major difficulty in speech prosody modelling is the number and the complexity of the variations to be modelled , and the huge number of contexts that may explain relevant speech prosody variations.

Finally, expert and statistical approaches adopt different methods to model speech prosody. Expert models provide a high-level formal description of speech prosody variations, but are generally limited to basic statistical methods or require the intervention of an expert linguist. Statistical models provide sophisticated methods to model speech prosody variations, but are generally limited to a crude linguistic description. Consequently, speech prosody modelling would

clearly benefit from the integration of the valuable formal description provided by expert linguists.

Current trends in speech prosody modelling intend to adapt conventional statistical methods to speech prosody modelling [Tokuda et al., 2003, Zen et al., 2004, Schmid and Atterer, 2004, Toda and Tokuda, 2007], model the speech variations over different temporal domains [Gao et al., 2008, Latorre and Akamine, 2008, Qian et al., 2009], improve the modelling of speech prosody variations in context [Yan et al., 2009], and extend to the statistical modelling and adaptation of speaking style [Yamagishi et al., 2004, Yamagishi, 2006, Bell et al., 2006].

1.3 Scope of the Thesis

The main objective of the thesis is to develop a speech prosody system that can be used to control, vary, and adapt the speaking style of a speaker in speech synthesis. The principle of the *MeLos* system that is presented in the thesis is to provide the speech prosody of a speaker that corresponds to a given text. The synthesized speech prosody is then used to control or adapt the speaking style of a speaker in speech synthesis. Speech prosody is a crucial issue in current speech synthesis systems, in which a natural, expressive, and varied speech prosody is desired either by improving the local/global dynamic of the speech prosody or the variety of speech prosody that can be used to control the prosodic strategy or the speaking style of a speaker. The research directions correspond with the necessity of integrating a rich description of the text characteristics, an appropriate description of speech prosody, and an adequate statistical modelling.

Special attention is paid to the combination of theoretical linguistic and statistical modelling to provide a complete speech prosody system that can be used in speech synthesis systems. In particular, speech prosody is described from signal variations to abstract representations. A unified discrete/continuous context-dependent HMM is used to model each of the linguistic levels separately. A linguistic processing chain is proposed to enrich the description of the text that is used to model the speech prosody variations in context. A context-dependent HMM is proposed to model the f_0 variations based on stylization and trajectory modelling over various temporal domains. The proposed method is used to model the prosodic strategies and the speaking-style that is specific to a speaker, and is extended to model the speaking of any arbitrary number of speakers using speaker normalization techniques. The proposed system is used either to model the speaking style of a speaker with a large read speech database or the speaking style that is shared among speakers with relatively small natural speech databases.

Additionally, the *MeLos* system is not restricted to the development of speech technologies only. The recent emergence of high-quality speech technologies at IRCAM (speech recognition, transformation, and synthesis) relates to a profound interest and demand by musicians and composers, and to a large range of artistic applications. In particular, speech prosody - the musical dimension of speech - is the central dimension that relates speech and music, and has fascinated composers and artists for a long time. Thus, the development of a speech prosody system found many applications in artistic creation during the thesis, from real-time control of speech material, control of musical phrasing driven by speech prosody, or control of speech synthesis driven by musical phrasing.

1.4 Major Contributions

In this thesis, the *MeLos* system is presented for the analysis and synthesis of speech prosody and speaking style. The major contribution of the present work is the special attention to combine theoretical linguistic and statistical modelling so as to provide a complete speech prosody system that can be used in speech synthesis systems. The main contributions consist of: 1) the design of a complete speech prosody system based on discrete/continuous context-dependent HMM models, 2) the enrichment of the linguistic description that is used in context-dependent modelling, 3) the symbolic modelling of speech prosody based on segmental HMM and Dempster-Shafer fusion, 4)

the acoustic modelling of speech prosody based on the stylization and the simultaneous modelling of short and long term speech prosody variations, and 5) the discrete/continuous modelling of speaking style. To a lesser extent, the modelling of speech prosody alternatives to vary speech prosody in speech synthesis is proposed. The proposed contributions are validated based either on subjective and/or subjective evaluations.

1.4.1 A unified discrete/continuous context-dependent model

A discrete/continuous context-dependent Hidden Markov Model (HMM) is proposed to model the speech prosody variations at the symbolic and the acoustic level. A discrete HMM is used to model the phonological variations in context (e.g., prosodic prominence, prosodic break). A continuous HMM is used to model the acoustic variations in context (e.g., melodic variations, and prosodic timing). Then, a context-dependent model is derived using a conventional context-clustering method based on *Maximum-Likelihood Minimum-Description-Length* (ML-MDL). The syllable is the minimal prosodic unit that is used for the modelling of speech prosody variations.

1.4.2 Rich linguistic context modelling

An automatic linguistic processing chain is used to enrich the linguistic description of a text in context-dependent HMM speech prosody modelling. The linguistic processing chain includes text pre-processing, surface parsing, and deep parsing. A preprocessing is conducted to segment a raw text into linguistic units that can be used by a linguistic parser (e.g., sentence and form). Surface parsing is processed to provide a morpho-syntactic analysis for each sentence. Then, Deep parsing is then achieved based on *Tree Adjoining Grammar* (TAG) which represents both the dependency graph and the constituency structure derived from each sentence. The extracted syntactic features are classified into different sets depending on their nature: morpho-syntactic features are extracted from the surface parsing, dependency and constituency features are extracted from the deep parsing, and adjunction features are additionally introduced which are retrieved from the deep parsing.

1.4.3 Symbolic Modelling of Speech Prosody Based on Segmental HMMs and Dempster-Shafer Fusion

A statistical method that combines linguistic and metric constraints in the modelling of prosodic breaks is proposed based on segmental HMMs and Dempster-Shafer fusion, and the relative importance of linguistic and metric constraints is assessed depending on the nature of the linguistic information. A discrete segmental HMM is used in which prosodic breaks are modelled conditionally to the linguistic context in which they are observed, and the distance between successive prosodic breaks (length of a prosodic group) is explicitly modelled. Dempster-Shafer fusion is used to balance the linguistic and metric constraints into the segmental HMM. The relative importance of the linguistic and metric constraints is assessed depending on the nature of the linguistic information.

1.4.4 Stylization and Trajectory Modelling of Speech Prosody

A trajectory model based on the stylization and the simultaneous modelling f_0 variations over various temporal domains is presented. First, the syllable is used as the minimal temporal domain for the description of speech prosody, and f_0 variations are stylized over various temporal domains which cover short-term and long-term variations (e.g., syllable, k-order syllable context, internal prosodic group, prosodic group) using a *Discrete Cosine Transform* (DCT). Then, the description of f_0 variations is formed by the joint description of short-term variations over the syllable and long-term variations that occur over long-term temporal domains. During the training, the joint short/long term description of f_0 variations is used to estimate context-dependent HMMs. During the context-clustering, the clustering of short-term characteristics is driven by long-term trajectories occurring over long-term temporal domains. During the synthesis, short-term f_0

characteristics are determined using the long-term variations as trajectory constraints.

1.4.5 Modelling and Adaptation of Speaking Style

Finally, a study on the modelling of speaking style for speech synthesis is presented, and the issue of speaking style from the cognitive description of speaking styles to the modelling in speech synthesis is addressed. First, the design of a speech database in four speaking styles that correspond to specific situations of communication - discourse genres (DGs) - is described. A preliminary experiment investigates whether listeners can distinguish speaking styles related to different communicative situations. The identification ability of speaking styles and the similarity that exists across different speaking styles is used to instantiate a reference for the evaluation of speaking style modelling in speech synthesis. In parallel, an average discrete/continuous context-dependent HMM is used to model the symbolic/acoustic characteristics of speaking style in speech synthesis. The ability of the model to model the speech characteristics of a speaking style is assessed. Finally, a speaker-independent modelling of speaking style based on shared context-dependent modelling and speaker normalization is presented to adapt the speaking style of a speaker in speech synthesis. The ability of listeners to distinguish speaking styles (natural speech and synthetic speech) is based on identification experiments using delexicalized speech or neutral text, and the identification obtained with synthetic speech is compared to that obtained with natural speech.

For clarity, various contributions on the automatic transcription of speech prosody [Obin et al., 2008c, Obin et al., 2008a, Obin et al., 2009b], the analysis [Obin et al., 2008d, **?**, Avanzi et al., 2011b] and modelling [Obin et al., 2009a] of speech prosody, and the classification of speaking style [Obin et al., 2008b] are not presented.

1.5 Outline of the Thesis

The document is organized into three parts: the state-of-the-art on the analysis and modelling of speech prosody is described in part I, the speaker-dependent speech prosody model is presented in part II, and modelling and adaptation of speaking style is presented in part III[4].

The state-of-the-art on the analysis and modelling of speech prosody is described in part I. The different levels of speech prosody analysis are presented in chapter 3, from the acoustic dimensions, the prosodic contours, to the symbolic representations. The architecture of a speech prosody system and the state-of-the-art on the modelling of speech prosody are described in chapter 4.

The discrete/continuous modelling of the symbolic/acoustic speech prosody characteristics of a speaker is presented in part II. The text and speech material used for the speaker-dependent modelling is described in chapter 5. The principles of the Hidden Markov Model (HMM) and the context-dependent HMM are described in chapter 6. The linguistic processing chain that is used for the rich description of a text structure is presented in chapter 7. The symbolic modelling of speech prosody based on context-dependent discrete HMM, segmental HMM, and information fusion, is presented and evaluated in chapter 8. The acoustic modelling of speech prosody based on context-dependent continuous HMM, stylization, and trajectory modelling, is presented and evaluated in chapter 9.

The application of the discrete/continuous model to the modelling of speaking style is presented in part III. The design of a speaking-style speech database and a preliminary study on the identification of speaking style are presented in chapter 10. The average modelling of speaking style and the ability of the discrete/continuous HMM to model a speaking style are addressed and evaluated in chapter 11. The speaker-independent modelling of speaking style based on stylization

[4]The utterance: "*Longtemps, je me suis couché de bonne heure.*" ("*For a long time I used to go to bed early.*") will be used as a simple study case over all of the document.

and speaker adaptive training is presented and evaluated in chapter 12.

Finally, the main contributions of the thesis are summarized and further directions are discussed.

Chapter 2

An Introduction to Speech Prosody: *The Music of Everyday Speech*

Contents

2.1 Prologue: The Voice or the "*Dialogue de l'Ombre Double*"

In the beginning: my mother's voice. Even before our birth, the human voice is one of the first experiences we have of life and is doubtless one of the first experience we have with the outside world. Through childhood, the voice becomes a space for the discovery and exploration of our body and its possibilities (babbling, vocal play, vocal mimicry), and a place for the gradual appropriation of our individuality. The voice is constitutive of our individuality; it is a part of the construction and affirmation of our identity, its reflection. The voice is also characteristic of our participation in a collective future; it carries the traces of our history, our origins, our milieu, and our culture carried through the accumulated layers of our life and our experiences. Lastly, the voice is the "mirror of the soul". Through its particular inflections and tiny modulations the voice reveals the workings of the soul, our emotional states, and our immediate feelings. From a singular experience to emotions, from the individual to the collective, the voice constitutes the articulation of the unique and the universal. In particular, the voice used to be considered in antiquity as the principal mediation between the human and God. The voice divides and multiplies: mirror of the soul, mirror of humanity, mirror of the divine.

How to Tell. What is the Word. The voice is co-substantial with our connection with the world and with others. The human being needs to express himself and be understood, he must *communicate* with others: voice turns into language and becomes *speech*. Speech is probably one of the most universally shared practices of humanity, the act the most commonly experienced by each of us, *the music of every day*. Banal, but rich: through speech, our representation of the world and our most intimate experiences can be expressed with a seemingly infinite degree of variation,

making each expression a unique act. Through the expression of speech, a "world", a "universe" [Proust, 1927] is revealed. Complex, speech remains an evasive object that continuously escapes definition, no doubt because its study aims at the understanding of mankind and its relationships.

Speech and the Poet. The poet understands the double nature of speech from an early age by distinguishing the musical dimension (*melos*) from the literal meaning (*lexis*): speech prosody [Aristotle, 0 BC], and the expressive potential of speech [Rousseau, 1781, Artaud, 1938, Deleuze and Guattari, 1975, Boulez, 2005]. For the poet, the voice has a primitive function that may be related to pre-linguistic expression forms "inarticulé" [Rousseau, 1781], "sons inclassables" [Boulez, 2005], "énergies vives" [Artaud, 1938] [1]) associated with sensorial content (des "vibrations" [Rousseau, 1781], des "sensations" [Artaud, 1938], des "résonances" [Boulez, 2005][2]) through a regressive process of purely organic pleasure ([Artaud, 1938, Deleuze and Guattari, 1975, Boulez, 2005]). This physical and plastic expression is incarnated by speech prosody ("prendre les intonations de manière concrète absolue" [Artaud, 1938]).

Speech prosody is commonly referenced as the "way of speaking" or, by analogy with music, referred to as: "the music of speech" [Wennerstrom, 2001]. From this original analogy, spun throughout the history of language, music, and philosophy, the relationship between speech and music finds its motivation in its acoustic substance and in its co-substantial ambiguity: speech prosody and music do not have a meaning, they mean; they suggest and evoke but do not represent anything; they relate to semiotics, not to semantics, to a connoted rather than a denoted meaning. This relationship still maintains all of its mystery and its origins will probably remain hidden. Speech prosody, the music of speech and the foundation of human communication, is the subject of the present study.

The rest of this part will give a comprehensive introduction to speech prosody. Fundamentals of speech prosody are presented without pretending to provide an exhaustive and definitive definition. The introduction is based on three axes. First, the notion of speech prosody is introduced in the context of speech communication. Then, the different levels of speech prosody (substance, form, and functions) are described from the acoustic continuum to the emergence of abstract linguistic objects. Finally, the different sources of speech prosody variations are described, and the notion of *speaking style* is introduced.

2.2 Speech Communication

Communication : Communication is a complex process that is characterized essentially by the transmission of a *message*. Information theory formalized communication early on in the 1940s, notably with the *Shannon and Weaver* model [Shannon, 1948] that essentially accounts for physical systems of information transmission: communication consists of the transmission of a message produced by a source to a target through a signal encoded by a sender and decoded by a receiver. This transmission is carried out through a canal through which the emitted signal is altered by parasite phenomena.

Human Communication: Numerous studies have been carried out to extend the mechanistic definition of the original model to the description of the specificities of *human communication* and especially of *inter-personal communication* [Schramm, 1954, Barnlund, 1968]. This enrichment is primarily due to the integration of *interaction* and *context* that are characteristic of inter-personal human communication.

The interactive dimension is based on the principles of *retroaction* and *enaction* [Condon, 1971, Gallese, 2003, De Jaegher and Di Paolo, 2007]. Retroaction means that the receiver of a message is not just the passive receiver of the message being conveyed, but actively

[1] "inarticulate", "unclassifiable sounds", "energies"
[2] "vibrations", "sensations", "resonances"

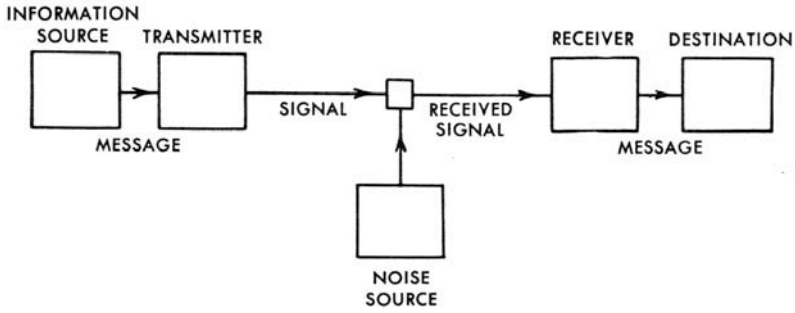


Figure 2.1: Schematic diagram of a communication system, after [Shannon, 1948].

contributes to the collaborative construction of the message in the communication process. Enaction means that the sender is both the sender and the receiver of his own message, the speaker and the listener, the actor and the spectator.

The contextual dimension refers to the contribution of the context of a communication in the communication process, in that the interpretation of a message may depend on the actual context of communication (e.g., immediate context of the communication, or mutual history of the individuals involved in the communication process).

In other words, inter-personal human communication is not reduced to a linear transmission system in which the sender and the receiver are passive elements, but constitutes a place for the interactive construction and transmission of mutually shared information.

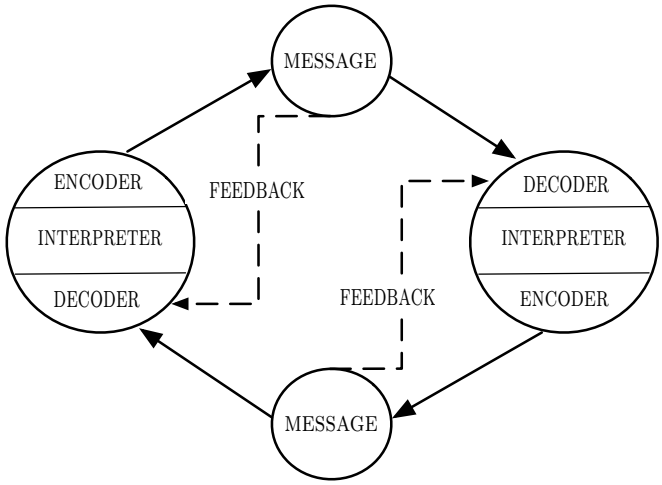


Figure 2.2: Schematic diagram of inter-individual communication, after [Schramm, 1954].

Human communication also presupposes the existence of a set of conventions that are shared by a group of individuals, i.e. the existence of a linguistic code that is prior to communication, and is required for the formulation and understanding of a message among a group of individuals.

Multi-modality of human communication : Human communication is, by essence, multi-modal insofar as the human communication process implies all the means of production and perception of the human being that can be used to communicate [Mehrabian, 1972, Andersen, 1999]:

gesture, voice, and touch are the principal dimensions for production, and symmetrically vision, hearing, and touch for perception. Other sensorial dimensions exist but remain little-known or are of less importance in human communication. Each of these channels is associated with a set of codes that is more or less universally shared and more or less conventional.

Human communication therefore resides in the co-production and co-integration of a set of signs conveyed through the different human sensory dimensions within a dynamic process involving all the participants in the communication.

Speech Communication: Among all the dimensions of human communication, the oral dimension is by far the most extensively studied [Bühler, 1934]. Several reasons explain this trend: firstly, speech is directly observable (originally with written transcriptions and later with audio recordings), and is also connected with the Western tradition of written language. In other words, speech historically relates to meaning, to the written meaning. From this point of view, the study of speech communication has largely benefited from the emergence and development of structural linguistic research carried out throughout the 20th century. In particular, the balance of writing and orality in speech communication has been gradually reversed, and today speech is no longer considered as the passive vehicle of a primary written meaning; rather *speech makes sense* through a dynamic construction of meaning. More recently, other dimensions of human communication, such as gesture and vision, and their interaction with speech have emerged as research domains. However, studies remain relatively scarce, and these domains still remain under-developed when compared to speech communication.

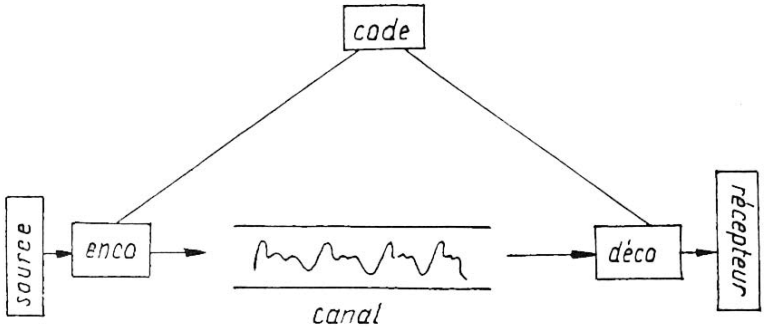


Figure 2.3: Schematic representation of speech communication, after [Fonagy, 1983]

2.3 Speech Domains

Studies in phonetics generally distinguish two domains in speech communication: the *linguistic domain* that is used to convey the primary linguistic message, and the *para-linguistic* domain that is used to convey the suggested meaning. The linguistic domain refers to the semantic meaning which explicitly refers to the linguistic system. The para-linguistic domain refers to the signs associated with the context of a communication, that can be used to interpret the suggested meaning (e.g., spatio-temporal context, intention of a speaker, emotional state of a speaker).
Finally, the extra-linguistic domain can be added as an additional dimension, and refers to the characteristics that are specific to an individual (e.g., individual characteristics, socio-professional and geographical origins) [Lacheret-Dujour and Beaugendre, 1999].

While the linguistic dimension has focused research on speech communication, recent studies have pointed out the importance of information conveyed by the non-linguistic domain in the communication process. For instance, para-verbal phenomena such as hesitations, reformulations, sighs,

breathing, and laughter used to be considered as parasitic phenomena of speech communication. However, recent studies on *spontaneous speech* have provided evidence that non-linguistic phenomena occupy a considerable place in speech communication, with respect either to the relative frequency of occurrence in speech [Schober and Brennan, 2001, Vettin and Todt, 2004] or to the information content conveyed [Nicholson et al., 2003, Campbell and Erickson, 2004].

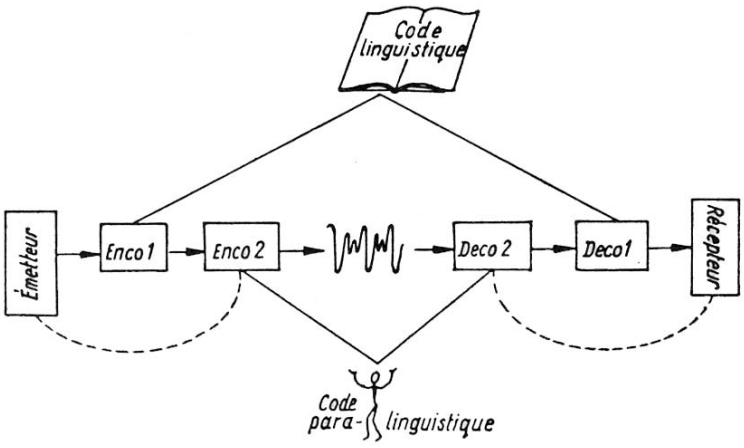


Figure 2.4: Schéma du double codage de la parole, d'après [Fonagy, 1983]

2.4 Speech Prosody: From Signal to Communicative Functions

Speech prosody is a challenge for a simple and shared definition. Speech prosody covers all of the speech communication domains from linguistic to extra-linguistic with multiple forms (substance, form, functions). The polyvalence of speech prosody makes a generic definition problematic, even impossible.

Speech prosody, historically considered as *marginal* [Martinet, 1956], was defined by a double opposition to conventional phonetics: the temporal domain and the linguistic domain [Fonagy, 1983].

temporal domain : the *phonetic* level refers to *segmental* variations, i.e. acoustic variations occurring over a short-term temporal segment (phoneme), while the *prosodic* level refers to *supra-segmental* variations, i.e. acoustic variations occurring over a long-term temporal segment.

linguistic domain : the phonetic level used to be associated with the linguistic domain, while the prosodic level used to be associated with the para-linguistic domain.

However, this definition remains partial, insofar as speech prosody is not limited to the para-linguistic domain, but includes all the domains of speech communication. Indeed, speech prosody conveys linguistic, paralinguistic, and extra-linguistic information, either to ensure the organization of a discourse and its semantic cohesion, to convey the intentions and emotions of a speaker, or to convey the habits, the socio-professional status, and the geographical origins of a speaker.

This original duality [Delattre, 1969, Fonagy, 1983, Lacheret-Dujour and Beaugendre, 1999] actually goes back to pre-linguistic origins. A brief look at studies on animal and emotion communication provides a better understanding of the ambivalence of speech prosody in human

communication [Ohala, 1996]: the emergence of linguistic conventions is the result of a gradual process from observation to abstraction, from the individual to the universal. This process may have originated in a biological mechanism in which a sound is associated with the specific conditions (e.g., physiological, situational) in which it is produced. The stabilization of this association may favour the emergence of an abstraction in which a sound and the condition of production can be arbitrarily substituted for each other. Finally, a distancing mechanism ultimately substitutes the concept for the position of the individual with regard to this concept. The original duality of expression and abstraction is co-substantial to speech prosody: crystallized in the conventions of human communication, speech prosody bears the traces of its potential primitive expression. The multiplicity of speech prosody results from the co-occurrence of different degrees of stratification from primitive expression forms to linguistic conventions. In this manner, [Morlec, 1997] has proposed a continuous classification of emotions, attitudes, and modalities.

The study of speech prosody reproduces this creation based on a bottom-up process from acoustic expression to abstract representations. This description distinguishes three levels of representation: *substance*, *form*, and *function*.

substance refers to the materiality of the speech signal and in particular to the acoustic dimensions used to convey prosodic information.

form refers to the articulation of the continuous and the symbolic. This articulation corresponds to the association of a set of distinctive formal categories to the variations of speech prosody substance. This association is achieved either in a bottom-up process through the emergence of distinctive forms from the prosodic substance, or in a top-down process through the mediation of expectations associated with high-level linguistic processing.

The formal description of French prosody originated in the description of elementary distinctive *contours* [Delattre, 1966], then various representations were formulated for the description of *contours* [Martin, 1975, Mertens, 1987] and *tones* [Pierrehumbert, 1980, Hirst and Di Cristo, 1998, Post, 2000, Jun, 2005].

The formal representation presupposes the determination of the acoustic dimensions that are involved in the prosodic substance, the definition of temporal domains over which relevant speech prosody variations occur, and the identification of a set of distinctive contours that can be associated with each of these domains. However, the complexity of speech prosody is such that the emergence of a comprehensive and exhaustive representation is extremely difficult.

function refers to the relation that relates a form and its meaning. Consequently, speech prosody consists of as many functions as there are domains in the process of speech communication. Linguistics usually distinguishes two principal components: the *linguistic function* and the *expressive function*, respectively associated with the linguistic and the para-linguistic domains. The linguistic function is associated with the markers that are used to instantiate and clarify the linguistic structures of the utterance and the discourse (syntactic structure structuring, semantic cohesion, and discourse organization). The expressive function is associated with the markers that are used to instantiate the suggested meaning (e.g., emotional state of the speaker).

2.5 Making Sense of Variations

Speech communication requires and presupposes the existence of an *invariant* system that guarantees the possibility of communication through shared competence (norms and conventions). Conversely, *variation* is actually co-substantial with speech communication [Labov et al., 1968]. Since any deviation is in essence significant, each language variation - insofar as a variation constitutes a deviation from a norm - conveys an information content and acquires a communicative function: *variations make sense*. This is how, for instance, a tiny inflexion in the voice potentially conveys the emotional state of a speaker.

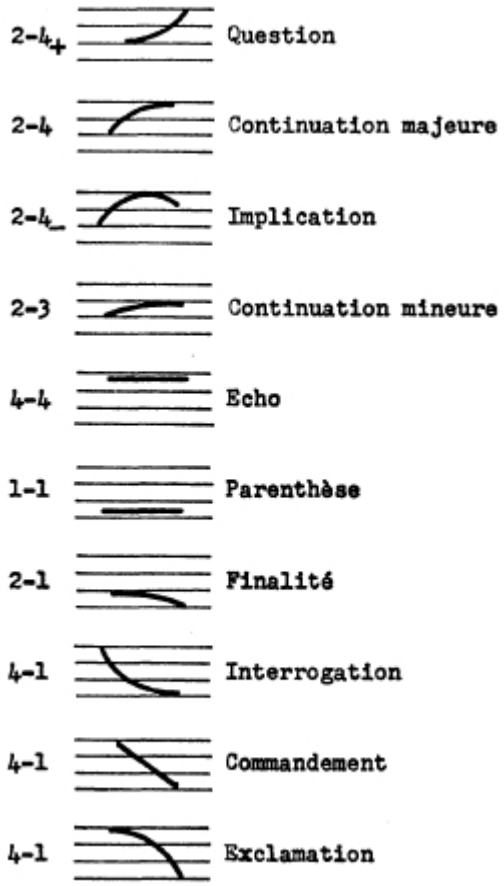


Figure 2.5: Original description of the ten elementary intonation contours of French and their associated functions, after [Delattre, 1966]

Above all, variation is constitutive of language: while language is a universal prerequisite for communication, and the principle of *invariance* is a *sine qua non* condition of a language, the principle of *variation* is co-substantial to it. Invariance and variation are the internal principles of all languages: invariance to guarantee the stability of shared communication, and variation to ensure differentiation of the meanings conveyed. Two types of variation exist: internal variations associated with the configuration of a specific linguistic system, and external variations associated with the differences among linguistic systems and their diversity (6,000 languages are, or have been, spoken; 300 to 500 language groups are genetically connected).

Variation actually conveys information relative to the message conveyed and its more or less immediate context but also information about the speaker, the his emotional state, intentions, habits, beliefs, socio-cultural and geographical background, etc. The principle of variation affects all domains of speech communication (linguistic, paralinguistic, and extra-linguistic) and in particular the linguistic domain (morphological, lexical, syntactic, phonemic, prosodic, etc.). By limiting the description to the oral dimension of speech communication (phonetics and prosody), a large variety of causes may explain the variations in speech, each associated with a specific domain.

The source of variations occurs through a variety of modalities, and conveys different types of information:

1. linguistic: variations are used as constraints that are directly related to the underlying lin-

guistic structure of the utterance. In speech prosody, the constraints are lexical (syllable stress in the case of lexical-stress languages such as English), syntactic (instantiation of the syntactic structure), semantic (cohesion and focus), and discursive (discourse organization).

2. para-linguistic: variations are used as indices associated with the context of the communication (e.g., position of the speaker with regard to the utterance, intention of the communication, implication of the speaker, relationship among the speakers), the emotional state of the speaker, and the specific situation of communication (e.g., spatio-temporal context, communication media).

3. extra-linguistic level: variations relate to orthogonal dimensions: synchronic (variations at a given time) and diachronic (variations over time).

A descriptive summary of the different sources of acoustic variations observed in speech communication is presented in table 2.1.

type	variable	description
linguistic	syntactic	speech prosody characteristics that are used to facilitate syntactic parsing [Selrik, 1984, Dell, 1984, Price et al., 1991, Ladd, 1996, Delais-Roussarie, 2000]
	semantic	speech prosody characteristics that are used to facilitate lexical access and semantic parsing [Shattuck-Hufnagel and Turk, 1996, Cutler, 1997]
	discursive	speech prosody characteristics that are used for discourse processing [Pierrehumbert and Hirschberg, 1990, Cohen et al., 2001]
	modality	modality of an utterance (e.g., question, exclamation) [Delattre, 1969, Kratzer, 1981]
	attitude	position of a speaker toward the linguistic content of an utterance (e.g., irony, doubt, surprise, evidence) [Morlec, 1997, Shochi et al., 2009]
para-linguistic	emotion	emotional state of a speaker (e.g., happiness, sadness, anger, fear)[Scherer et al., 1991, Ohala, 1996, Bachorowski, 1999]
	pragmatics	suggested meaning related to the communicational context [Austin, 1962]: speaker's intention, communicative's intention, speaker's belief, speaker's involvement, implicatures, spatio-temporal context, as well as mutual relationship
extra-linguistic	physiological	gender, age, intrinsics and co-intrinsics characteristics related to articulation and co-articulation [Di Cristo, 1985]
	idiolectal	variety of a language that is unique to an individual: *"the language of the individual, which because of the acquired habits and the stylistic features of the personality differs from that of other individuals and in different life phases shows, as a rule, different or differently weighted [communicative means]."* [Dittmar, 1996]
	geographical	variety of a language that is characteristic of a group of a language's speakers associated to a particular geographical region [Walter, 1982, Delais-Roussarie and Durand, 2003, Gilles and Peters, 2004].
	sociological	variety of a language that is characteristic of a group of a language's speakers associated to a particular social group [Labov et al., 1968, Labov, 1972].
	situational	variety of a language that is characteristic of a group of a language's speakers associated to a specific situation of communication [Koch and Oesterreicher, 2001, Simon et al., 2009]
	temporal	variations of a language over time [Ohala, 1993, Boula de Mareuil et al., 2008]

Table 2.1: Typology and domains of speech variations.

2.6 Speaking Style: a matter of Identity, Situation & Time

"*Les gestes vocaux qui constituent le style se transforment: ils disparaissent en tant que gestes phonatoires, porteurs de messages, pour réapparaître comme manière de parler individuelle. Cela revient à dire que les gestes vocaux [...] seront directement rattachés, sans analyse sémantique préalable, à la personne du locuteur pour faire partie de son signalement, au même titre que la couleur de ses cheveux, sa taille, son nom.*".

IVAN FONAGY, LA VIVE VOIX.

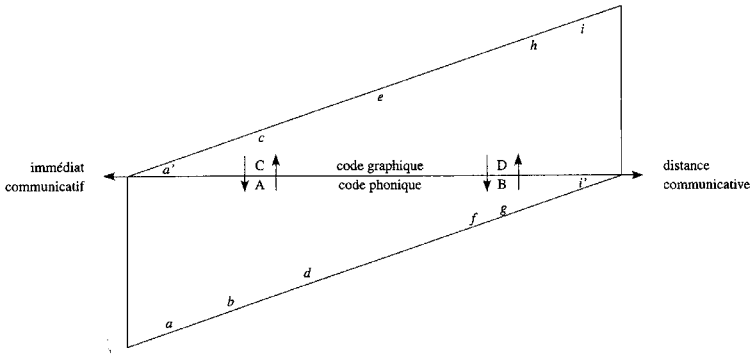
The concepts of style and genre find their origins in the foundations of modern Western society [Aristotle, 0 BC, Plato, 0 BC]. Their definition and their mutual relationship have been the subject of continual consecutive studies in the fields of theater, poetry, literature, philosophy, and linguistics [Hegel, 1835, Benvéniste, 1966, Todorov, 1978, Genette et al., 1986, Rastier, 1989] and more recently in speech communication [Bakhtin, 1984, Halliday, 1985, Biber, 1988]. The difficulty of a proper definition is such that some even deny the reality of these concepts. Faced with the difficulty of defining these concepts, the present description will be limited to a general definition, and confined to the description of speech communication.

Each individual is characterized by a variety of vocal characteristics that are unique to him and distinguish him from others, depending on his physiological characteristics, his idiolect, and the variety of his dialectal origins: a *speaking style* which constitutes his *vocal signature*, and contributes in the construction of his *identity* [Fonagy, 1983, Léon, 1993, Lacheret-Dujour and Beaugendre, 1999].

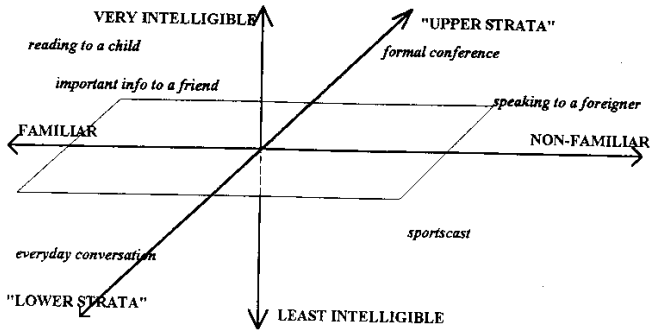
However, the style of an individual is not invariable, and each individual continuously adapts his style depending on specific situations of communication. Each situation of communication determines a specific mode of production that is associated with it: a *genre*. A genre is defined by a set of conventions of form and content that are shared among all of its productions [Bühler, 1934, Benvéniste, 1966, Bakhtin, 1984, Koch and Oesterreicher, 2001]. In particular, a specific discourse genre relates to a specific *speaking style* [Fonagy, 1983, Léon, 1993, Lacheret et al., 2009, Simon et al., 2009, Degand and Simon, 2009]. Consequently, an individual adapts his speaking style to a specific situation depending on the formal conventions that are associated with the situation, his a-priori knowledge about these conventions, and his ability to adapt his speaking style. In other words, each communicative act instantiates a style composed of a *unique* speaking style that is constitutive of the individual identity, and a *shared* speaking style that is conditioned by a specific situation. The individual and the genre, the unique and the shared, constitute the stylistic dimensions of speech communication.

Finally, style is not invariable over time, but changes from one epoch to another: each epoch has its own styles depending on the evolution of the language and the cultural and linguistic conventions. Just as an individual, a community, or a genre can be identified from its style, it is also possible to identify an epoch by its styles. Additionally, styles and genres have continuously moving forms, constantly updated throughout time, in the dialectic between the individual and the collective. The emergence, transformation, and disappearance of styles and genres are conditioned by the complex evolution of collective practices with regard to individual productions and collective conventions, depending on the socio-cultural context. For instance, the emergence of new forms of expression can result from the contingency of divergent individual productions, and a socio-cultural context that favors their development and their sedimentation in collective practices.

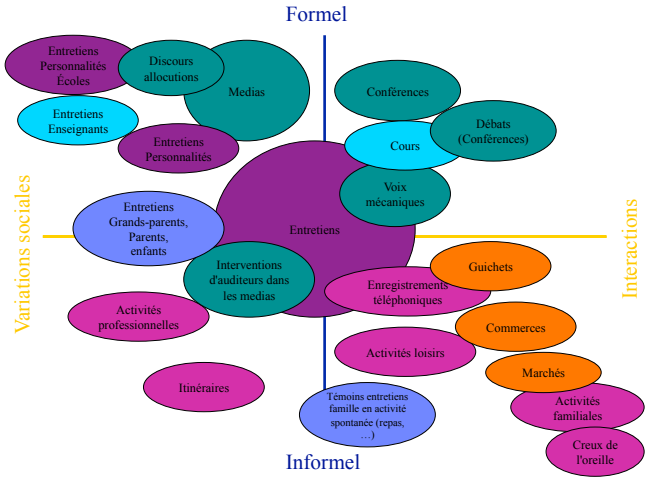
For a general definition, *speaking style* will refer to the speech characteristics that are specific to an individual, a linguistic community, a situation, or an epoch; and a *prosodic strategy* will refer to the specific use of speech prosody by an individual depending on a specific situation.



(a)



(b)



(c)

Figure 2.6: Multi-dimensional representations of speaking style. (a) medial/conceptual representation, after [Koch and Oesterreicher, 2001] ; (b) intelligibility/familiarity/social representation, after [Eskenazi, 1993]; (c) formal/interactional representation, informal description after Olivier Baude.

Part I

Analysis and Modelling of Speech Prosody - a Survey

Abstract

In this part, a description of the state of the art in the analysis and modelling of speech prosody is presented.

For the analysis of speech prosody, the description of speech prosody is presented from the determination of the acoustic dimensions to the abstract representations. The five acoustic dimensions of speech prosody are described (e.g., f_0, timing, and to a lesser extent intensity, vocal quality, and degree of articulation). The temporal domains that are used for the description of relevant speech prosody variations are described (e.g., syllable, prosodic phrase), and stylization methods for the description of speech prosody contours (based on production, perception, linguistic, or signal processing) are presented and discussed. Finally, methods for the symbolic transcription and/or the phonological representation of speech prosody are presented and discussed for the transcription of French prosody. In particular, the Rhapsodie transcription system that has been developed in parallel to this study is presented and discussed.

For the modelling of speech prosody, the architecture of a speech prosody system is presented and conventional methods are described. The principle of the modelling of speech prosody in context is briefly summarized including linguistic, para-linguistic, and extra-linguistic contexts. Conventional methods used for the text analysis based on surface syntactic parsing (POS, chunks) are described. Then, conventional methods for the analysis of speech prosody (automatic transcription of speech prosody) and the temporal domains considered (e.g., phoneme, syllable, prosodic phrase) are described. Finally, conventional methods for the symbolic and acoustic modelling of speech prosody based on the combination of expert and statistical models are described. In particular, a comparison of the short-term and long-term modelling of speech prosody is presented and discussed.

The present part is organized as follows: the analysis of speech prosody is described in chapter 3, and the modelling of speech prosody is presented in chapter 4.

Chapter 3

Analysis of Speech Prosody: From Signal to Abstract Representations

Contents

3.1 Dimensions of Speech Prosody

A speaker may potentially use any acoustic characteristic of the voice to convey information in speech communication. Some are denotative and referential to a linguistic system that is assumed to be shared among speakers, some are connotative and rather suggest than refer to an explicit meaning. Five acoustic dimensions of speech prosody are currently referenced in literature:

F_0 : refers to the variations of the fundamental frequency of the source excitation (f_0) over time [Maeda, 1974, Fujisaki, 1981]. Intonation - also referenced as melodic variations or melodic phrasing [Hirst and Espesser, 1993, Mertens, 2004a] - is the acoustic dimension that has been the most widely studied in speech prosody. In particular, this results into the description of tones and melodic contours, and the elaboration of complex intonation systems and intonational phonology.

timing : Timing is the acoustic anchor of speech rhythm - which is probably the most difficult prosodic dimension to describe in speech [Campbell, 2000]. Many acoustic correlates actually

interact in the production and the perception of speech rhythm. The syllable is widely referenced as the minimal prosodic unit to describe speech timing in syllable-based languages [Ladd and Campbell, 1991]. Other syllable-like timing measures exist, such as vocalic onset [Hermes, 1987, Mertens, 2004a], or perceptual centres [Morton et al., 1976, Scott, 1993] ([Barbosa, 2004] for French). More complex acoustic anchors exist, such as prosodic prominence ([Pasdeloup, 1992, Delais, 1994] for French). Finally, continuous measurements of speech rhythm variations have been proposed, referenced as local speech rate [Ohno and Fujisaki, 1995, Pfitzinger, 1998].

intensity : refers to the intensity of the speech signal. While intensity is generally assumed to encode prosodic information, few studies on the used of intensity in speech prosody exists [Delattre, 1938]. Physiologically, intensity variations partially correlate with fundamental frequency variations in case of intonational prominences [Atkinson, 1978, Beckmann, 1986, Campbell, 1992]. Nevertheless, recent studies pointed out the role of intensity variations into prosodic organization [Tseng and Lee, 2004], and information focus [Beaver et al., 2007] independently of fundamental frequency variations. Several measurements of the speech intensity exist, from the conventional short-term intensity measure, to the long-term integration over prosodic units. Additionally, a refinement of the intensity measure which accounts for perception (loudness, [Fletcher and Munson, 1933]) has been proposed, in which the perceived intensity is measured with respect to the frequency content and the duration of the speech segment.

voice quality [Campbell and Mokhtari, 2003]: refers to the characteristics of the glottal excitation (e.g. breathy, pressed, tense, whispered, creaky, rough). Measures of the voice quality have been recently proposed to describe breathiness and tension in the voice [Mokhtari and Campbell, 2003, Degottex et al., 2010]

articulation degree : refers to the phonetic quality of a phoneme, i.e. the extent to which a given target sound is realized (hypo and hyper articulation) [Lindblom, 1983]. The degree of articulation results from the combination of phonetic context (co-articulation), speech rate, and spectral dynamic (i.e. the velocity of the articulatory movements in the vocal tract). Methods have been proposed for the measurement and modification of the degree of articulation based on the analysis of formant trajectories and spectral dynamics [Wouters and Macon, 2001, Beller et al., 2008]. The degree of articulation actually reflects the articulatory effort produced by a speaker, and recent studies support the the degree of articulation as an additional prosodic dimension [Fougeron, 1998, Pfitzinger, 2006].

In this study, the conventional speech prosody parameters (f_0 variations and syllable duration) will be considered and modelled. Nevertheless, the other prosodic dimensions would be required to provide a high-quality modelling of speech prosody.

3.2 Stylization of Speech Prosody

Speech prosody organizes into prosodic contours that occur on specific temporal domains and which convey specific information. The principle of stylization is to decompose the observed speech prosody into contours that are relevant for the description of speech prosody, and residual variations. Relevant variations refer to the long-term variations that are used to convey prosodic information, i.e. to organize the meaning being conveyed. Residual variations are associated with short-term variations related to intrinsic and co-intrinsic phonetic variations [Di Cristo, 2004]. For this purpose, stylization methods provide a decomposition of prosodic variations into a limited set of relevant elementary contours. Stylization methods generally prerequisite a segmentation into temporal domains over which relevant prosodic contours will be described. Stylization methods are all invertible, i.e. the prosodic variations can be synthesized from the sequence of stylized contours. In particular, parametric stylization methods can be efficiently used in speech prosody modelling and speech synthesis.

A variety of methods have been proposed for the stylization of speech prosody - almost exclusively for the description of the f_0 variations - which can be divided with regard to the method and the temporal domains used for stylization of speech prosody. Historically, studies on the stylization of f_0 contours oppose sequential and hierarchical decomposition methods. In the first class of methods, the f_0 variations are decomposed into a sequence of linear segments, either based on signal analysis and joint segmentation-stylization method [Kloker, 1976] or on a perception ['t Hart et al., 1990] model. In parallel, a method for the decomposition of f_0 variations into stress and phrasal components was proposed based on a physiological model of intonation production [Fujisaki, 1981].

Stylization methods can be divided into: *production*-motivated models [Fujisaki, 1981], *perception*-motivated models ['t Hart et al., 1990, House, 1990, Beaugendre, 1992, d'Alessandro and Mertens, 1995], *linguistically*-motivated models [Hirst and Espesser, 1993], and *signal* models [Kloker, 1976, Grabe et al., 1994, Taylor, 1994, Taylor, 2000, Mishra et al., 2006, Lolive et al., 2006, Teutenberg et al., 2008].

A large number of signal methods have been proposed to gradually refine the description of speech prosody, and increase the number of temporal domains considered for stylization. Firstly, methods have been proposed for the decomposition of prosodic contours into linear segments[Kloker, 1976, 't Hart et al., 1990, d'Alessandro and Mertens, 1995], parabolic segments ['t Hart, 1991], and various decomposition bases [Grabe et al., 1994, Mishra et al., 2006, Lolive et al., 2006, Teutenberg et al., 2008]. Secondly, methods have been proposed for the representation of f_0 variations over various temporal domains [Grabe et al., 1994].

In the following, stylization methods are shortly presented and discussed in the case of f_0 variations.

3.2.1 ProsoGram

Historically, stylization of speech prosody based on the perception of intonational variations has been initiated at the IPO (Institute for Perception Research) and extended at the KUL (Katholieke Universiteit Leuven) into the ProsoGram [Mertens, 2004a].

A variety of stylization methods have been proposed in which f_0 variations are stylized in order to account for the actual perceived variations. Thus, perception models have been elaborated to model accurately the perception of f_0 variations. In [d'Alessandro and Mertens, 1995], the perception model is based on three assumptions about the perception of f_0 variations: the integration of pitch perception over time (WTA) [d'Alessandro and Castellengo, 1994], the perception of pitch changes (glissando) ['t Hart et al., 1990], and the differential perception of pitch change (differential glissando) ['t Hart et al., 1990, d'Alessandro and Mertens, 1995]. In particular, perception models can be used to describe compound tone segments, i.e. segments which are composed of elementary tones.

The perception of a change in f_0 over time is locally averaged according to a short-term integration process;

The threshold of pitch change is the minimum difference in f_0 necessary to perceive a change in pitch. The semi-tone per second ratio (ST/s) was proposed as the optimal unit to measure the glissando threshold;

The differential threshold of pitch change is the minimum difference in slope necessary to distinguish between two successive tonal segments.

Perception-based stylization methods generally require a segmentation into minimal temporal domains, such as vowel onset or syllable segmentation [Hermes, 1987], prior to the stylization.

From the speech segmentation and the perception hypotheses, the stylization of f_0 contours is achieved in a recursive manner. First, f_0 variations are integrated over time according to the WTA filter in a pre-processing step. Then, a recursive method is used to decompose f_0 variations into a sequence of tonal segments based on glissando and differential glissando thresholds.

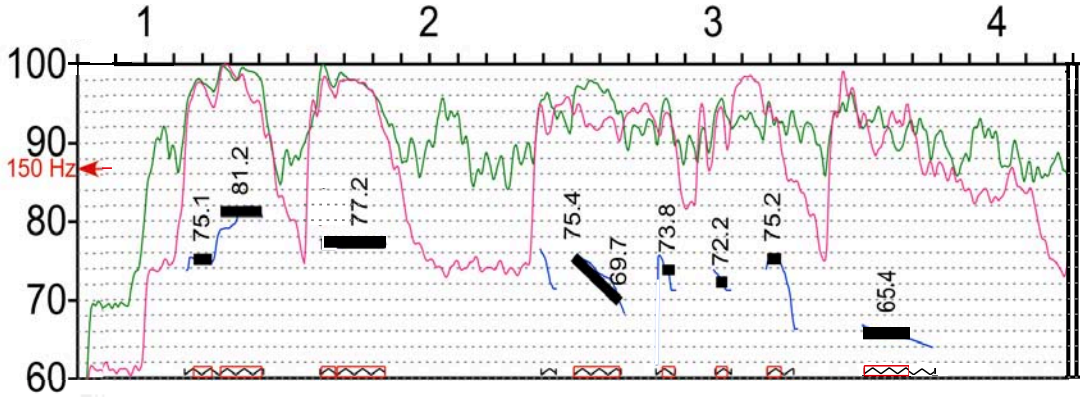


Figure 3.1: Illustration of the PROSOGRAM stylization for the utterance: "*Longtemps, je me suis couché de bonne heure.*" ("*For a long time I used to go to bed early.*"). Stylized f_0 contours are represented with bold lines (ST). The blue curves represent f_0 variations (ST), the green curve intensity variations (dB), and the magenta curve loudness variations (dB). On the bottom, black triangled lines indicate voiced regions and red rectangles indicate regions used for stylization.

3.2.2 MoMel

The MoMel *(Modelling Melody)* stylization method has been developed at the IPA (*Institut Phonétique d'Aix-en-Provence*) [Hirst and Espesser, 1993, Hirst et al., 2000].

Contrary to most of the conventional stylization methods, the anchor of prosodic variations is formulated in terms of *prosodic targets* rather than prosodic segments. In particular, no prosodic segmentation is required prior to the stylization. Thus, the stylized contour accounts for variations occurring over variable temporal domains that are defined by the prosodic targets solely.

The principle of the stylization is to estimate the sequence of prosodic targets, then to stylize the f_0 variations given the prosodic targets. Prosodic targets are defined as the salient inflections of the f_0 variations, which are estimated using a local parabolic curve fitting method. Then, f_0 variations are stylized using a quadratic spline function given the set of prosodic targets and a set of knots heuristically defined[1]. The quadratic spline function is determined under the additional constraint that the function is zero-derivative at each prosodic target. The set of knots is simply set to the center of the successive prosodic targets.

3.2.3 TILT

The TILT model has been developed at the University of Edinburgh in the context of the FESTIVAL speech synthesis system [Taylor, 1998].

In the TILT model, *intonational events* are considered as the minimal temporal domain for the description of intonation. Firstly, intonational events have to be identified from the observed f_0 variations. Then, each intonational event is decomposed with respect to the TILT parameters: amplitude A, duration D, tilt T, f_0 position A_0, and time position t_0. The f_0 position A_0 is the f_0 value at the center of the event, and the time position t_0 is the time of the beginning of the event.

[1]The decomposition of prosodic variations using a basis of spline functions is presented in more details in the following.

The TILT parameters are directly derived from the RFC (Rise/Fall/Connection) model [Taylor, 1994] in which an intonational event E is decomposed into a rising part E_{rise} and a falling part E_{fall}, that are described in term of f_0 excursion and duration $E_{rise} = (A_{rise}, D_{rise})$ and $E_{fall} = (A_{fall}, D_{fall})$.

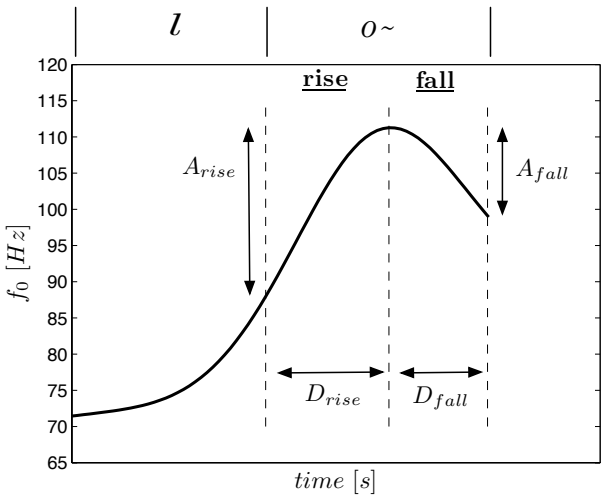


Figure 3.2: RFC decomposition of a melodic contour on the first syllable of the utterance: "*Longtemps, je me suis couché de bonne heure.*" ("*For a long time I used to go to bed early.*"). .

The TILT parameters are simply estimated as follows:

$$A = |A_{rise}| + |A_{fall}| \tag{3.1}$$

$$D = D_{rise} + D_{fall} \tag{3.2}$$

$$T = \frac{T_A + T_D}{2} \tag{3.3}$$

where T_A and T_D are the amplitude and duration tilts, respectively:

$$T_A = \frac{|A_{rise}| - |A_{fall}|}{|A_{rise}| + |A_{fall}|} \tag{3.4}$$

$$T_D = \frac{|D_{rise}| - |D_{fall}|}{|D_{rise}| + |D_{fall}|} \tag{3.5}$$

The f_0 contour of an intonational event is obtained by the inverse transform of the TILT parameters:

$$f_0(t) = \begin{cases} \mathrm{A}_{abs} + A - 2A\left(\frac{t}{D}\right)^2 & 0 \leq t \leq D/2 \\ \mathrm{A}_{abs} + 2A\left(\frac{1-t}{D}\right)^2 & \mathrm{D}/2 \leq t \leq D \end{cases} \tag{3.6}$$

3.2.4 Parametric Decomposition of Prosodic Contours

The principle of parametric decomposition of prosodic contours is to decompose speech prosody on a limited basis of adequately selected elementary contours. In particular, the decomposition aims at distinguishing the macro and the micro prosodic variations, thus the basis is usually chosen as a set of slowly time-varying functions. Various bases have been proposed for the decomposition, including polynomial, cosine, and spline decompositions.

Formally, the principle of parametric decomposition is to decompose a signal sequence $\mathbf{y} = [y_1, \ldots, y_T]$ with real values at discrete times $\mathbf{x} = [x_1, \ldots, x_T]$ on a basis of elementary functions $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_K)$.

Formally, the optimal approximation of a signal sequence $\mathbf{y}$ at discrete times $\mathbf{x}$ on the basis $\boldsymbol{\phi}$ in the Mean Square Error (MSE) sense is given by:

$$y_t = \sum_{k=1}^{K} c_k \phi_k(x_t) \qquad k \in [1, K] \tag{3.7}$$

where c_k is the projection of the finite signal $\mathbf{y}$ on the elementary function ϕ_k:

$$c_k = \langle \mathbf{y}, \phi_k \rangle \qquad k \in [1, K] \tag{3.8}$$

This can be simply formulated in a matrix form:

$$\mathbf{c} = \left(\mathbf{\Phi}^\top \mathbf{\Phi}\right)^{-1} \mathbf{\Phi}^\top \mathbf{y} \tag{3.9}$$

where:

$$\mathbf{\Phi} = \begin{bmatrix} \phi_1(x_1) & \phi_2(x_1) & \ldots \phi_K(x_1) \\ \phi_1(x_2) & \phi_2(x_2) & \ldots \phi_K(x_2) \\ \vdots & \vdots & \vdots \\ \phi_1(x_T) & \phi_2(x_T) & \ldots \phi_K(x_T) \end{bmatrix} \tag{3.10}$$

3.2.4.1 Polynomial Transform

The polynomial transform consists of the decomposition of a contour $\mathbf{y} = [y_1, \ldots, y_T]$ at discrete times $\mathbf{x} = [x_1, \ldots, x_T]$ on a basis of polynomial functions $\boldsymbol{\phi} = (P_1, \ldots, P_K)$, where P_k is a polynomial function.

The popularity of the polynomial decomposition is due to the easy interpretation of each polynomial component, from linear [Mishra et al., 2006] and quadratic decompositions [Taylor, 1994, Taylor, 2000] to higher-order decompositions [Grabe et al., 1994]. Among the variety of possible decompositions, the Legendre decomposition is one of the most popular polynomial decomposition method used for the description prosodic contours [Grabe et al., 1994]. In particular, the Legendre basis is orthogonal so that the elements of the decomposition are decorrelated.

$$\langle \phi_i, \phi_j \rangle = \begin{cases} 1, & \text{i} = \text{j} \\ 0, & \text{i} \neq j \end{cases} \tag{3.11}$$

This property ensures the non-redundancy of the information carried out by each elementary function of the basis.

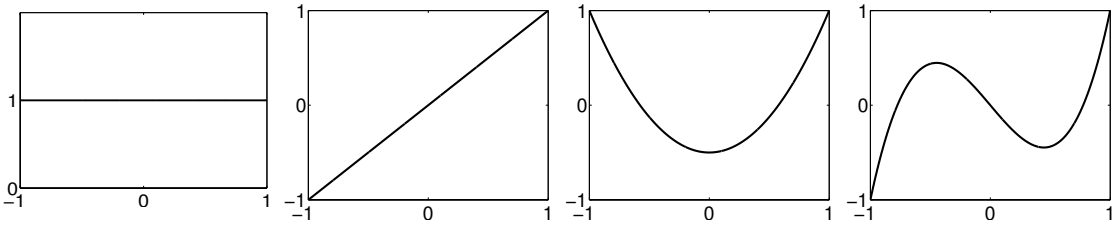


Figure 3.3: Fourth-order elementary Legendre polynomials functions.

3.2.4.2 Discrete Cosine Transform (DCT)

More recently, the Discrete Cosine Transform has been proposed for the stylization of speech prosody [Teutenberg et al., 2008].

The Discrete Cosine Transform consists of the decomposition of a contour $\mathbf{y} = [y_1, \ldots, y_T]$ at discrete times $\mathbf{x} = [x_1, \ldots, x_T]$ on a basis of zero-phase cosine functions $\boldsymbol{\phi} = (\cos(\omega_1), \ldots, \cos(\omega_T))$ at discrete frequencies $\omega_k = \dfrac{\pi}{2T}(2k+1)$:

$$c_k = \alpha_k \sum_{t=1}^{T} x_t \cos(\omega_k n) \qquad k \in [1, T] \tag{3.12}$$

where

$$\alpha_k = \begin{cases} \sqrt{\dfrac{1}{T}} & k = 1 \\ \sqrt{\dfrac{2}{T}} & k \in [2, T] \end{cases}$$

The Discrete Cosine Transform is an invertible transform with perfect signal reconstruction:

$$x_t = \sum_{k=1}^{T} \alpha_k c_k \cos(\omega_k t) \qquad t \in [1, T] \tag{3.13}$$

In particular, the truncation of the Discrete Cosine Transform $\mathbf{c}$ at order $K \in [1, T]$ (and the associated discrete frequency $\omega_c = \dfrac{\pi}{2T}(2K+1)$) constitutes the optimal approximation of the original signal sequence $\mathbf{x}$ in the Mean Square Error (MSE) sense.

The Discrete Cosine Transform can be efficiently used to decompose speech prosody into macro and micro prosodic variations, then removing the micro prosodic variations by the appropriate selection of the truncation order according to the desired cut-off frequency ω_c.

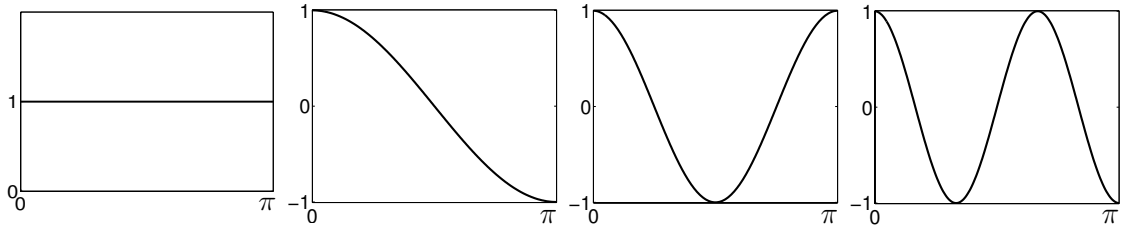
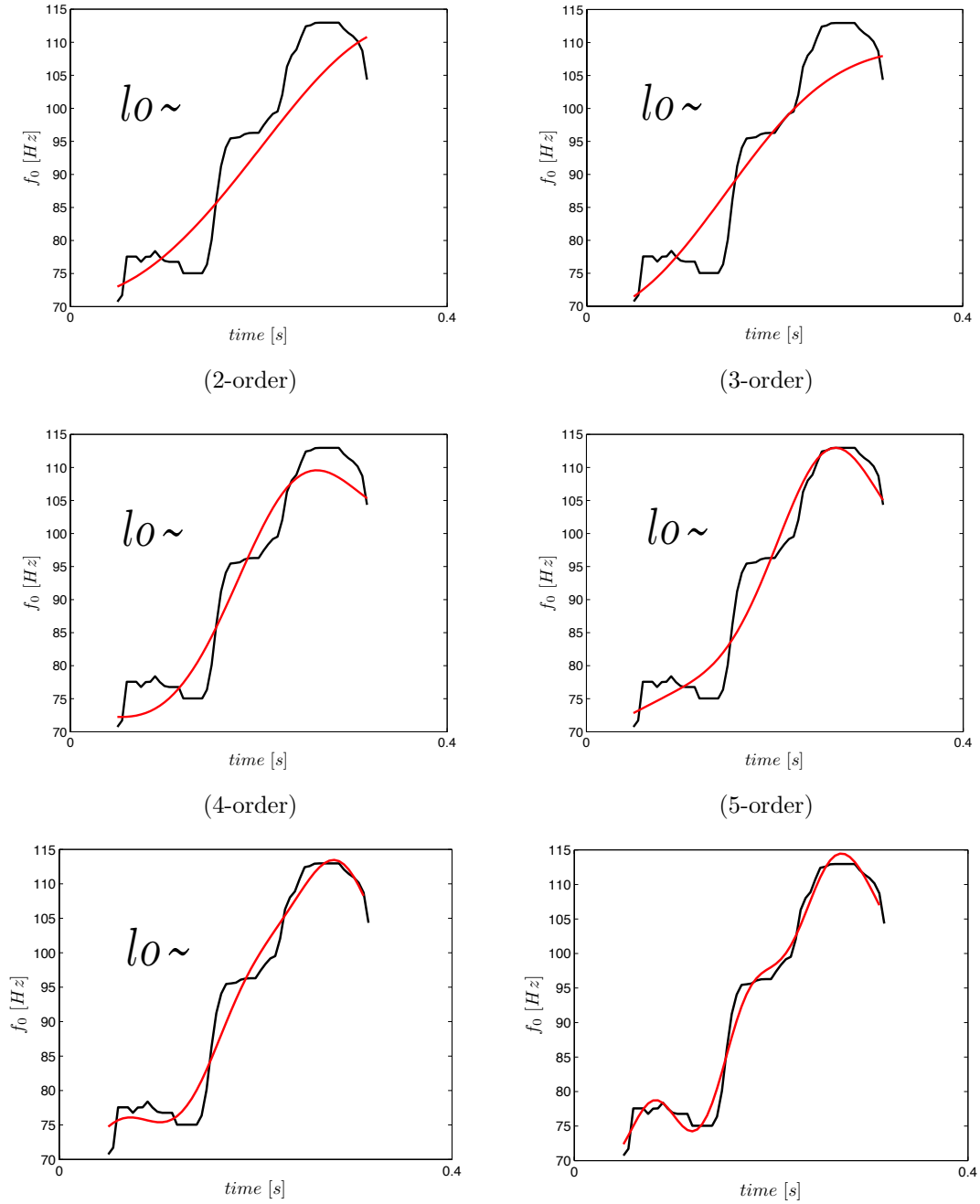


Figure 3.4: Fourth-order elementary cosine functions.



(2-order) (3-order)

(4-order) (5-order)

(7-order) (10-order)

Figure 3.5: Stylization of melodic contours with various order of Discrete Cosine Transform.

3.2.4.3 Spline Transform

While the above transforms decompose a contour on a set of functions that are globally defined on the temporal segment, the spline transform decomposes a contour over a set of overlapping functions that are locally defined in time [de Boor, 1978]. Consequently, each part of the contour can be decomposed over locally time-defined functions [Lolive et al., 2006].

A spline function $S_{k,t}$ of order k is a piecewise polynomial function defined on an interval $\mathbf{x} = [x_1, \ldots, x_N]$ that is partitioned into a sequence of m ordered knots $\mathbf{t} = [t_1, \ldots, t_m]$ ($x_1 = t_1 \leq \cdots \leq t_{k-1} \leq t_m = x_N$) each associated with a polynomial function B_k^i:

$$S_{k,\mathbf{t}} = \left(\mathbf{t} = [t_1, \ldots, t_m],\ \left\{B_k^i\right\}_{i=1}^{m-k-1}\right) \tag{3.14}$$

The spline transform consists of the decomposition of a contour $\mathbf{y} = [y_1, \ldots, y_T]$ at discrete times $\mathbf{x} = [x_1, \ldots, x_T]$ on a basis of of B-spline functions $\boldsymbol{\phi} = (B_k^1, \ldots, B_k^{m-k-1})$, where B_k^i is a k-order B-spline function associated with the $[t_i, t_{i+k+1}[$ interval.

B-spline functions of order k are recursively defined:

$$\begin{aligned} B_0^i(t) &= \begin{cases} 1\ , & \mathrm{t} \in [t_i, t_{i+1}[\\ 0\ , & \text{else} \end{cases} \\ B_k^i(t) &= \frac{t - t_i}{t_{i+k} - t_i} B_{k-1}^i(t) + \frac{t_{i+k+1} - t}{t_{i+k+1} - t_{i+1}} B_{k-1}^{i+1}(t) \end{aligned} \tag{3.15}$$

Each B-spline $B_k^i$of degree k is locally defined on the time interval $[t_i, t_{i+k+1}[$:

$$B_k^i(t) \quad \begin{cases} > 0\ , & \mathrm{t} \in [t_i, t_{i+k+1}[\\ = 0\ , & \text{else} \end{cases} \tag{3.16}$$

so that each segment of a contour is decomposed over locally time-defined polynomial functions.

The B-spline decomposition requires the determination of an appropriate knot sequence $\mathbf{t} = [t_1, \ldots, t_k]$ and the sequence of coefficients $\mathbf{c} = [c_1, \ldots, c_{m-k-1}]$ that are associated with the B-splines.

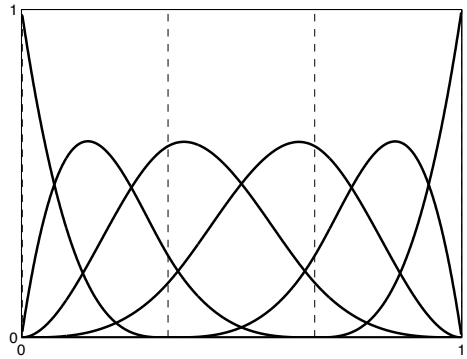


Figure 3.6: Fourth-order elementary B-spline functions with two knots.

3.2.5 Discussion

Stylization of speech prosody presents various advantages for the analysis and the modelling of speech prosody in speech synthesis. However, no study exists on the comparison of existing stylization methods due to the variety of assumptions, objectives, and properties of the respective methods. Additionally, speech prosody analysis and modelling methods substantially differ in their objective: on the one hand, the analysis of speech prosody aims at providing the more accurate

and comprehensive description of the speech prosody variations, objectively or perceptually. On the other hand, the modelling of speech prosody aims at optimizing the accuracy of the statistical modelling, i.e. the naturalness and the variety of the synthesized speech prosody. However, there is no evidence for the correlation of the accuracy of a stylization and the accuracy of the statistical modelling. Thus, a comparison of stylization methods in speech prosody modelling would require to evaluate and to compare the accuracy of the stylization in analysis and synthesis.

In the absence of a comparative study in speech prosody analysis and modelling, a comparison of stylization methods is here shortly discussed. Conventional speech prosody model and speech synthesis methods are based on the statistical description of the acoustic variations on a set of specified temporal domains, and their associated linguistic characteristics. Additionally, statistical modelling methods require a unified description of the acoustic variations, either in terms of the dimensionality or the homogeneity of the acoustic description. However, some of the mentioned stylization methods suffer from a inadequacy for the statistical modelling, such as the Fujisaki, the MoMel, the ProsoGram, and to a lesser extent the TILT stylization methods. Firstly, the Fujisaki and the MoMel stylization methods require to model the precise temporal location of prosodic events (such as impulses or targets) over an utterance. Secondly, the MoMel stylization method describes the prosodic variations on variable and non linguistically-defined temporal domains. Thirdly, the ProsoGram stylization described prosodic contours in terms of simple and compound contours that are described with a variable number of parameters. Finally, the TILT stylization describes prosodic variations only for prosodic events, while other segments are defined with linear interpolation of consecutive prosodic events.

Parametric decomposition methods are generally preferred for the statistical modelling of speech prosody and speech synthesis. An informal study on the stylization of speech prosody revealed no qualitative difference in the accuracy of the stylization. However, the B-spline stylization requires to define an adequate sub-segmentation of the temporal domains on which prosodic contours will be decomposed during the analysis, and to model the temporal location of the sub-segments during the synthesis. Finally, the Discrete Cosine Transform was observed to perform a more adequate decomposition of the prosodic contours than the Polynomial Transform, especially for the stylization of long-term contours - such as the prosodic phrase contour[2]. Consequently, the Discrete Cosine Transform was chosen for the stylization of speech prosody in this study.

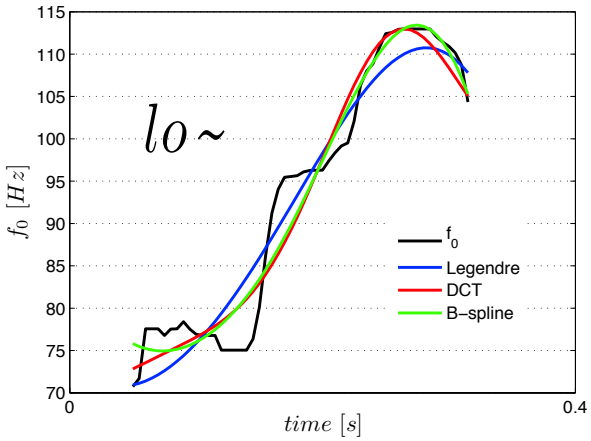


Figure 3.7: Comparison of parametric stylization methods on the voiced region of a syllable. The black line represents the observed f_0 variations, the blue line a 3-order Legendre polynomial decomposition, the red line a 5-order Discrete Cosine Transform, the green line a 4-order B-spline decomposition with a knot situated at the center of the voiced region.

[2]Naturally, this observation is informal and remains to be evaluated.

3.3 Transcription of Speech Prosody

Phonology is the study of the phonetic system of a language, whose objective is to determine the phonetic primitives of the language. Primitives are defined as a finite set of minimal and distinctive units that can be used for the description of the phonetic variations of a language. The primitives constitute the alphabet of the phonetic system, which are generally associated with a grammar that can be used to derive the prosodic structure of any utterance of the language. The phonological representation is the first level of abstraction in the description of speech prosody, in which the continuous acoustic variations are converted into a discrete sequence of contrastive symbols.

Historically, the core contribution to the study of intonational phonology originates from the Massachusetts Institute of Technology (MIT) for the study of English and American-English [Liberman, 1975, Pierrehumbert, 1980]. In particular, the phonological theory proposed in [Pierrehumbert, 1980] results in the development of the ToBI (Tone and Break Indices) standard for the transcription of English prosody [Silverman et al., 1992, Beckman and Ayers, 1997]. This theory has strongly influenced the description of European [Ladd, 1983, Gussenhoven, 1984] and Asiatic [Jun, 1993] languages, and the ToBI annotation was extended and adapted for the description of a large variety of languages [Venditti, 1995, Beckman and Jun, 1996, Baumann et al., 2000, Ladd, 1996, Gussenhoven, 2004, Jun, 2005]. Additionally, ToBI is the most widely representation used for the description of prosodic variations in speech prosody modelling and speech synthesis systems [Black and Taylor, 1997b, Zen et al., 2007, Schröder and Trouvain, 2003].

Nevertheless, a number of studies have pointed out some inconsistency of the ToBI standard. In terms of transcription feasibility and reliability, the ToBI transcription standard suffers from major drawbacks [Wightman, 2002]:

language-dependency : ToBI requires the complete inventory of the intonational system of a language prior to the transcription.

expertise : ToBI requires experts or highly trained individuals;

time : ToBI manual transcription is extremely time-consuming (about 100 to 200 times real time [Syrdal et al., 2001]);

reliability : inter-transcriber agreement is relatively low while using the complete ToBI alphabet [Syrdal and McGory, 2000].

Other studies have questioned the theoretical limits of the ToBI standard:

prosodic substance : ToBI is a standard for the transcription of intonation, solely. Consequently, none of the other prosodic dimensions are used in the transcription, such as local speech rate, intensity, voice quality, or articulation degree. In particular, prosodic events in the absence of a pitch event are commonly observed in many languages, while described with the conventional ToBI system;

prosodic domain : ToBI assumes the syllable as the minimal anchor for the description of speech prosody;

prosodic structure : the ToBI transcription is explicitly sequential while the prosodic structure is rather hierarchical;

prosodic genericity : ToBI was originally designed for the transcription of non-elicited speech only, and does not account for many prosodic phenomena occurring in spontaneous speech, such as hesitations and reformulations;

Various alternatives to the ToBI standard have been proposed for the description and the transcription of French prosody [Campione et al., 2000, Post et al., 2006, Avanzi et al., 2007, Lacheret et al., 2010] (see [Delais-Roussarie et al., 2006] for a review in French). However, the

transcription of French prosody remains under debate, and no standard exists. Among the different transcription systems, ToBI is the only that provides a phonological representation *stricto sensu*, while the others provide a surface description of speech prosody without the explicit inventory of a system. Transcription systems divide into acoustic, perceptual, and functional methods, depending on the strategy adopted for the description of speech prosody. Acoustic methods are based on the description of the acoustic variations solely (ToBI, INTSINT). Perceptual methods are based on the perception of prosodic events and prosodic prominences (Rhapsodie). Functional methods account for top-down processes and the integration of higher-levels of linguistic processing (lexical, syntactic, semantic, discursive) [Wagner, 2005]. While transcription methods assume that the description of speech prosody emerges from a bottom-up integration (from the acoustic description), most of them remain functional since the acoustic description and the linguistic processing remain hardly distinguishable[3].

In this section, methods for the transcription of French prosody are shortly described and discussed. In particular, a transcription standard [Lacheret et al., 2010] recently developed in the Rhapsodie Project[4] for the transcription of French prosody is presented, that will be used in this thesis for the description of speech prosody.

3.3.1 ToBI

The ToBI representation describes the melodic variations in terms of *pitch events* and *intonational breaks*. The intonational units used for the description are: syllable, intonational form, intermediate intonational phrase, and intonational phrase. Intonational breaks describe the degree of junctures among successive forms using a continuous break-index scale. Pitch events divide into pitch accents and phrasal tones: pitch accents are pitch events associated with accented syllables, and phrasal tones are pitch accents associated with intonational boundaries. Phrasal tones are further distinguished into phrase accents (intermediate phrase boundary) and boundary tones (full intonation phrase boundary). The description of tones is based on an elementary alphabet composed of tones (Low and High tones) that is used to derive complex tone structures. The transcription of pitch events of a language requires the inventory of contrastive tones that are observed in this language. Additional symbols have been further introduced to manage uncertainty, underspecification, and spontaneous speech.

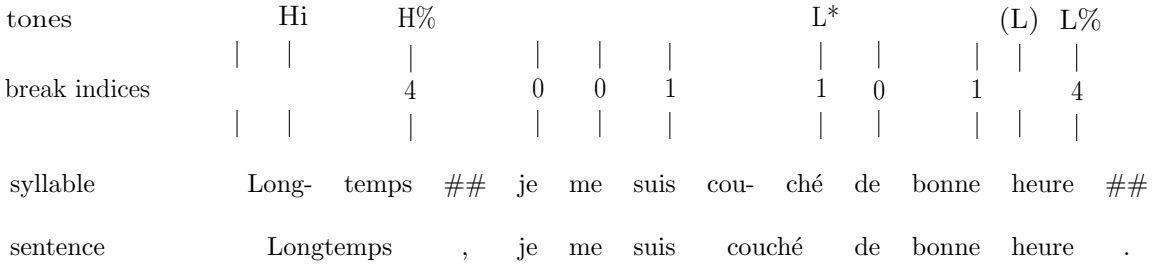


Figure 3.8: Illustration of the ToBI transcription for the utterance: "*Longtemps, je me suis couché de bonne heure.*" ("*For a long time I used to go to bed early.*")

3.3.2 INTSINT

The INTSINT (*INternational Transcription System for INTonation*) is a phonological intonation system that has been developed at the IPA (*Institut Phonétique d'Aix-en-Provence*) [Hirst and Di Cristo, 1998, Campione et al., 2000]. This INTSINT representation is directly

[3] with the exception of the INTSINT method which is strictly descriptive.
[4] *Rhapsodie: Reference Prosody Corpus of Spoken French.*

derived from the MoMel stylization in which the INTSINT representation is the phonological side of the Momel phonetic description.

The INTSINT representation is composed of a set of 9 symbols that are used to describe intonational variations over and across intonational units. The intonational unit is defined as the maximal unit for the description of speech prosody, which is usually equated to the inter-pausal unit for convenience. The INTSINT representation accounts for two levels of intonational variations: absolute and relative variations. The absolute description accounts for the intonation range within an intonational unit with respect to the register of the speaker (Top, Middle, Bottom). The relative description accounts for the intonational temporal structure, i.e. the relative position of a tone with respect to that of the previous one (Higher, Same, Lower, and Upstepped, Downstepped).

3.3.3 IVTS

The IVTS (*Intonation Variations Transcription System*) is a phonological intonation transcription system that has been developed at the LLF (*Laboratoire de Linguistique Formelle*, Université Paris Diderot) in the PFC Project (Phonologie du Français Contemporain) [Post et al., 2006]. The IVTS representation is directly derived from the IViE transcription system for English [Grabe et al., 2001] and studies on phonology in English, German, and French [Grabe, 1998, Post, 2000].

The IVTS transcription decomposes speech prosody into fours levels of representation, from the perception of prosodic prominences to the phonological representation: *prominence*, *local phonetic variations*, *global phonetic variations*, and the *phonological representation*. The minimal temporal domain used for the description is the syllable, and the maximal intonational unit is the intonational phrase. The transcription of prosodic events is based on the perception of an acoustic prominence (P) that occurs on a syllable, including but not restricted to pitch events. The intermediate levels describe the intonational local and global variations based on the perception of the intonational variations. The local description accounts for the perceived intonational variations (low, middle, high) relative the intonational domain. The intonational domain is defined as the speech segment which is left and right bounded by a prosodic prominence. The description is capitalized when an intonational target is aligned with a prosodic prominence. The global description accounts for the perceived intonational variations that occur on a temporal domain that is larger than the intonation domain (Reset, Downstep). Finally, the phonological representation is derived from the ToBI system and describes the tone structure of pitch accents and intonational boundaries depending on the tone inventory of the language.

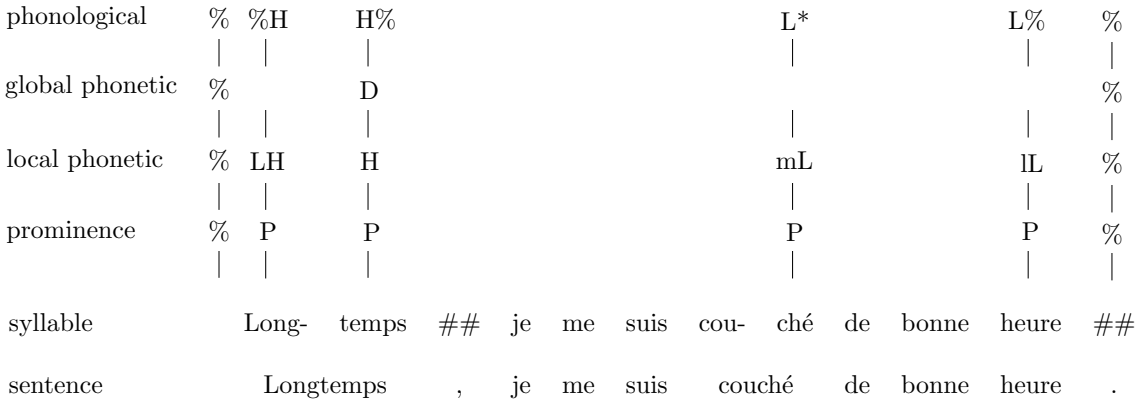


Figure 3.9: Illustration of the IVTS transcription for the utterance: "*Longtemps, je me suis couché de bonne heure.*" ("*For a long time I used to go to bed early.*")

3.3.4 Rhapsodie

The RHAPSODIE transcription system is a standard that has been developed in the Rhapsodie Project (Rhapsodie: Reference Prosody Corpus of Spoken French) [Lacheret et al., 2010] for the transcription of French prosody.

The RHAPSODIE transcription system intends at providing a simple and unified transcription ground that can be shared among the existing phonological theories and description systems. The description of the prosodic variations is based on the perception of prosodic events that are implicitly shared among the phonological theories, such as *prosodic prominence* and *prosodic grouping*. Prosodic prominence is defined as an acoustic saliency, and covers prosodic events that are marked by intonation or by any other acoustic cue. The perceptual description of prosodic variations presents various advantages over more sophisticated systems. Firstly, a perceptual description does not require expert knowledge, and can be processed by moderately trained individuals. Secondly, the transcription can be easily integrated into most of the existing models for further phonetic and phonological descriptions. In particular, the perceptual level provides a minimal description of prosodic events that can be used to precise and describe the acoustic dimensions that may be phonetically and phonologically relevant.

The minimal prosodic unit used for the description is the syllable, and the maximal prosodic unit is the prosodic period [Avanzi et al., 2008]. The transcription of prosodic events is based on the perception of prosodic prominences (P) and prosodic grouping (minor and major prosodic boundaries). The transcription is processed recursively to account for the hierarchical organization of prosodic events. For this purpose, a variable temporal resolution is used to manage the relative perception of prosodic prominences and to refine gradually the prosodic description. Firstly, segmentation into major prosodic groups (MPGs) is achieved within a large integration domain (typically 5-10 s. depending on the speaking style). The segmentation is based on the perception of a major prosodic prominence that is associated with the end of a major prosodic group. Secondly, segmentation into minor prosodic groups (mPGs) is achieved within each prosodic group. The segmentation is based on the perception of a minor prosodic prominence that are associated with the end of a minor prosodic group. Finally, residual prosodic prominences (P) are identified as the remaining prosodic prominences that occur within the minor prosodic group. Additional symbols are used to manage uncertainty and underspecification on the presence and nature of a prosodic boundary, and on the presence of a prosodic prominence. Speech disfluencies are transcribed in parallel to the prosodic transcription on a separate tier[5].

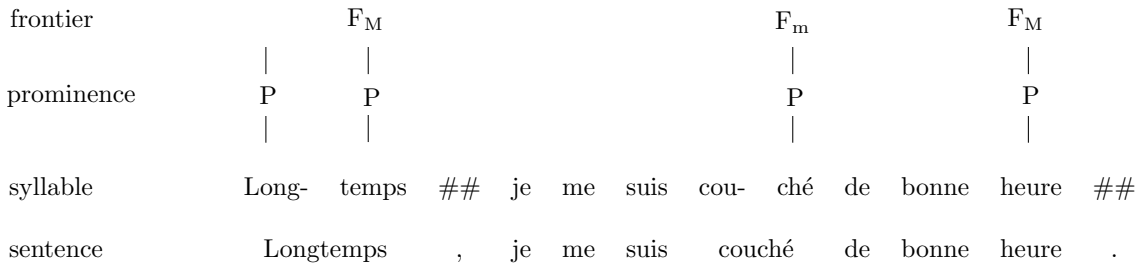


Figure 3.10: Illustration of the RHAPSODIE transcription for the utterance: "*Longtemps, je me suis couché de bonne heure.*" ("*For a long time I used to go to bed early.*")

[5]prosodic prominences and prosodic disfluencies are assumed to be independent prosodic phenomena: a disfluency can be prosodically marked with a prominence or not.

Chapter 4

Modelling of Speech Prosody: State-of-the-Art

Contents

4.1 Introduction

Speech prosody is a core module in a speech synthesis system. Alongside the development of speech synthesis systems, a large number of studies on the analysis and modelling of speech prosody have been developed to improve the intelligibility and the naturalness of speech synthesis systems, from the design of early expert models [Aubergé, 1991] to the development of statistical models [Yoshimura et al., 1999, Tokuda et al., 2003]. Speech prosody has gradually emerged as a central concern in speech synthesis to improve the variety and the liveliness of speech synthesis systems [Bulyko and Ostendorf, 2001, Toda and Tokuda, 2007, Yan et al., 2009], and to control and adapt the speaking style of a speaker in speech synthesis [Yamagishi et al., 2004, Tachibana et al., 2005, Yamagishi, 2007].

The principle of speech prosody systems is to model the speech prosody characteristics of a speaker or a speaking style, and to synthesize the sequence of speech prosody variations given an input text, and eventually para-linguistic (e.g., emotional state) or extra-linguistic information (e.g., specific situation of a communication). In fact, the synthesis of prosodic parameters is the inverse problem to that of speech prosody analysis: while speech prosody

analysis is a bottom-up process in which a symbolic representation is described from the signal variations, the synthesis of speech prosody is a top-down process in which the signal variations are synthesized from a symbolic description. A large number of methods have been proposed for the symbolic [Veilleux et al., 1990, Ostendorf and Veilleux, 1994, Ross and Ostendorf, 1996, Black and Taylor, 1997a, Schmid and Atterer, 2004] and acoustic [Yoshimura et al., 1999, Dusterhoff et al., 1999, Toda and Tokuda, 2007, Gao et al., 2008, Qian et al., 2009, Yan et al., 2009] statistical modelling of speech prosody, and the modelling and adaptation of speaking style [Bell et al., 2006, Yamagishi, 2007].

In this chapter, the state-of-the-art on the modelling of speech prosody is briefly presented. The architecture of a speech prosody system is presented in section 4.2. The principle of context-dependent analysis is presented in section 4.3. State-of-the-art on the linguistic description and the analysis of speech prosody are presented in sections 4.4 and 4.6. Finally, the state-of-the-art on the symbolic and acoustic modelling of speech prosody are briefly presented in sections 4.7 and 4.8 and will be discussed in more details in chapters 8 and 9 devoted to the discrete and continuous modelling of speech prosody.

4.2 Architecture of a Speech Prosody System

The architecture of a speech prosody system is generally decomposed into separate modules that each models a level of the prosodic variations, in particular by distinguishing the *symbolic* and the *acoustic* variations. The macro and the micro prosodic variations are generally integrated into a single acoustic module, either by modelling the macro-prosodic variations only, or by modelling simultaneously the micro and macro prosodic variations. Thus, a speech prosody system is composed of:

a **symbolic** module: in which the symbolic characteristics of speech prosody are modelled conditionally to their context,

an **acoustic** module: in which the acoustic characteristics of speech prosody are modelled conditionally to their context.

Each module is either based on expert knowledge or statistical modelling, and more commonly on a combination of expert and statistical modelling.

During the synthesis, the text is analyzed and a set of para-linguistic and extra-linguistic information are eventually automatically extracted or manually described. Then, the corresponding speech prosody variations are determined in a top-down process, from the symbolic to the acoustic variations. First, the sequence of symbolic parameters is determined given the linguistic and the additional information. Then, the sequence of acoustic parameters is inferred given the extracted information and the symbolic sequence of prosodic events.

The architecture of a HMM-based speech prosody system is illustrated in figure 4.4.

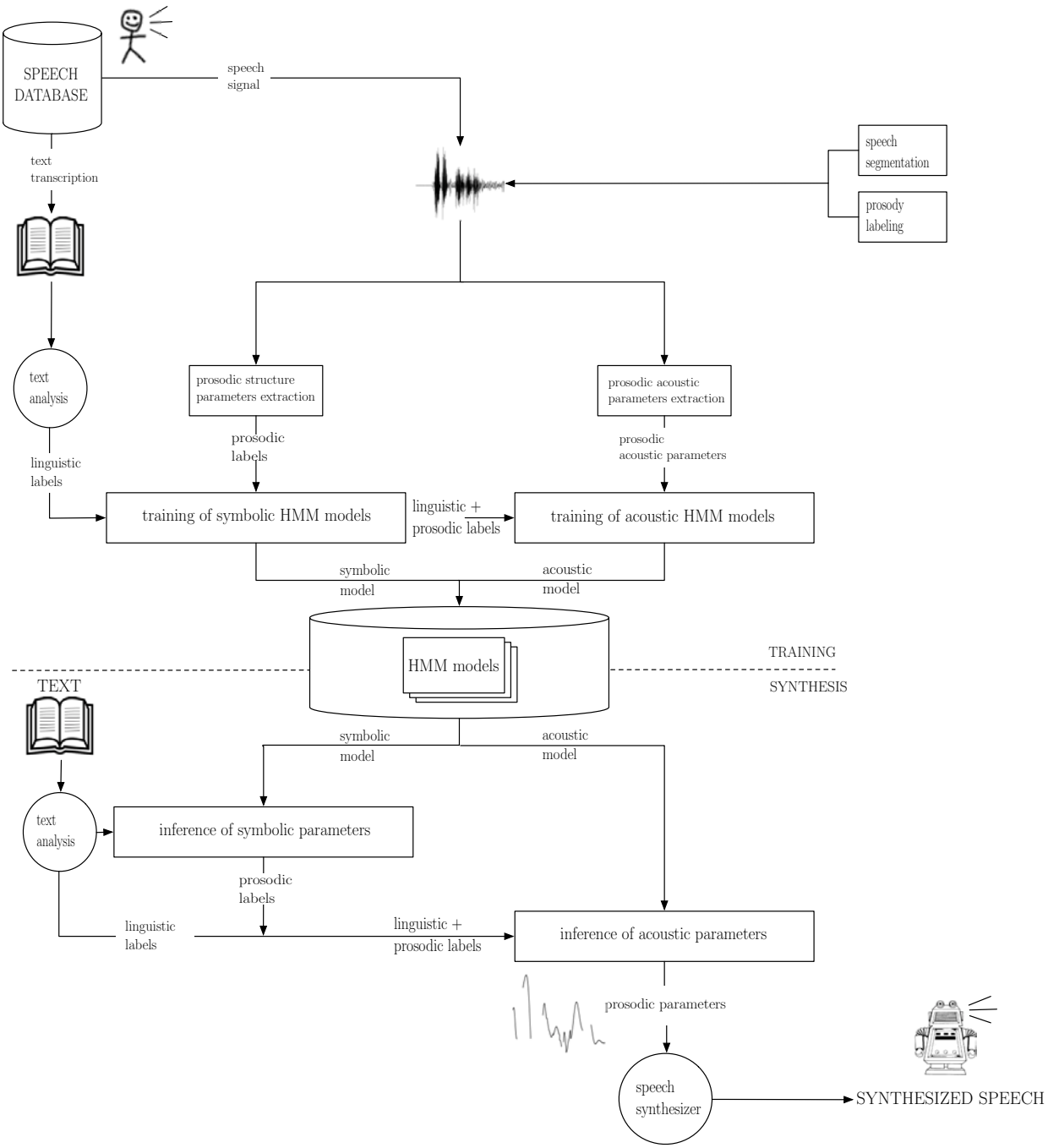


Figure 4.1: Architecture of a HMM-based speech prosody system for speech synthesis.

4.3 Modelling Speech Prosody in Context

As mentioned in section 2, speech prosody is essentially *variation*, and a large number of linguistic, para-linguistic, and extra-linguistic factors are involved and interacts to produce variations. In consequence, a speech prosody system primarily consists of the identification of relevant sources of speech prosody variations, then to adequately model these variations. Thus, context-dependent analysis is a core requirement for the analysis and modelling of speech prosody. The principle of context-dependent analysis is precisely to model the variations of speech prosody *in context*, i.e. depending on the context in which the variation is observed. Context-dependent analysis was originally introduced in speech recognition to model adequately the acoustic variations associated with co-articulation [O'Dell, 1995], and later extended to speech synthesis and speech prosody [Yoshimura et al., 1999]. In speech prosody modelling, the type and the complexity of the contexts to be considered has considerably increased due to the number of temporal domains that is to be considered (syntactic / prosodic units), and their hierarchical organization (syntactic structure / prosodic structure).

A *context* is a generic term that describes a symbolic information that is associated with a temporal segment on which a variation is observed. More formally, a context is associated with a fixed *target* unit (e.g., phoneme, syllable) over which speech prosody variations are observed, and a symbolic information that is inherited from the context of this unit. A context may refers to the linguistic structure of a text (e.g., sequence of phonemes, sequence of syllables, prosodic structure, syntactic structure, discursive structure), para-linguistic information (e.g., emotional state of the speaker, intention of the speaker, pragmatic information), and extra-linguistic information (e.g., idiolectal, geographical, sociological characteristics of the speaker, and the situation in which the speech is produced). Various types of contextual information exist in conventional speech synthesis and speech prosody systems:

the *immediate* context : describes the characteristics that are strictly associated with the target unit (e.g., phoneme label, prominent status of a syllable, syllable type, syllable structure),

the *linear* context : describes the characteristics that are associated with the sequential context of a given unit (e.g., characteristics of the preceding and succeeding syllable),

the *hierarchical* context : describes the characteristics of a higher-level unit that are inherited from a hierarchical representation of various units (e.g., lexical category of a form, morpho-syntactic category of a chunk in which a syllable is included),

the *global* context : describes the characteristics that are globally associated with an utterance (e.g., emotional state, speaking style).

Additionally, the order of a context dependency has to be defined depending on the dependency that is expected to be relevant for the description of speech prosody, and the contextual information can be combined (e.g., position of a syllable within a prosodic group, position of a prosodic group within an utterance).

However, context-dependent analysis suffers from various problems for the modelling of speech prosody. First, most of the contexts cannot be automatically extracted and require a manual labelling. Secondly, the large number of potentially relevant contexts is computationally unreachable and would require huge labelled speech databases that are currently not available. Thirdly, a single model cannot accurately model simultaneously all of the contexts. Thus, it is generally assumed to model the observed variations with separated models, each modelling a part of the variations: the linguistic, para-linguistic, and extra-linguistic variations are generally modelled separately. Finally, speech prosody modelling suffers from a double problem for context-dependent analysis: speech prosody modelling requires to define the temporal domains on which relevant variations are observed, and to determine the contexts that are relevant to explain the observed variations over each temporal domain.

4.4 Text Analysis & Linguistic Contexts

Among the source of variations, the linguistic variations are the most widely used. First, a number of theoretical linguistic studies proved strong evidence for the dependency of the speech prosody variations and the syntactic structure of a sentence. More recently, studies described speech prosody over larger temporal domains, such as macro-syntactic and discursive structures. Second, studies in computational linguistic provided algorithms that can be used for the automatic extract of linguistic information from a text: from the phonemes and syllable conversion, to surface and deep syntactic parsing. Unfortunately, no method exists for the automatic extraction of linguistic information to describe the linguistic characteristics of a text over larger domains. Consequently, text analysis is generally limited to the sentence domain, and only speech prosody variations occurring within a sentence are modelled.

Most of the conventional speech synthesis and speech prosody systems are based on elementary text analysis and remain limited to a surface syntactic parsing, only: the text is segmented into forms that are associated with a morpho-syntactic category (Part-of-Speech or POS). Eventually, the surface analysis provides segmentation into chunks associated with a syntactic category. Expert and deep syntactic parsers exist that provide a rich description of the text structure, but they are currently not used in speech synthesis and speech prosody systems.

FORM	POS	CHUNK
longtemps	adverb	AdvP
,	punctuation	-
je	nominative clitic	VP
me	reflexive clitic	
suis	auxiliary verb	
couché	verb	
de	preposition	NP
bonne	adjective	
heure	common noun	
.	final punctuation	-

Table 4.1: Description of surface syntactic information used in conventional speech synthesis and speech prosody systems for the sentence: "*Longtemps, je me suis couché de bonne heure.*" ("*For a long time I used to go to bed early.*"). Form, form segmentation, Part-of-Speech, chunk segmentation, chunk category. AdvP, VP, and NP respectively denote adverbial, verbal, and nominal phrase.

4.5 Prosodic Analysis & Prosodic Contexts

The extraction of prosodic information based on acoustic analysis required the segmentation into prosodic units, the identification and eventually the description of prosodic events.

The prosodic segmentation is usually based on the alignment of the text and the utterance through conventional phonetic segmentation and syllabification. First, the text is aligned with the utterance based on conventional speech segmentation into phonemes. Then, a phoneme-to-syllable conversion system is used to convert the phoneme sequence into syllable sequence, either based on symbolic and/or acoustic analysis. Finally, a symbolic/acoustic analysis is processed so as to identify and/or describe relevant prosodic events, either based on manual or automatic transcription.

In conventional speech prosody systems: phoneme, syllable, and prosodic phrases are used as prosodic units for the description of speech prosody. The prosodic transcription is used for the identification and the description of relevant prosodic events, and prosodic units are used to in-

stantiates relative prosodic contexts (position/number of a prosodic unit within a higher prosodic unit). Finally, 1-order context are conventionally used for the description of the linear context (e.g., prosodic label of the preceding/succeeding syllable, number of syllables within the preceding/succeeding prosodic phrase).

	unit	description
prosodic description	syllable	prosodic label (ToBI, TILT,...)
	unit	**parent unit**
prosodic/syntactic unit	phoneme	syllable form prosodic phrase utterance
	syllable	form prosodic phrase utterance
	form	prosodic phrase utterance
	prosodic phrase	utterance

Table 4.2: Description of prosodic information used in conventional speech synthesis and speech prosody systems.

4.6 Segmental Analysis & Segmental Contexts

The modelling of speech prosody for speech synthesis additionally requires to model the variations the fine micro variations of speech prosody (articulation, co-articulation) due to the segmental context (phoneme, syllable structure).

The description of the segmental context includes phoneme (label of the phoneme, phonological description of the phoneme, and class of the phoneme) and syllable structure. The label of the phoneme depends on the inventory of the phoneme in a language, the phonological description of a phoneme relates to the distinctive characteristics of a phoneme (e.g., bilabial, labiodental, dental, obstruent, sonorant), and the class of a phoneme relates to general classes (liquid, nasal, plosive, fricative, vowel, glide, schwa). The syllable structure refers to the structure of the syllable (onset, nucleus, coda).

4.7 Discrete Modelling of Speech Prosody

The principle of discrete modelling of speech prosody is to determine the sequence of prosodic labels (associated with relevant prosodic events) that correspond to a given text.

Two main approaches can be distinguished for the discrete modelling of speech prosody: on the one hand, *expert approaches* attempt at elaborating formal models that account for the observed prosodic variations with respect to linguistic, para-linguistic, and extra-linguistic constraints. On the other hand, *statistical methods* attempt at elaborating a statistical model which accounts for the prosodic variations from the observation of statistical regularities on large speech databases.

sentence	Longtemps	,	je	me	suis	couché	de	bonne	heure	.

⇓

prosodic structure																						
F_M			*								*					*						
F_m			*								*					*						
P	*		*								*					*						
syllable	Long-		temps		##		je		me		suis		cou-		ché		de		bonne		heure		##

Table 4.3: Illustration of a discrete modelling of speech prosody for the sentence: "*Longtemps, je me suis couché de bonne heure.*" ("*For a long time I used to go to bed early.*").

4.7.1 Expert Models

On the one hand, expert approaches mostly concern the hierarchical organization of speech prosody, and in particular prosodic boundaries ([Cooper and Paccia-Cooper, 1980, Gee and Grosjean, 1983, Selrik, 1984, Ferreira, 1988, Abney, 1992, Watson and Gibson, 2004] for English; [Dell, 1984, Bailly, 1989, Monnin and Grosjean, 1993, Ladd, 1996, Delais-Roussarie, 2000, Mertens, 2004b] for French, [Barbosa, 2006] for some other languages). Expert models assume that a prosodic structure results from the integration of various and potentially conflictual constraints, in particular *syntactic* and *rhythmic* constraints.

The linguistic module mostly concerns the extraction of prominent syntactic boundaries from deep syntactic parsing, based on syntactic constituency (*Constituent-Depth* [Cooper and Paccia-Cooper, 1980], *ϕ-phrases* [Gee and Grosjean, 1983, Delais-Roussarie, 2000], *Left-hand-side / Right-hand-side Boundary* [Watson and Gibson, 2004]), syntactic dependency (*Dependency-Grammar-based local markers* [Bailly, 1989, Barbosa, 2006]), or a combination (*Chunks-and-Dependencies* [Abney, 1992]). Some studies attempt at integrating higher-level linguistic constraints, such as syntactic/semantic constraints - defined in terms of the degree of syntactic dependency across successive syntactic constituents [Ferreira, 1988]. A score is associated with each of the considered syntactic cues, and combined to provide the likelihood or the strength of a prosodic boundary conditionally to the observed syntactic cues. The rhythmic module is used as a regularization process to adjust the produced prosodic structure with respect to the size of the prosodic constituent candidates [Gee and Grosjean, 1983, Bailly, 1989, Delais-Roussarie, 2000, Barbosa, 2006], either in parallel or in cascade with the linguistic module.

However, expert approaches design a *universal* model in which general principles are formulated so as to account for the inter-speaker variations of a given language. Consequently, such models do not account for the variations associated with a specific speaker or speaking style, and can not simply be adapted to a specific speaker or a specific speaking style.

4.7.2 Statistical Models

On the other hand, statistical approaches aims at elaborating a statistical model which accounts for the statistical dependencies that relate prosodic variations and linguistic cues from the observation of their relative co-occurrence on large speech databases. Formally, statistical models are used to estimate the likelihood of a prosodic structure conditionally to the observed information extracted from text. In a similar manner as for expert models, statistical approaches mostly

concern the modelling of prosodic boundaries[1]. Thus, for a given sentence, each form is associated with the *likelihood* that a prosodic boundary exists between this form and the following.

Statistical approaches mainly divide into *static* (Decision-Tree-Based [Hirschberg, 1991, Black and Taylor, 1994]), *dynamic* (HMM-based [Veilleux et al., 1990, Ross and Ostendorf, 1996, Black and Taylor, 1997a, Sun and Applebaum, 2001, Atterer and Klein, 2002, Schmid and Atterer, 2004, Bonafonte and Agüero, 2004, Bell et al., 2006]), and *hierarchical* (Hierarchical HMM, Weighted Tree Automata [Ostendorf and Veilleux, 1994, Rangarajan Sridhar et al., 2008]) methods.

static methods models the likelihood of a prosodic structure given the observed linguistic information, solely [Hirschberg, 1991, Black and Taylor, 1994].

dynamic methods additionally regularize the likelihood of a prosodic structure given the observed linguistic information with that of the prosodic structure.

Additionally, statistical models differ in the representation of the prosodic structure:

sequential methods assume the prosodic structure as a sequential structure [Veilleux et al., 1990, Ross and Ostendorf, 1996, Black and Taylor, 1997a, Schmid and Atterer, 2004, Bonafonte and Agüero, 2004, Sun and Applebaum, 2001];

hierarchical methods explicitly model the prosodic structure as a hierarchical structure [Ostendorf and Veilleux, 1994, Rangarajan Sridhar et al., 2008].

Linguistic dependencies used to be statistically modelled based on syntactic information extracted from surface parsing, such as lexical category (Part-Of-Speech or POS) or lexical class (content and function forms) of a form, and punctuation markers. Rare studies exist on the integration of a rich syntactic description into speech prosody modelling, without significant improvements [Ingulfen et al., 2005].

Statistical models present various advantages over expert models. Firstly, parametric models can adequately model and adapt to the prosodic strategies associated with a speaker or speaking style. Secondly, statistical models can accurately model various and complex linguistic dependencies in a proportion and a time that would be unreachable for the expert.

Finally, expert and statistical models do not oppose to each other and benefit of their mutual advances: statistical models are introduced into expert models ([Barbosa, 2006]), and statistical models benefit of expert knowledge. In particular, recent statistical models have been proposed to explicitly account for rhythmic constraints (segmental models [Ostendorf and Veilleux, 1994, Schmid and Atterer, 2004, Bell et al., 2006]), or for the modelling of the hierarchical organization of the prosodic structure [Ostendorf and Veilleux, 1994, Rangarajan Sridhar et al., 2008].

4.8 Continuous Modelling of Speech Prosody

The principle of continuous modelling of speech prosody is to determine the sequence of acoustic parameters that corresponds to an input text and the corresponding sequence of prosodic events.

Statistical methods are commonly used to model the acoustic variations of speech prosody. The prosodic parameters to be modelled are generally limited to the conventional prosodic parameters: f_0 variations, and durations.

[1] In particular, lexical stress is imposed in stress-based languages such as English, and does not require statistical modelling. Additionally, residual prosodic prominences such as prosodic focus are not used to be modelled. Consequently, most of the studies focus on major prosodic boundaries (or prosodic break) modelling.

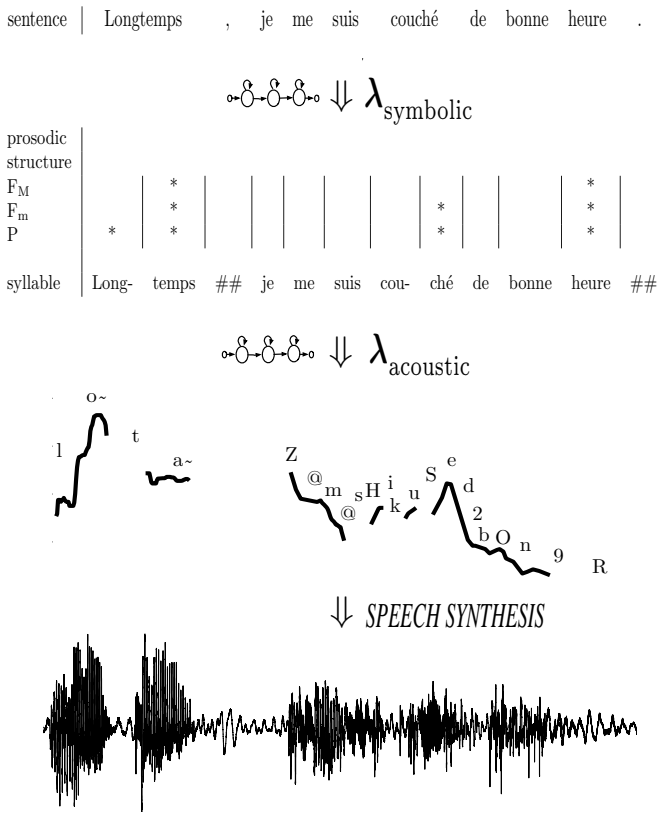


Figure 4.2: Synthesis of speech prosody parameters.

Statistical modelling of prosodic variations includes short [Tokuda et al., 1995, Yoshimura et al., 1999, Tokuda et al., 2003] and long-term [Latorre and Akamine, 2008, Qian et al., 2009] modelling, stylization methods, multi-level methods [Latorre and Akamine, 2008, Qian et al., 2009], and unsupervised statistical techniques [Morlec, 1997, Holm, 2003]. Continuous models of speech prosody vary depending on the signal model and the temporal domains that are used to model the speech prosody variations. Statistical continuous models decompose into:

short-term approaches is which the short-term variations of speech prosody are modelled over the sub-states of the phoneme [Yoshimura et al., 1999, Tokuda et al., 2003, Zen et al., 2004] (*frame-based*);

long-term approaches in which long-term variations of speech prosody are described and modelled over linguistically-motivated [Latorre and Akamine, 2008, Qian et al., 2009] or data-driven [Morlec, 1997, Holm, 2003] temporal domains (*multiple-levels*).

The prosodic parameters are generally modelled separately, with the exception of the hidden-semi-Markov-model (HSMM) in which the f_0 variations and the temporal structure (state-duration) are jointly modelled [Zen et al., 2004].

4.8.1 Short-Term Modelling

The HMM-based modelling is the most popular method used to model speech prosody in parallel to the development of the HMM-based speech synthesis [Tokuda et al., 1995, Yoshimura et al., 1999, Tokuda et al., 2003, Zen et al., 2004].

The HMM-based speech synthesis presents the advantage of a unified statistical framework in which the speech parameters and their temporal structure are simultaneously modelled with a hidden Markov model in context (context-dependent HMMs). However, the conventional HMM-based speech synthesis system suffers from a major problem in speech prosody modelling that is primarily due to its original design. Indeed, the HMM-based speech synthesis system originally derives from the inversion of the statistical methods used in automatic speech segmentation systems for speech synthesis. The acoustic parameters are represented as the instantaneous speech characteristics that are used to model the short-term phonatory characteristics of a speaker (articulation and co-articulation). More precisely, each phoneme in context is modelled by a HMM which describes the short-term variations of the acoustic parameters over the phoneme. Consequently, the conventional HMM-based speech synthesis system does not account for long-term variations that is required to model appropriately speech prosody variations.

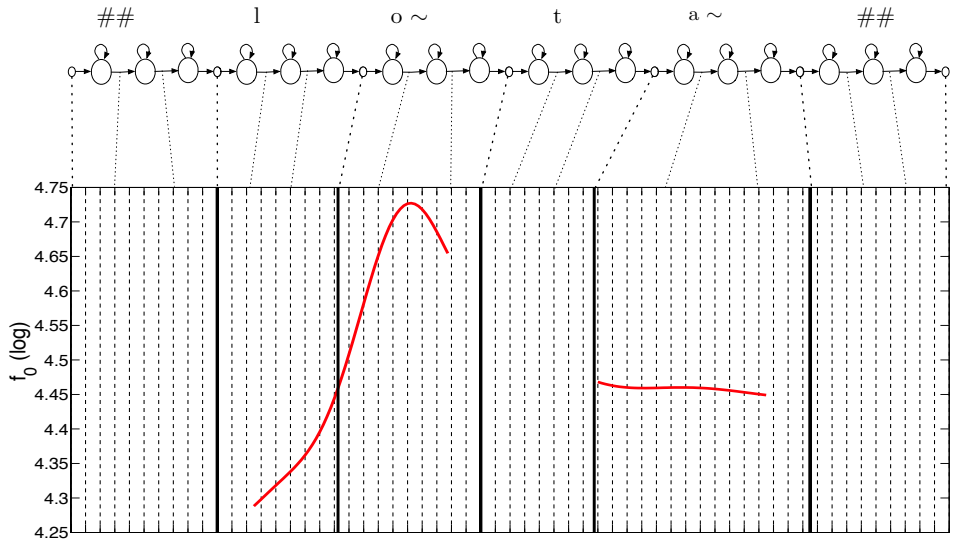


Figure 4.3: Short-term modelling of the f_0 variations with 3-states left-to-right context-dependent HMMs.

4.8.2 Long-Term Modelling

Statistical methods generally assume a temporal domain that is considered relevant for the description of the acoustic variations. The phoneme [Yamagishi et al., 2008], the syllable [Ladd and Campbell, 1991, Dusterhoff et al., 1999, Chen et al., 2003, Sreenivasa Rao and Yegnanarayana, 2007, Shuang et al., 2009], syllable-like [Barbosa, 2004], or the prosodic phrase are commonly considered as relevant temporal domains for the description of speech prosody. The description of the speech prosody variations is either based on short-term representation [Shuang et al., 2009] or on stylization methods [Dusterhoff et al., 1999].

4.8.3 Simultaneous Modelling over Various Temporal Domains

The multiple-level approach is the extension of the long-term methods to model variations simultaneously over various temporal domains that are considered as relevant for the description of speech prosody. Theoretical studies on speech prosody generally support that speech prosody consists in the co-occurrence of acoustic variations occurring over different temporal domains [Fujisaki, 1983, Van Santen and Moebius, 1999] which are associated to different communicative functions. Following these studies, various methods have been proposed to model speech prosody variations over various temporal domains, in particular for

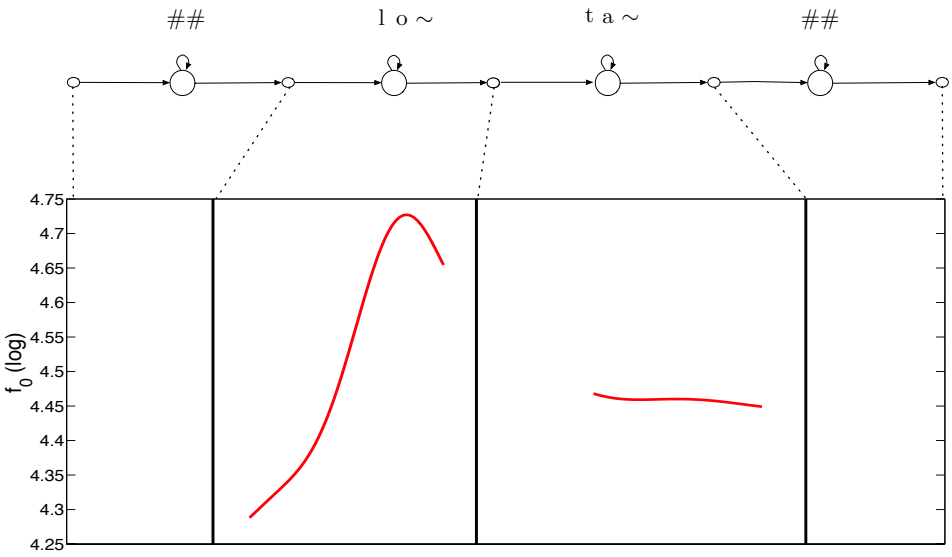


Figure 4.4: Modelling of the f_0 variations with stylization of the f_0 contours over the syllable.

the modelling of f_0 variations [Aubergé, 1991, Morlec, 1997, Holm, 2003]. In parallel to the development of HMM-based speech synthesis [Zen et al., 2007], there is currently a trend for the integration of multi-level approaches into HMM-based speech prosody modelling, either for the modelling of f_0 variations [Latorre and Akamine, 2008, Qian et al., 2009], or state-durations [Gao et al., 2008, Obin et al., 2009a].

Multi-levels methods divide into:

superpositional methods in which speech prosody variations are iteratively decomposed as the superposition of elementary contours occurring over different temporal domains,

joint methods in which speech prosody variations that occur over different temporal domains are *jointly* described and modelled,

unsupervised methods in which speech prosody variations are modelled over non-specified and data-driven temporal domains.

A description of the different multi-levels methods is shortly presented in the following.

4.8.3.1 Superpositional model

Superpositional model is historically the first attempt to decompose speech prosody over several temporal domains, either physiologically [Fujisaki, 1983] or linguistically motivated [Aubergé, 1991, Van Santen and Moebius, 1999, Obin et al., 2009a].

In the superpositional model, speech prosody is decomposed as a superposition of elementary contours occurring over different temporal domains. The superposition refers either to a additive or multiplicative decomposition. More precisely, speech prosody is decomposed recursively over a set of linguistically-motivated temporal domains, generally from the larger to the smaller. For a given temporal domain, a prosodic contour is estimated that describes the speech prosody characteristics over the temporal segment. Then, the estimated contour is subtracted from the observed variations so as to form a residual that will be used to describe the remaining variations over further temporal domains. During the training, the variations are modelled over each temporal domain separately. During the synthesis, the variations are inferred over each temporal domain separately

and then superposed. The syllable and the prosodic phrase are commonly used as prosodic units to decompose speech prosody variations.

4.8.3.2 Joint model

The joint model is the extension of the HMM-based speech synthesis to account for the long-term variations of speech prosody [Gao et al., 2008, Latorre and Akamine, 2008, Qian et al., 2009].

In the joint model, speech prosody variations are *jointly* described and modelled over a set of linguistic-defined temporal domains. The joint model presents several advantages over the superpositional model. Firstly, no decomposition is required for the description of speech prosody, thus no bias is introduced due to the decomposition. Secondly, the joint model adequately models the covariation that may exists over different temporal domains. During the synthesis, speech prosody variations are inferred so as to maximize the short-term variations under the constraint of the long-term variations. The joint model is either used for the modelling of f_0 variations [Latorre and Akamine, 2008, Qian et al., 2009] and extended to the temporal structure [Gao et al., 2008], and different temporal domains were experimented from the phoneme, the syllable, and the prosodic group.

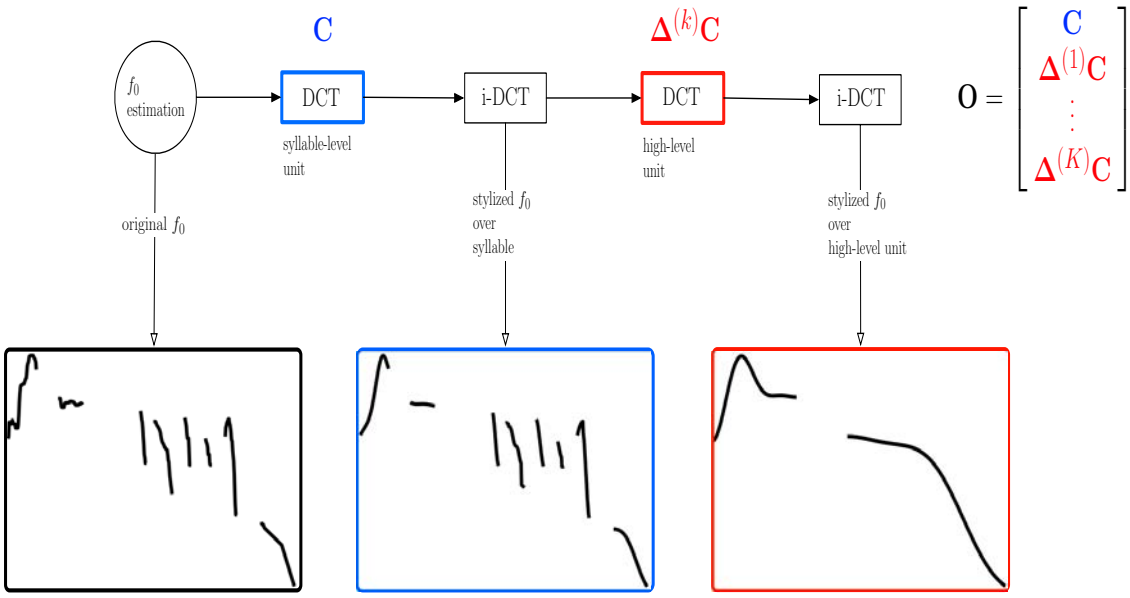


Figure 4.5: Simultaneous modelling of f_0 variations with stylization over various temporal domains. Left: observed f_0 variations, middle: f_0 stylization over syllable, right: f_0 stylization over prosodic phrase.

4.8.3.3 Unsupervised model

Unsupervised statistical methods have been developed to alleviate the problem caused by the definition of relevant prosodic domains. In the unsupervised model, speech prosody is modelled based on the unsupervised segmentation of speech prosody into relevant temporal domains, in which segmentation and modelling are jointly determined. Thus, the temporal domains used for the modelling of speech prosody directly emerge from the analysis [Morlec, 1997, Holm, 2003].

Finally, the temporal structure modelling of speech prosody remains generally static, i.e. the long-term variations of speech rate are generally not considered for the estimation neither

for the modelling of state durations [Zen et al., 2004, Sreenivasa Rao and Yegnanarayana, 2007, Yamagishi et al., 2008] - with the exception of [Gao et al., 2008].

Part II

Discrete/Continuous Modelling of Speech Prosody

Abstract

In this part, a complete statistical model is proposed to model the speech prosody characteristics of a speaker. In the proposed approach, the symbolic/acoustic characteristics of a speaker are statistically modelled based on context-dependent discrete/continuous HMMs. The main advances in speech prosody modelling are the following:

1. a complete system is proposed to model the symbolic and acoustic speech prosody characteristics of a speaker.
2. a unified context-dependent HMM is proposed based on the discrete/continuous statistical modelling of speech prosody characteristics.
3. a rich linguistic description is proposed to refine context-dependent modelling of speech prosody.
4. a trajectory model based on the stylisation of speech prosody variations over various temporal domains is proposed.
5. To a lesser extent, a method to vary the speech prosody of a speaker is proposed.

The architecture of the proposed speech prosody system is presented and briefly discussed below.

Architecture of the Speech Prosody System

The architecture of the proposed speech prosody system is designed so that the symbolic and the acoustic characteristics of a speaker are explicitly distinguished: on the one hand, the symbolic description accounts exclusively for the observation of relevant prosodic events and the corresponding abstract prosodic structure - without the specification of a phonetic characteristic or a precise contour. On the other hand, the acoustic description accounts exclusively for the actual prosodic contours and acoustic variations. This presents advantages in flexibility and modularity over conventional speech prosody systems (e.g. in which the use of the ToBI transcription actually describes specific prosodic contours). In particular, the proposed decomposition can be efficiently used to vary the speech prosody of a speaker: a variety of prosodic alternatives may be potentially associated with a given prosodic structure. More generally, speech prosody can be varied independently over the symbolic or the acoustic module, thus relevant speech prosody alternatives may be obtained by symbolic variations, acoustic variations, and their combination.

Description of Speech Prosody

Symbolic Description of Speech Prosody

The Rhapsodie transcription system was chosen for the description of prosodic events and prosodic structure as an alternative to the conventional ToBI transcription system for the transcription of French speech prosody. Firstly, the Rhapsodie transcription system efficiently describes the hierarchy of prosodic events with a very limited set of symbols (prosodic prominence, minor and major prosodic boundaries). Secondly, the Rhapsodie transcription guidelines do not require

expert-knowledge and can be easily used for manual transcription and automatic labelling based on acoustic analysis. Finally, the RHAPSODIE transcription is primarily based on the perception of acoustic saliency and is not confined to the description of intonational prominences or the identification of specific intonational contours only. Thus, the transcription can be easily integrated into most of the existing models for further phonetic and phonological descriptions.

Acoustic Description of Speech Prosody

The conventional speech prosody parameters are used for the modelling of the acoustic variations: f_0 variations and the temporal structure. The syllable is chosen as the minimal prosodic unit for the description of speech prosody variations. The f_0 variations are described and stylized over various temporal domains using a Discrete Cosine Transform (DCT). The temporal structure is described with respect to the sequence of syllable durations.

Modelling of Speech Prosody

Rich Linguistic Description

An automatic linguistic processing chain is used to enrich the linguistic description of a text in context-dependent HMM speech prosody modelling. The linguistic processing chain includes text pre-processing, surface parsing, and deep parsing. A preprocessing is achieved in order to segment a raw text into linguistic units that can be used by a linguistic parser (such as form and sentence segmentation). Surface parsing is processed to provide a morpho-syntactic analysis of the text. Then, Deep parsing is achieved based on Tree Adjoining Grammar (TAG) which represents both dependency graph and constituency structure derived from the text analysis. The extracted syntactic features are classified into different sets depending on their nature: morpho-syntactic, dependency and constituency features. Additionally, adjunction features which cover a large variety of syntactic constructions (e.g., relative clauses, incises) are additionally introduced for comparison.

Discrete Modelling of Speech Prosody

A context-dependent discrete HMM is presented in which the symbolic characteristics of speech prosody are modelled conditionally to the linguistic context in which they are observed. A method that combines linguistic and metric constraints for prosodic break modelling is proposed based on segmental HMMs and Dempster-Shafer fusion, and the relative importance of the linguistic and the metric constraints is assessed depending on the nature of the linguistic information. Finally, a method to vary the speech prosody of a speaker based on the *General Viterbi Algorithm* (GVA) is presented.

Continuous Modelling of Speech Prosody

A context-dependent continuous HMM is proposed in which f_0 variations are stylized and jointly modelled over various temporal domains. The syllable is defined as the minimal prosodic unit for the description of speech prosody variations: syllable durations are used to explicitly represent prosodic timing, and f_0 variations are stylized over various temporal domains using a *Discrete Cosine Transform* (DCT). Each of the prosodic dimensions is modelled separately using a context-dependent continuous HMM. Syllable duration is modelled using a conventional context-dependent continuous HMM, and f_0 variations are modelled using a *Joint Trajectory Model*.

Discrete and continuous modelling of speech prosody are evaluated separately, and the role of linguistic contexts in the context-modelling of speech prosody is assessed depending on the nature of the linguistic information. The proposed models are either objectively or subjectively evaluated, and the evaluation of speech prosody is discussed with regard to the existence of a large variety

of speech prosody alternatives that exist.

The present part is organized as follows: the text and speech material used to model the speech prosody of a speaker is presented in chapter 5. The basic principles of the hidden Markov model (HMM) and context-dependent modelling are briefly described in chapter 6 . The linguistic processing chain and the deep syntactic parser used for the enrichment of the linguistic description are described in chapter 7. The discrete modelling of speech prosody is presented and evaluated in chapter 8.1 . The continuous modelling of speech prosody is presented and evaluated in chapter 9.

Chapter 5

Text & Speech Material

Contents

5.1 Speech Material

In this study, two French read-speech speech databases were exploited: a *laboratory* and a *multi-media* corpus.

The laboratory corpus is composed of short sentences, automatically retrieved from the internet in order to design a phonetically well-balanced speech database for speech synthesis. Each sentence was read separately by a non-professional French speaker and recorded in an anechoic room.

The multi-media corpus is the novel "*Du côté de Chez Swann*" ("*Swann's Way*"), the first volume of "*A la Recherche du Temps Perdu*" ("*In Search of Lost Time*") [Proust, 1913] by the French writer Marcel Proust. The text was read by the French professional actor *André Dussolier* in the context of story telling in the audio-book format. The text was recorded in home made conditions, and interpreted by the actor according to his understanding of the narrative structure.

The laboratory corpus consists of simple linguistic structures and controlled speech prosody, while the multi-media corpus consists of complex linguistic structures and rich speech prosody. In particular, French writer Marcel Proust is famous for the high syntactic complexity of his style (e.g., long-term syntactic dependencies, embedded clauses), and the professional actor uses a wide variety of prosodic strategies.

The speech material is composed of speech utterances and their corresponding text transcriptions. The audio material was recorded with a high quality microphone and a 16 bits 44.1 kHz analogue-to-digital converter.

Table 5.1 summarizes the characteristics of the speech databases.

corpus	speech type	style	speaker expertise	corpus size	linguistic complexity	prosodic complexity
laboratory	read	neutral	naïve	9h	-	-
multi-media	read	story-telling	professional	7h	+	+

Table 5.1: Description of the speech material.

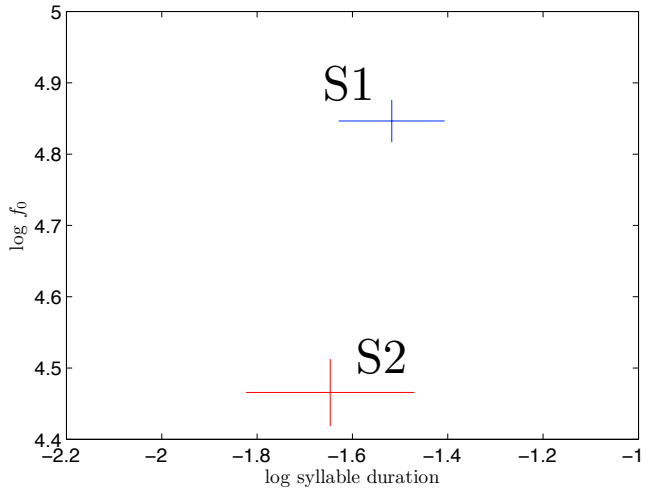


Figure 5.1: Prosodic description of the speakers: mean and variance of the speaker. S1: non-professional speaker and read laboratory speech; S2: professional speaker and story telling.

5.2 Speech Segmentation

The speech material was phonetically aligned to the text transcription using the HMM-based phoneme segmentation IRCAMALIGN system [Lanchantin et al., 2008] and the *hidden Markov model Toolkit* (HTK, [Young et al., 2002]).

During the training, a speaker-dependent HMM model was estimated using a bootstrap of manually aligned speech utterances. The typology of the model is a tri-phone model, in which each phoneme is model by 7 left-to-right with-no-skip states and 5 normal distributions per state.

During the alignment, the text is first segmented into forms and a multi-pronunciation phonetic lattice using LIA_PHON, a rule-based text-to-phoneme conversion system [Béchet, 2001]. Then, the sequence of phonemes is determined and aligned with the speech signal using conventional Viterbi algorithm.

5.3 Transcription of Speech Prosody

The RHAPSODIE transcription was adopted as an alternative to TOBI [Silverman et al., 1992] and other transcription systems for French prosody labelling [Lacheret et al., 2010]. The description of the prosodic variations is based on the perception of prosodic events that are implicitly shared among phonological theories, such as *prosodic prominence* and *prosodic grouping*.

The description is based on the following assumptions:

1. the prosodic structure is hierarchical rather than sequential;
2. the syllable is the minimal unit that convey prosodic information;
3. a prosodic prominence is defined as a perceived acoustic saliency, regardless to a precise acoustic dimension, contour, and/or function that may be associated;

The prosodic grammar used for description is composed of: major prosodic boundary (F_M), minor prosodic boundary (F_m), and prosodic prominence (P). Then, the prosodic structure is recursively transcribed from the maximal prosodic unit (F_M) to the local variations (P).

Prosodic units are defined in correspondence with the prosodic labels:

utterance : maximal unit of current speech synthesis systems. However, the utterance may not be precisely considered as a prosodic unit.

major prosodic group : maximal prosodic unit associated with a major prosodic boundary (F_M). A prosodic group is defined as the prosodic unit that ends with a major prosodic boundary, and is used for prosodic segmentation;

minor prosodic group : intermediate prosodic unit associated with a minor prosodic boundary (F_m). A minor prosodic group is defined as the prosodic unit that ends with a minor prosodic boundary, and is used for rhythmic grouping that is typical of French.

syllable : minimal prosodic unit potentially associated with a prosodic prominence (P).

The RHAPSODIE system was used for the automatic transcription of speech prosody, in a similar manner to that used for the manual transcription: syllable prosodic prominence are first identified and then organized in a hierarchy based on a combination of acoustic and linguistic information.

Firstly, syllabification is used to convert the phonetic sequence into a syllable sequence, based on a rule-based phoneme-to-syllable conversion system [Boula de Mareüil, 1997].

Then, automatic prosodic prominence transcription is performed using the IRCAMPROM system [Obin et al., 2008c, Obin et al., 2009b][1]. Short-term acoustic features are extracted from the speech signal including f_0 variations, syllable duration, intensity, spectral information, and vocal quality. Then, acoustic features are computed over the syllable and normalized with respect to those observed on larger temporal domains (e.g., surrounding syllables, prosodic phrase). Firstly, a *feature selection* method based on *Inertia Ratio Maximization and Feature Space Projection* (IRMFSP) [Peeters, 2003] is used to select the most discriminant acoustic features [Obin et al., 2008c]. Secondly, a *feature transform* method is used based on *Discriminant Analysis* to determine a set of linear combinations of the acoustic features that maximizes the discrimination of prosodic prominences [Obin et al., 2009b]. Finally, a prominence model is estimated using a *Gaussian Mixture Model* (GMM). The prominence model was estimated on a set of manually labelled speech utterances.

Finally, the prosodic hierarchy is retrieved using simple heuristics: a major boundary (F_M) is a prosodic prominence that is followed by a pause; a minor boundary (F_m) is a prosodic prominence that ends a syntactic chunk; a prosodic prominence (P) is a residual prosodic prominence, mostly associated with a semantic or a discursive focus.

The prosodic hierarchy is represented into a prosodic tree in which the relationships of the prosodic units are retrieved from the symbolic alignment and text information (e.g., syntactic chunk). Finally, the prosodic tree is aligned with the speech utterance according the phonetic segmentation.

[1] For reasons of space and clarity, the studies on automatic prosodic prominence labelling are not presented in this thesis.

5.4 Extraction of Prosodic Parameters

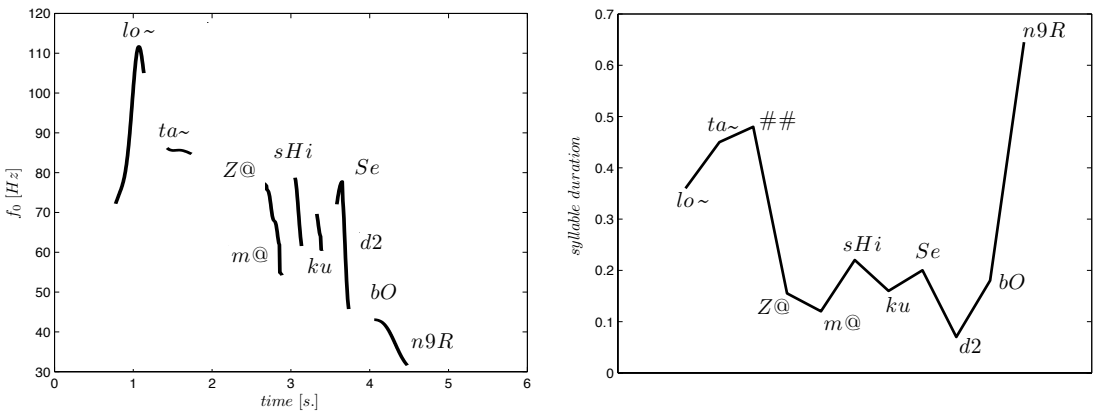
5.4.1 Fundamental Frequency (f_0)

The fundamental frequency f_0 and periodicity measure were estimated using the STRAIGHT algorithm, a frequency-based fundamental frequency estimation method based on instantaneous frequency estimation and fixed-point analysis [Kawahara et al., 1999a].

The analysis was performed using a 50-ms. Blackmann window and a 5 ms. frame rate. F_0 boundaries set for the analysis were manually adapted depending on the characteristics of the speaker. The voiced/unvoiced regions were determined using the aperiodicity measure.

5.4.2 Syllable duration

Syllable durations were simply extracted with respect to the phonetic alignment and the syllabification.



Description of f_0 and syllable duration characteristics for the utterance: : "*Longtemps, je me suis couché de bonne heure.*" ("*For a long time I used to go to bed early*").

corpus	laboratory			multi-media		
acoustic						
segmentation						
utterance	9280			2184		
syllable	#	# / utt.		#	# / utt.	
	163194	18		86050	**40**	
phone	#	# / utt.		#	# / utt.	
	456962	50		263134	**120**	
prosodic analysis	#	# / utt.	%	#	# / utt.	%
T	41199	4.5	**25**	19797	9.1	23
F_M	17975	**1.9**	11 (**44**)	5812	**2.7**	7 (**30**)
F_m	15871	1.7	9.5 (38)	9503	4.4	11 (48)
P	7353	**0.8**	4.5 (**18**)	4482	**2.0**	5.2 (22)
lingusitic						
segmentation						
sentence	9280			2184		
forms	89231	9.5		52358	**27**	
parsing						
coverage %	80			52		
ambiguity	0.69			1.2		

Table 5.2: Descriptive analysis of text and speech materials.

Chapter 6

The Hidden Markov Model

Contents

The *hidden Markov model* (HMM) [Baum and Petrie, 1966] is one of the most commonly used statistical models for stochastic processes and in particular time-series processes. Hidden Markov models have been widely used in many domains (speech processing [Rabiner, 1989], natural language processing [Manning and Schütze, 1999], machine translation [Vogel et al., 1996], handwriting recognition [Kundu and Bahl, 1988], image processing [Geman and Geman, 1984], music [Cont, 2010]...). In particular, hidden Markov models have been introduced into speech processing systems such as speech recognition ([Bahl et al., 1983, Lee, 1988, Leggetter and Woodland, 1995, Ostendorf et al., 1996]), speaker identification [Matsui and Furui, 1992, Reynolds and Carlson, 1995], and emotion recognition [Nogueiras et al., 2001, Nwe et al., 2003]. More recently, hidden Markov models have been successfully extended into Text-To-Speech synthesis with various refinements (HMM-based speech synthesis [Tokuda et al., 1995, Yoshimura et al., 1999], multi-space distributions [Tokuda et al., 1999], segmental model [Yoshimura et al., 1998], trajectory model [Tokuda et al., 2003], speaker adaptation [Yamagishi, 2006], and global variance [Toda and Tokuda, 2007]). In this chapter, we briefly describe the HMM framework and notations that will be used in the rest of the thesis.

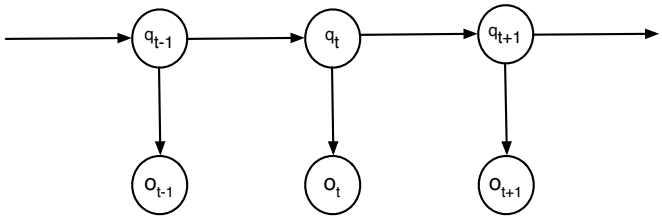


Figure 6.1: Representation of the dependence structure of an hidden Markov model.

6.1 Definition

Let define $\mathbf{o} = [o_1, \ldots, o_T]$ an observation sequence of length T and $\mathbf{q} = [q_1, \ldots, q_T]$ an associated state sequence, $q_t \in [1, N]$, $\forall t \in [1, T]$. A hidden Markov model is a bivariate stochastic time process ($\mathbf{q} = [q_1, \ldots, q_T]; \mathbf{o} = [o_1, \ldots, o_T]$) where $\mathbf{q}$ is a hidden Markov process and, conditional on $\mathbf{q}$, $\mathbf{o}$ is an observed stochastic process of independent random variables such that the conditional distribution of o_t only depends on q_t at time t. A Markov process is a stochastic process with the property that the next state depends only on the present state; that is, given the present, the future does not depend on the past.

$$\mathrm{p}(q_{t+1}|q_t, \ldots, q_1) = \mathrm{p}(q_{t+1}|q_t) \qquad t \in [1, T] \tag{6.1}$$

A hidden Markov model $\boldsymbol{\lambda}$ is defined by the triple $\boldsymbol{\lambda} = (\mathbf{\Pi}, \mathbf{A}, \mathbf{B})$, where:

- Π is the initial state probability distribution: $\mathbf{\Pi} = \{\pi_i\}_{i=1}^N$

$$\pi_i = \mathrm{p}(q_1 = i) \qquad i \in [1, N] \tag{6.2}$$

  where:

$$\pi_i \geq 0$$

  and

$$\sum_{i=1}^{N} \pi_i = 1$$

- $\mathbf{A}$ is the state transition probability distribution: $\mathbf{A} = \{a_{i,j}\}_{i,j=1}^N$

$$a_{i,j} = \mathrm{p}(q_t = j|q_{t-1} = i) \qquad \begin{matrix} t \in [1, T] \\ i, j \in [1, N] \end{matrix} \tag{6.3}$$

  where:

$$a_{i,j} \geq 0$$

  and

$$\sum_{j=1}^{N} a_{i,j} = 1$$

- $\mathbf{B}$ is the output probability distribution: $\mathbf{B} = \{b_i(o_t)\}_{i=1}^N$

$$b_i(o_t) = \mathrm{p}(o_t|q_t = i) \qquad \begin{matrix} t \in [1, T] \\ i \in [1, N] \end{matrix} \tag{6.4}$$

The output probability $b_i(o_t)$ of the observation o_t can be either discrete or continuous depending on the nature of the observations. In the case of continuous observations, the output probability density is usually modelled by a *Gaussian Mixture Model* (GMM), i.e. a weighted mixture of gaussian distributions:

$$b_i(o_t) = \sum_{m=1}^{M} \alpha_{i,m} \mathcal{N}(o_t|\mu_{i,m}, \Sigma_{i,m}) \qquad i \in [1, N] \tag{6.5}$$

where:

$$\int_{o_t} b_i(o_t)\, do_t = 1 \qquad\qquad t \in [1, T]$$

and M is the number of mixture components, $\alpha_{i,m}$ is the weight of the m-th mixture component of state i, $\mu_{i,m}$ and $\sigma_{i,m}$ are the parameters of the m-th Gaussian distribution $\mathcal{N}$ of state i.

A D-dimensional Gaussian distribution is defined by:

$$\mathcal{N}(o_t|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D |\boldsymbol{\Sigma}|}}\, e^{-\frac{1}{2}(o_t-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(o_t-\boldsymbol{\mu})} \tag{6.6}$$

where $\boldsymbol{\mu}$ is the (1xD) mean vector and $\boldsymbol{\Sigma}$ the (DxD) covariance matrix of the Gaussian distribution.

There are three common problems related to hidden Markov models [Rabiner, 1989]: (1) the evaluation of the probability $\mathrm{p}(\mathbf{o}|\boldsymbol{\lambda})$ of an observation sequence $\mathbf{o} = [o_1, \ldots, o_T]$ given the model $\boldsymbol{\lambda}$; (2) the determination of the optimal state sequence $\widehat{\mathbf{q}} = [\widehat{q_1}, \ldots, \widehat{q_T}]$ given an observation sequence $\mathbf{o} = [o_1, \ldots, o_T]$ and the model $\boldsymbol{\lambda}$; (3) the estimation of the model parameters $\boldsymbol{\lambda}$ which optimize a given objective function of an observation sequence $\mathbf{o} = [o_1, \ldots, o_T]$ given the model $\boldsymbol{\lambda}$.

6.2 Probability Estimation

The probability $\mathrm{p}(\mathbf{o}|\boldsymbol{\lambda})$ of the observation sequence $\mathbf{o} = [o_1, ..., o_T]$ given the model $\boldsymbol{\lambda}$ is obtained by marginalizing the joint probability $\mathrm{p}(\mathbf{o}, \mathbf{q}|\boldsymbol{\lambda})$ of the observation sequence $\mathbf{o}$ and the state sequence $\mathbf{q}$ given the model $\boldsymbol{\lambda}$:

$$\mathrm{p}(\mathbf{o}|\boldsymbol{\lambda}) = \sum_{\forall \mathbf{q}} \mathrm{p}(\mathbf{o}, \mathbf{q}|\boldsymbol{\lambda}) \tag{6.7}$$

Using Bayes' theorem:

$$\mathrm{p}(\mathbf{o}, \mathbf{q}|\boldsymbol{\lambda}) = \mathrm{p}(\mathbf{o}|\mathbf{q}, \boldsymbol{\lambda})\mathrm{p}(\mathbf{q}|\boldsymbol{\lambda}) \tag{6.8}$$

According to the statistical independence of observations, the conditional probability $\mathrm{p}(\mathbf{o}|\mathbf{q}, \boldsymbol{\lambda})$ of observations given the state sequence and the model is:

$$\mathrm{p}(\mathbf{o}|\mathbf{q}, \boldsymbol{\lambda}) = \prod_{t=1}^{T} \mathrm{p}(o_t|q_t, \boldsymbol{\lambda}) \tag{6.9}$$

$$= \prod_{t=1}^{T} b_{q_t}(o_t) \tag{6.10}$$

The probability $\mathrm{p}(\mathbf{q}|\boldsymbol{\lambda})$ of the state sequence is:

$$\mathrm{p}(\mathbf{q}|\boldsymbol{\lambda}) = \pi_{q_1} \prod_{t=2}^{T} a_{q_{t-1}, q_t} \tag{6.11}$$

Finally, one can formulate $\mathrm{p}(\mathbf{o}|\lambda)$ as:

$$\mathrm{p}(\mathbf{o}|\boldsymbol{\lambda}) = \sum_{\forall \mathbf{q}} \pi_{q_1} b_{q_1}(o_1) \prod_{t=2}^{T} a_{q_{t-1}, q_t} b_{q_t}(o_t) \tag{6.12}$$

This probability could thus be efficiently calculated using forward or backward probabilities defined as:

$$\alpha_t(i) = \mathrm{p}(o_1, \ldots, o_t, q_t = i|\lambda) \qquad \begin{array}{l} i \in [1, N] \\ t \in [1, T] \end{array} \tag{6.13a}$$

$$\beta_t(i) = \mathrm{p}(o_{t+1}, \ldots, o_T|q_t = i,\lambda\,) \qquad \begin{array}{l} i \in [1, N] \\ t \in [1, T] \end{array} \tag{6.13b}$$

The forward/backward probabilities can be recursively calculated as follows:

- **initialization:**

$$\alpha_1(i) = \pi_1 b_i(o_1) \qquad i \in [1, N] \tag{6.14a}$$

$$\beta_T(i) = 1 \qquad i \in [1, N] \tag{6.14b}$$

- **recursion:**

$$\alpha_t(j) = [\sum_{i=1}^{N} \alpha_{t-1}(i) a_{i,j}] b_j(o_t) \qquad \begin{array}{l} j \in [1, N] \\ t \in [2, T] \end{array} \tag{6.15a}$$

$$\beta_t(i) = \sum_{j=1}^{N} a_{i,j} b_j(o_{t+1}) \beta_{t+1}(j) \qquad \begin{array}{l} j \in [1, N] \\ t \in [2, T] \end{array} \tag{6.15b}$$

- **termination:**

$$\mathrm{p}(\mathbf{o}|\boldsymbol{\lambda}) = \sum_{i=1}^{N} \alpha_T(i) \tag{6.16a}$$

$$\mathrm{p}(\mathbf{o}|\boldsymbol{\lambda}) = \sum_{i=1}^{N} \pi_1(i) \beta_1(i) \tag{6.16b}$$

6.3 Optimal State Sequence

The second problem is to determine the optimal state sequence $\widehat{\mathbf{q}} = [\widehat{q}_1, \ldots, \widehat{q}_T]$ given an observation sequence $\mathbf{o} = [o_1, \ldots, o_T]$. Many criteria to define the optimality of a state sequence, and the most widely used criterion is to find the single most likely state sequence, i.e. to determine the state sequence which maximizes the conditional probability $\mathrm{p}(\mathbf{q}|\mathbf{o}, \boldsymbol{\lambda})$ of the sequence $\mathbf{q}$ given the observation sequence $\mathbf{o}$ and model $\boldsymbol{\lambda}$.

This criterion is referred as the *Maximum A Posteriori* (MAP):

$$\widehat{\mathbf{q}} = \underset{\mathbf{q}}{\operatorname{argmax}}\ \mathrm{p}(\mathbf{q}|\mathbf{o}, \boldsymbol{\lambda}) \tag{6.17}$$

This problem is formally solved using dynamic programming procedure referred as the *Viterbi Algorithm* (VA) ([Forney, 1973])

Let define:
$\delta_t(i) = \max\ \mathrm{p}(q_1, ..., q_t = i, o_1, ..., o_t|\boldsymbol{\lambda})$ the probability corresponding to the most likely state sequence which accounts for the partial observation sequence $[o_1, ...o_t]$ and ends in state i at time t;
$\psi_t(i)$ the state sequence associated with $\delta_t(i)$, i.e. the most likely state sequence which accounts for the partial observation sequence $[o_1, ...o_t]$ and ends in state i at time t;
$\Gamma_t(i,j) = \mathrm{p}(q_1, ..., q_{t-1} = i, q_t = j, o_1, ..., o_t|\boldsymbol{\lambda})$ the probability corresponding to the most likely state sequence which accounts for the partial observation sequence $[o_1, ...o_{t-1}]$ and ends in state i at time t-1 and in state j at time t.

Then, the structure of the Viterbi algorithm can be written as follows:

- **initialization**:

$$\delta_1(i) = \pi_i b_i(o_1) \qquad i \in [1, N] \tag{6.18a}$$
$$\psi_1(i) = 0 \qquad i \in [1, N] \tag{6.18b}$$

- **recursion**:
  - induction:

$$\Gamma_t(i,j) = \delta_{t-1}(i) a_{i,j} \qquad i, j \in [1, N],\ t \in [2, T] \tag{6.19}$$

  - selection:

$$\delta_t(j) = \left[\max_{\mathbf{i}} \Gamma_t(i,j)\right] b_j(o_t) \qquad j \in [1, N],\ t \in [2, T] \tag{6.20a}$$

$$\psi_t(j) = \operatorname*{argmax}_{\mathbf{i}} \Gamma_t(i,j) \qquad j \in [1, N],\ t \in [2, T] \tag{6.20b}$$

- **termination**:

$$\mathrm{p}(\mathbf{o}, \widehat{\mathbf{q}}|\lambda) = \max_{\mathbf{i}} \delta_T(i) \tag{6.21a}$$
$$\widehat{q}_T = \operatorname*{argmax}_{\mathbf{i}} \delta_T(i) \tag{6.21b}$$

- **sequence backtracking**:

$$\widehat{q}_t = \phi_{t+1}(\widehat{q}_{t+1}) \qquad t \in [T-1, 1] \tag{6.22}$$

6.4 Model Parameters Estimation

The third problem is the estimation of the model parameters $\boldsymbol{\lambda}$ which optimize a given objective function of the observation sequence $\mathbf{o} = [o_1, \ldots, o_T]$ given the model $\boldsymbol{\lambda}$. One of the most popular

objective function is the *Maximum-Likelihood* (ML), i.e. estimate the model parameters $\widehat{\boldsymbol{\lambda}}$ which maximize the probability of the observation sequence $\mathbf{o} = [o_1, \ldots, o_T]$ given the model $\boldsymbol{\lambda}$:

$$\widehat{\boldsymbol{\lambda}} = \arg\max_{\boldsymbol{\lambda}} \mathrm{p}(\mathbf{o}|\boldsymbol{\lambda}) \tag{6.23}$$

As there actually is no known way to analytically solve this problem, it is approximated using constrained optimization procedures. Being an optimization problem from incomplete observations, it is expensive to find a global solution, i.e. to estimate model parameters that globally maximize the likelihood of the observation sequence. However, methods have been derived to match a local solution, i.e. to estimate model parameters that locally maximize the likelihood of the observation sequence.
One of the most popular methods is the *Expectation-Maximization* (EM) algorithm, also referred as the *Baum-Welch* algorithm [Baum et al., 1970] in the context of Hidden Markov Model, which is an iterative procedure of model parameters reestimation.

6.4.1 Baum's Auxiliary Function

To do so, it is convenient to define an auxiliary function referred as the Baum's auxiliary function:

$$Q(\boldsymbol{\lambda}, \boldsymbol{\lambda}') = \sum_{\mathbf{q}} \mathrm{p}(\mathbf{q}|\mathbf{o}, \boldsymbol{\lambda}) \log(\mathrm{p}(\mathbf{o}, \mathbf{q}|\boldsymbol{\lambda}')) \tag{6.24}$$

It has been proved that the auxiliary function has a unique critical point that is a global maximum and that solution leads to increased value of the objective function, i.e.:

$$\max_{\bar{\lambda}} Q(\boldsymbol{\lambda}, \bar{\boldsymbol{\lambda}}) \Rightarrow \mathrm{p}(\mathbf{o}|\bar{\boldsymbol{\lambda}}) \geq \mathrm{p}(\mathbf{o}|\boldsymbol{\lambda}) \tag{6.25}$$

Thus, iteratively updating the auxiliary function can be proved to monotonically increase the objective function with convergence to a unique critical point.

6.4.2 Maximization of the Baum's Auxiliary Function

Finally, model parameters which maximize the auxiliary function under the constraints $\sum_{i=1}^{N} \pi_i = 1$ and $\sum_{j=1}^{N} a_{i,j} = 1$, $\forall i \in [1, N]$ can be formally derived using Lagrange multipliers and partial derivatives:

$$\bar{\pi_i} = \gamma_1(i) \tag{6.26}$$

$$\bar{a_{i,j}} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \tag{6.27}$$

$$\bar{\mu_i} = \frac{\sum_{t=1}^{T} \gamma_t(i,j) o_t}{\sum_{t=1}^{T} \gamma_t(i)} \tag{6.28}$$

$$\bar{\Sigma_i} = \frac{\sum_{t=1}^{T} \gamma_t(i,j)(o_t - \mu_i)(o_t - \mu_i)^T}{\sum_{t=1}^{T} \gamma_t(i)} \tag{6.29}$$

where:

$\gamma_t(i)$ the probability of being in state i at time t given the observation sequence and the model:

$$\gamma_t(i) = \mathrm{p}(q_t = i|\mathbf{o}, \boldsymbol{\lambda}) \quad (6.30)$$

$$= \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{N} \alpha_t(j)\beta_t(j)} \quad (6.31)$$

$\xi_t(i,j)$ the probability of being in state i at time t, and in state j at time $t+1$ given the observation sequence and the model:

$$\xi_t(i,j) = \mathrm{p}(q_t = i, q_{t+1} = j|\mathbf{o}, \boldsymbol{\lambda}) \quad (6.32)$$

6.5 Decision-Tree-Based Context-Clustering

As mentioned in section 2, many linguistic or para-linguistic factors affects speech prosody variations. The principle of context-dependent modelling is to model the characteristics of speech prosody depending on a specific linguistic context. Consequently, context-dependent modelling requires the estimation of the speech characteristics for each of the observed linguistic contexts. However, The huge amount of the linguistic contexts observed in natural language causes problems to robustly estimate the model parameters associated with each of the linguistic contexts. For instance, the simple tri-phone context modelling commonly used in speech recognition requires the estimation of $36^3 = 46656$ HMMs in French[1]. Additionally, the linguistic description used in speech prosody modelling is much more complex than it is in the case in speech recognition. This causes several computational issues: first, the estimation of the models parameters is time-consuming; secondly, the estimation of the model parameters has to deal with spare observations of the linguistic contexts: some of the linguistic contexts are poorly observed, and some of the linguistic contexts remain unseen to the observation. To overcome the estimation of context-dependent models in the case of large vocabulary contexts, context-clustering techniques have been proposed [O'Dell, 1995, Yoshimura et al., 1999, Shinoda and Watanabe, 2000].

Context-clustering aims at clustering linguistic contexts that are associated with similar observations, and sharing model parameters among these contexts. Such a method is used to ensure a robust estimation of the model parameters of the clustered linguistic contexts, in particular in the case of spare observations. Additionally, top-down context-clustering methods enable to model unseen linguistic contexts since any context can be clustered with the set of contexts which share at least one of its linguistic dimension.

Decision-tree-based context-clustering consists in estimating the tree derivation of contexts and associated model parameters which maximizes a given criterion, usually defined as the maximization of an objective function. However, estimating the tree that globally maximizes the objective function appears unrealistic since it requires an exhaustive search through all of the possible tree structures. Hopefully, a local solution to this problem consists in iteratively maximizing the objective function at each node of the tree.

The tree derivation can be summarized as follows:

1. The tree is initialized with the root node S_0 in which all the contexts share the same model parameters.
2. The tree is derived by iteratively finding the context that locally maximizes the objective function, then splitting model parameters into child nodes.
3. The tree derivation is stopped according to a local *model selection criterion*.

Methods to derive a context-dependent model in the case of discrete and continuous HMMs will be presented in chapters 8 and 9.

[1] French is composed of a set of approximatively 36 phonemes depending on the considered variants.

Chapter 7

Integration of Rich Linguistic Contexts [1]

Contents

7.1 Introduction

Among the large variety of source of variations, the linguistic dimension is the most commonly used for the modelling of speech prosody. First, a number of theoretical linguistic studies pointed out that speech prosody is produced by speakers and can be used by listeners to clarify the meaning and the structure of an utterance. More recently, studies described speech prosody over larger temporal domains, such as macro-syntactic and discursive structures. Second, studies in computational linguistic provided algorithms that can be used for the description of the syntactic structure of a text, including surface and deep syntactic parsing. The rich description of the syntactic characteristics of a text would qualitatively improve the naturalness and the variety of speech prosody in speech synthesis. Since most of the current speech prosody systems are based on a surface linguistic description only (such as part-of-speech and chunks [Ostendorf and Veilleux, 1994, Ross and Ostendorf, 1996, Black and Taylor, 1997a, Schmid and Atterer, 2004, Black and Taylor, 1997b, Zen et al., 2007]), there is a definitive need for the enrichment of the syntactic description used for the modelling of speech prosody.

In this chapter, an automatic linguistic processing chain is presented and described in order to enrich the linguistic description of a text for the modelling of speech prosody. The linguistic processing chain includes text preprocessing, surface parsing, and deep parsing (section 7.2). A

[1] In collaboration with: Eric Villemonte de la Clergerie (*INRIA, France*).

preprocessing is achieved in order to segment a raw text into linguistic units that can be used by a syntactic parser (such as sentence segmentation and form segmentation) (section 7.2.1). Surface parsing is used to provide a morpho-syntactic description of a sentence (section 7.2.1). Then, deep parsing is achieved based on Tree Adjoining Grammar (TAG), and used to represent both dependency graph and constituency structure derived from a sentence (section 7.2.2). The extraction of syntactic features from the linguistic analysis is presented in section 7.3. The extracted syntactic features are classified into different sets depending on their nature: morpho-syntactic features are extracted from the surface parsing, dependency and constituency features are extracted from deep parsing, and adjunction features are additionally introduced which are retrieved from deep parsing. The linguistic processing chain presented provides an enriched description of the text characteristics that will further be used to refine the context-dependent modelling of speech prosody. In particular, the relevancy of the syntactic characteristics will be compared and discussed for the symbolic and acoustic modelling of speech prosody in chapters 8 and 9.

7.2 Linguistic Analysis

An input text (sentence, set of sentences or raw text) is processed by a linguistic parser in order to provide a description of the text characteristics (surface and deep syntactical parsing) over the sentence. The *Alpage Linguistic Processing Chain*[2] is a full linguistic processing chain for French that is organized as a sequence of processing modules:

- a *lexer* module (LE*fff*: a French Morphological and Syntactic Lexicon [Sagot et al., 2006, Sagot, 2010]; SXPIPE: a full linguistic preprocessing chain for French [Sagot and Boullier, 2005]);
- a *parse* module (DyALog: a parser compiler and logic programming environment [Villemonte de La Clergerie, 2005a]; FRMG: a FRench Meta Grammar [Villemonte de La Clergerie, 2005b]),
- and a *post-processing* module.

7.2.1 Text Pre-Processing: Sentence Segmentation, Form Segmentation, and Surface Parsing

The lexer module uses SXPIPE to convert the input text into form lattices (represented as Direct Acyclic Graphs (DAGs)) that are combined with lexical information retrieved from LE*fff*. SXPIPE [Sagot and Boullier, 2005] is used to segment a raw text into forms and sentences that can be used by a parser. Text segmentation is achieved according to punctuation markers and local context [Grefenstette and Tapanainen, 1994]. Additionally, SXPIPE manages spelling error correction and complex forms processing, such as compound forms (e.g., "pomme de terre" = "pomme_de_terre") and agglutinates (e.g., "du" = "de le"). Then, LE*fff* [Sagot et al., 2006, Sagot, 2010] is used to provide morpho-syntactic and syntactic information retrieved from morphological and syntactic lexicons for each output sentence of SXPIPE. Finally, the output of the lexer associates each sentence to a form lattice enriched with the morpho-syntactic and syntactic information.

7.2.2 Text Analysis: Deep Parsing Based on Tree Adjoining Grammar (TAG)

Deep parsing is performed by the FRMG parser, a symbolic parser based on a compact *Tree Adjoining Grammar* (TAG) for French that is automatically generated from a Meta-Grammar (MG) [Villemonte de La Clergerie, 2005b, Villemonte de La Clergerie, 2005a].

A *Tree Adjoining Grammar* (TAG) [Joshi et al., 1975, Abeillé, 1988] consists of a finite set of *elementary trees*, and a set of *operations* that are used to derive trees from elementary trees. Thus,

[2] `http://alpage.inria.fr/alpc.en.html`

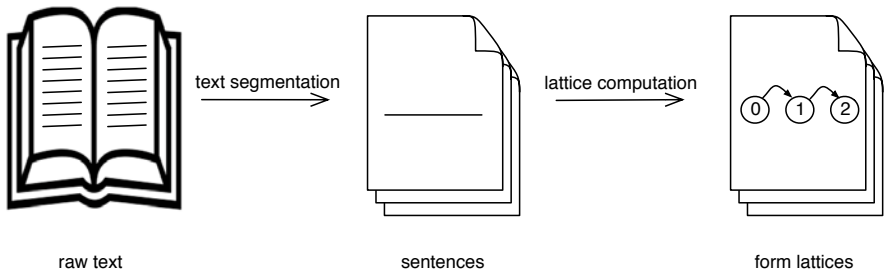


Figure 7.1: General Architecture of the SXPipe module.

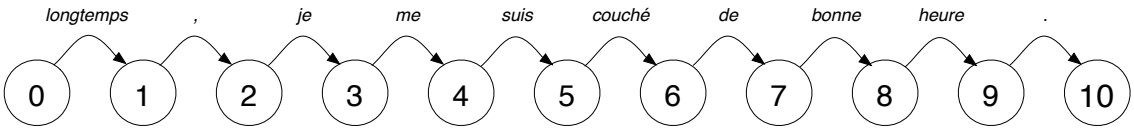


Figure 7.2: Direct Acyclic Graph (DAG) associated with the sentence: "*Longtemps, je me suis couché de bonne heure.*" ("*For a long time, I used to go to bed early.*").

Tree Adjoining Grammar is a tree automaton which accounts for all of the linguistic structures that can be derived from any combination of elementary trees by successive application of any of the operations that are included in the grammar.

Elementary trees are divided in *initial trees* and *auxiliary trees*. An initial tree α is a tree of which the non-terminal nodes are all labelled with non-terminal symbols, and the terminal nodes are either labelled with terminal symbols, or with non-terminal symbols. An auxiliary tree β is an initial tree which is constrained to have exactly one of its terminal nodes labelled with a non-terminal symbol which is the same as the label of the root node. These trees constitute the units of the grammar that can be interpreted as minimal linguistic structures.

Operations of *substitution* and *adjunction* are defined so as to derive trees from elementary trees. Substitution inserts an initial tree or a tree derived from an initial tree into an elementary tree. Adjunction inserts an auxiliary tree at one of the corresponding nodes of an elementary or a derived tree.

Feature structure is optionally associated with each node of an elementary tree to provide additional constraints on tree derivations.

Meta-Grammar and *Grammar Factorization* were introduced in order to reduce the amount of elementary trees needed to account for linguistic structures while preserving the generability of the Tree Adjoining Grammar.

Following the definition of the operations used in the Tree Adjoining Grammar [Abeillé, 1988]: 1) adjunction has recursive property, and 2) adjunction can insert a complete structure at a non-terminal node of another complete structure. Consequently, adjunction can derive complex linguistic structures and covers a large amount of various linguistic constructions as observed in natural language, from a single form adjunction such as adverbial or adjectival adjunction, to complete adjunction structure such as clauses, and even to complex adjunction structure such as embedded clauses.

The compilation and execution of the parser is performed within the DyALog system. The output of FRMG is a shared derivation forest that represents all of the possible derivation structures that the grammar can derive from the input sentence, and indicates which TAG *operation* (substitution,

adjunction, anchoring) took place on a given node of a given tree for a given constituent. The shared derivation forest is finally converted into a shared dependency forest by converting each anchor of the derivation structure into a dependency [Villemonte de La Clergerie, 2010].

A dependency forest is represented into a DEP XML format that incorporates the following items:

clusters that are associated with the forms of the sentence;

```
<cluster id="c_0_1" left="0" right="1" token="longtemps" lex="F1|Longtemps"/>
```

nodes that point to a given cluster. Nodes are associated with a form, a lemma, a syntactic category, an anchored tree, the maximal syntactic category of the anchored tree, and a set of derivations;

```
<node cluster="c_0_1" form="longtemps" lemma="longtemps" cat="adv" xcat="S"
       id="n019"  deriv="d000015" tree="153 adv_s arg0:adv_subcat
       modifier_at_S_level modifier_before_S modifier_before_x shallow_auxiliary"/>
```

edges that connect a source node with a target node. Edges are associated with a label, a type, and a set of *derivations* which instantiate this edge to connect a source constituent with a target constituent in the derivation structure.

```
<edge id="e001" source="n019" target="n003" type="adj" label="incise" >
      <deriv names="d000015" source_op="o8" target_op="o21" span="0 2 0 1"/>
</edge>
```

Finally, the forest is disambiguated by an heuristic-based module that outputs a single dependency tree. Normally, The parser tries to find *complete parses* covering the full sentence. However, in cases of failure, the parser switches to *partial parsing* to retrieve the best sets of partial parses covering the sentence.

The output of the parsing is then enriched by a series of post-processing modules whose role is to organize all of the information retrieved along the whole linguistic processing. An example of output ambiguous and disambiguated dependency graphs is presented in figure 7.4.

As mentioned above, Tree Adjoining Grammars essentially rely on two types of operations on elementary trees [Joshi et al., 1975]. *Substitution* replaces a non-terminal leaf node by a tree and is mostly used to handle arguments, for instance the verbal arguments, dependency labels associated to substitution generally reflect syntactic functions, such as subject or object. *Adjunction*, adjoins an auxiliary tree around a (possibly internal) node. It is mostly used to handle non-essential modifiers, for instance an adjective, an adverb or a subordinate clause. In fact, the grammar has a specific notion of x-modifier, handled by adjunction, with optional or strict parenthesizing of the modifier with commas, dashes or parentheses.

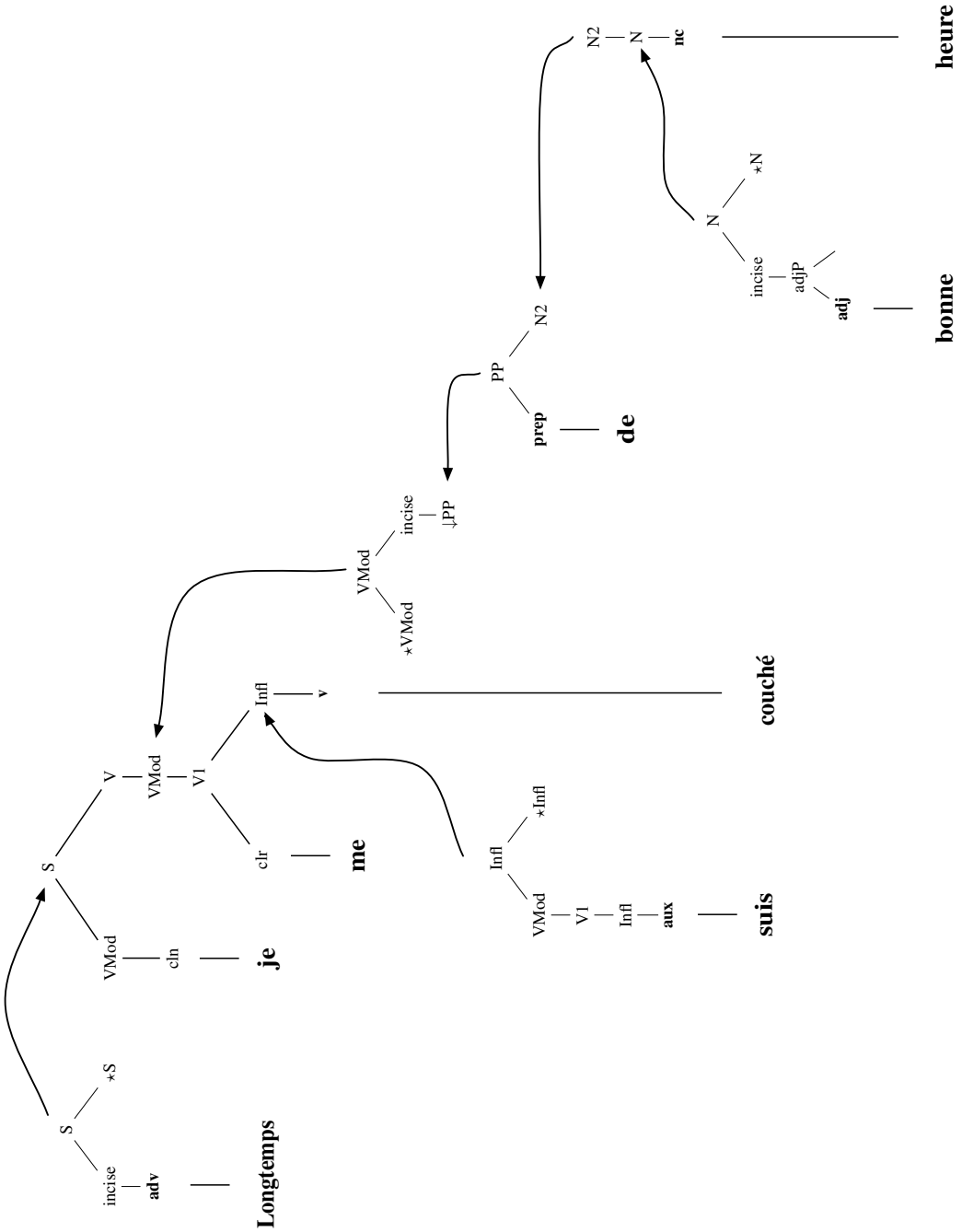


Figure 7.3: Derived tree for the sentence: "*Longtemps, je me suis couché de bonne heure.*" ("*For a long time I used to go to bed early*").

7.2.3 Reliability of the Syntactic Analysis

The FRMG parser has participated in two campaigns for the evaluation of French syntactic parsers: EASY [Paroubek et al., 2008] and PASSAGE [Villemonte de La Clergerie et al., 2008]. Some of the performances of the FRMG syntactic parser are reported in tables 7.1, 7.2, and 7.3. The ambiguity rate of a sentence reflects the average number of edges entering a form, minus 1. Thus, a non-ambiguous sentence has a null ambiguity rate.

text corpus	#sentence	ambiguity	coverage (%)
EUROTRA	334	0.81	100
TSNLP	1161	0.48	95.18
EasyDev	3879	1.10	69.01
JRCacquis	1.1M	1.10	51.26
Europarl	0.8M	1.36	70.19
EstRep	1.6M	0.92	67.05
Wikipedia	2.2M	0.87	69.11
Wikisource	1.5M	0.89	61.08
AFP	1.6M	1.06	52.15

Table 7.1: Coverage and parsing ambiguity of FRMG obtained for different text databases.

evaluation	chunk F-measure	dependency F-measure
2004	69	41
2007	89	63

Table 7.2: Performance of FRMG on the EASY treebank.

text corpus	mode	chunk F-measure	dependency F-measure
EasyRef	full	91.90	69.15
	partial	83.70	57.52

Table 7.3: Performance of FRMG in case of full and partial parsing.

The rest of the chapter is dedicated to the integration of deep syntactic analysis as provide by the syntactic parser into the context-dependent analysis.

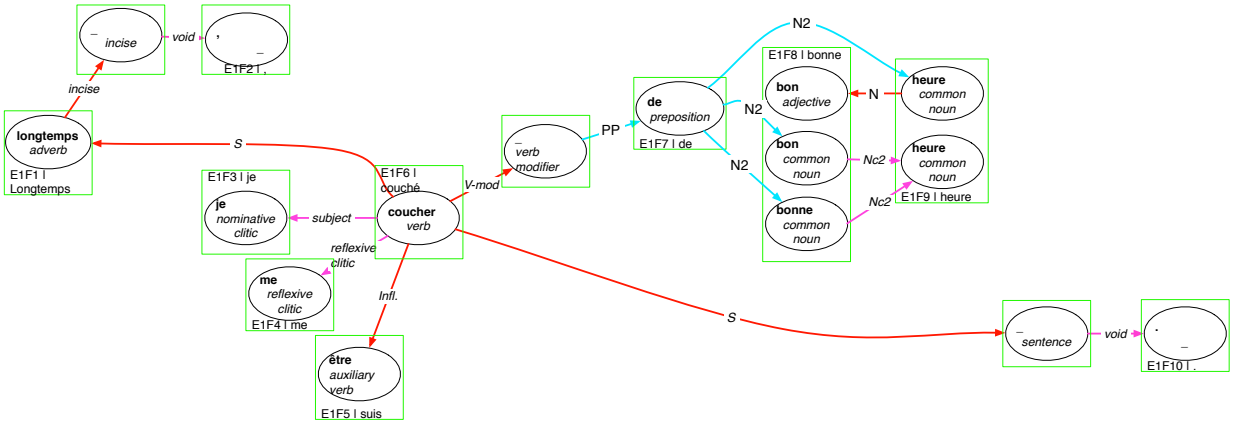
7.2.4 Syntactic Analysis of the Text Material

The syntactic parsing of the speech databases used in this study is summarized in table 7.4.

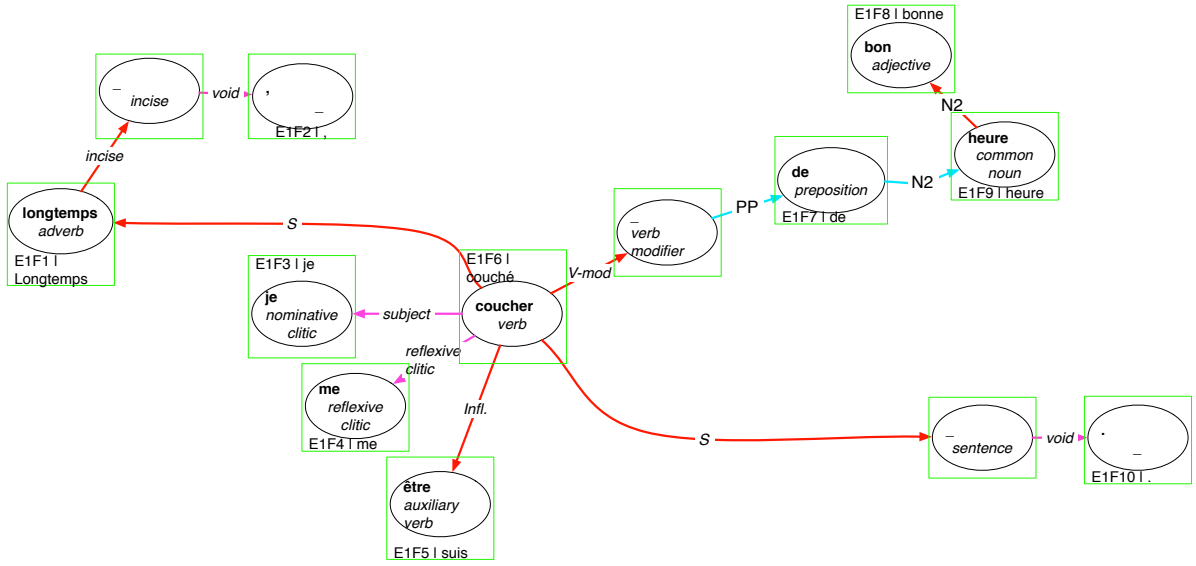
text corpus		laboratory	multi-media
segmentation	sentence	9281	2184
	form	89321	52358
	form/sentence	**9.5**	**27**
parsing	coverage (%)	**80**	**52**
	ambiguity	**0.69**	**1.2**

Table 7.4: Description of the text segmentation and syntactic parsing for the text material used.

The syntactic analysis confirms evidence for the relative simplicity of the laboratory text and the complexity of the multi-media text. Firstly, complete parsing was achieved for 80% of the parsed sentences of the laboratory text while only for 52% of the multi-media text. Secondly, the laboratory text has a 0.69 ambiguity rate only, while the multi-media text has a 1.2 ambiguity rate, which locates the text used in extreme position compared with other text databases (minimum = 0.48, maximum = 1.36, and mean ambiguity = 0.95, figure 7.1).



(a) ambiguous graph



(b) disambiguated graph

Figure 7.4: Dependency graph for the sentence: "*Longtemps, je me suis couché de bonne heure.*" ("*For a long time I used to go to bed early*"). Squares represent *clusters* attached to a token and carrying a form, circles represent *nodes* associated with a lemma and a morpho-syntactic category, and *edges* represent *dependencies* that connect a source node (*governor*) to a target node (*governee*), and associated with the label of the dependency and the type of the TAG operation (edge color).

7.3 Extraction of Rich Syntactic Features

The automatic linguistic chain presented in section 7.2 combines a surface and deep syntactic analysis of a given sentence. Contrary to other syntactic parsers that exclusively assume one or the other representation, the dependency trees provided by FRMG also include information about constituency (category and span of maximal syntactic constituents such as nominal phrases or clauses). Since there is no evidence that speech prosody depends rather on constituency or on dependency, this representation presents the advantage of a unified syntactic description that will be used to compare their relevancy for the modelling of speech prosody in chapters 8 and 9). Additionally, while the substitution operation may slightly relate to speech prosody, some specific adjunction operations may be relevant syntactic cues of speech prosody variations (e.g., relative clauses, incises). Thus, syntactic adjunctions will be specifically studied and evaluated in speech prosody modelling.

In this section, the extraction of the syntactic features that will be used for the speech prosody modelling is presented. The extracted syntactic features are classified into different sets according to the nature of the syntactic information. The feature sets are composed of the three main syntactic classes: morpho-syntactic (section 7.3.1), dependency (section 7.3.2), and constituency (section 7.3.3). An additional feature set that covers adjunctions is introduced and discussed (section 7.3.4). The first features set is retrieved from surface parsing while the others are extracted from deep parsing. Following is a description of the different syntactic feature sets extracted from the linguistic parser. An exhaustive description of the features used is presented in appendix 13.2.

7.3.1 Form

Morphological and morpho-syntactic form features are extracted from the surface parsing (figure 7.5).

- form segment;
- form *morpho-syntactic category* (Part-Of-Speech)
- form *morpho-syntactic class*: function vs. content form;

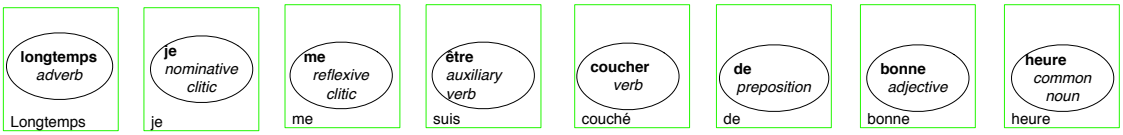


Figure 7.5: Representation of morpho-syntactic features for the sentence: "*Longtemps, je me suis couché de bonne heure.*" ("*For a long time I used to go to bed early*"): form, lemma, and morpho-syntactic category.

7.3.2 Dependency

Form dependencies are extracted from the deep parsing and the derived dependency graph (figure 7.6).

- {governor, current, governee} form *morpho-syntactic category* and *class*;
- *edge type* and *label* between current form and {governor, governee} form;
- *signed dependency distance* between current form and {governor, governee} forms (in forms and in chunks);

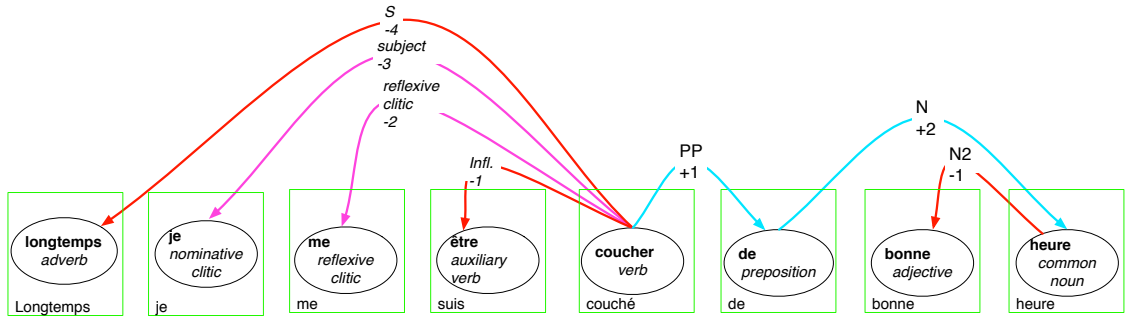


Figure 7.6: Representation of morpho-syntactic and dependency features for the sentence: "*Longtemps, je me suis couché de bonne heure.*" ("*For a long time I used to go to bed early*"): form category, edge label, edge type, and edge distance in forms.

7.3.3 Constituency

Constituent structure is not directly represented in the output of the syntactic parser. Nevertheless, constituent structure is retrieved in a top-down process from the operations and associated derivations. Constituent structure is initialized with root operations, i.e. operations that are not derived from any operation. This coincides with the complete sentence in case of complete parsing and with a set of partial elements that cover the complete sentence in case of partial parsing. Then, the complete constituent structure is iteratively retrieved from successive derivations conducted from the initial constituents.
As a constituent can be associated with an arbitrary number of derivations, constituents are stacked from left to right in order to provide a binary constituent tree representation. This is achieved in order to transform the original constituent tree into a constituent binary tree and thus providing a more convenient representation for representing the constituent dominances. Finally, terminal constituents are converted into ϕ-phrases, by merging each of the terminal heads with all of its specifiers [Selrik, 1981].

For the sentence: "*Longtemps, je me suis couché de bonne heure.*" ("*For a long time I used to go to bed early*"), the complete constituent structure is:

$$(S\ (AdvP\ \text{Longtemps}\ )\ (\ (VP\ \text{je me suis couché}\ )\ (NP\ \text{de bonne heure}\ )\ )\ )$$

where S, $AdvP$, VP and NP denote respectively sentence, adverbial, verbal and nominal phrases.

Finally, the constituent structure is converted into a constituent sequence (chunks) that is simply retrieved from the terminal constituents of the constituent tree. However, the chunk sequence do not account for the hierarchical organization of the constituent structure, e.g., intermediate constituents and derivations. Moreover, constituent recursion prohibits the inheritance of an information associated with an intermediate constituent onto its derived chunks. Consequently, the constituent structure will remain partially described through the inheritance of non-recursive information and direct dependencies. Finally, the constituent structure is represented by the chunk sequence associated with a partial description of the constituent structure.

The following constituency features are extracted (figure 7.7):

- form maximal syntactic category;
- {governor, current, governee} *chunk category*;
- *edge type* and *label* between the current chunk and {governor, governee} chunks;

- *signed dependency distance* between the current chunk and {governor, governee} chunks (in forms and in chunks);
- *chunk depth* in the constituent tree;
- *chunk depth difference* between the current chunk and the next chunk;

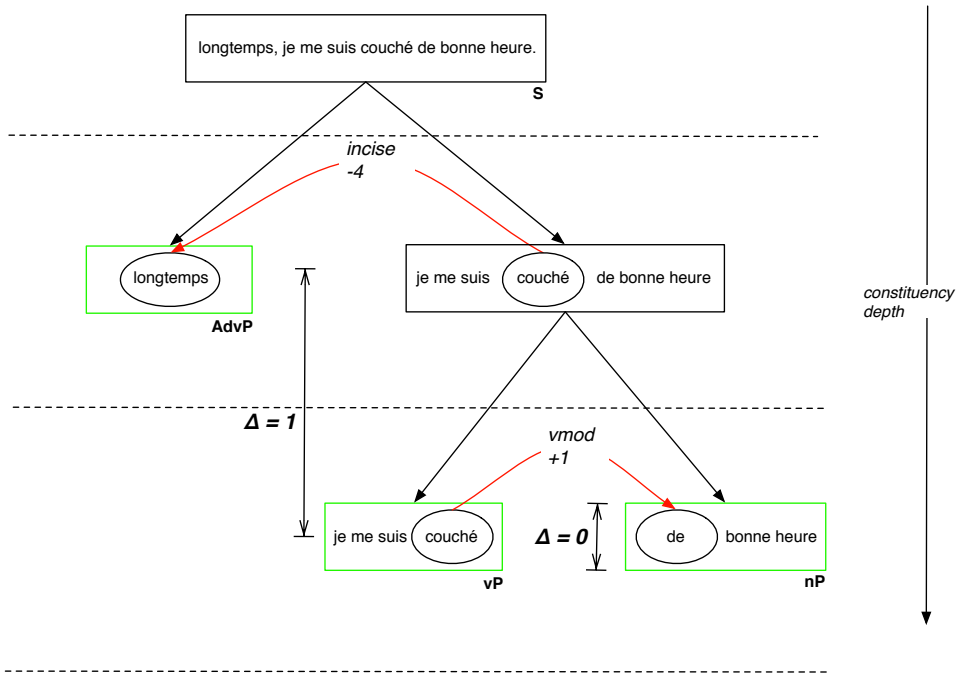


Figure 7.7: Representation of constituency features for the sentence: "*Longtemps, je me suis couché de bonne heure.*" ("*For a long time I used to go to bed early*"): intermediate constituents (black box), terminal constituents (green box), constituent category, chunk dependency (red edge), edge type, edge label, and dependency distance in forms, constituent depth, and chunk depth difference.

7.3.4 Adjunction

Adjunction refers to a specific syntactic operation in the Tree Adjoining Grammar formalism. Adjunction is particularly expressive in the sense that adjunction can derive complex linguistic structures and covers a large amount of various linguistic constructions as observed in natural language, from a single form adjunction (e.g., adverbial or adjectival adjunction), to complete adjunction structure (e.g., clauses), and even to complex adjunction structure (e.g., embedded clauses). Interestingly, adjunction covers a large amount of syntactic constructions - such as incises, parentheses, subordinate and coordinate clauses - that may be relevant for the modelling of speech prosody[3].

Adjunctions can be easily extracted according to specific pattern matching in the syntactic parser formalism (figure 7.8), with some variations depending on the nature of the adjunction. Full adjunction is then extracted by retrieving the full dependency descendency from the introducer. Moreover, the specific type of an adjunction can be identified from the governee, introducer, and governor category, and edges type and label. For instance, a relative clause is defined as a phrase that modifies a noun as represented in figure 7.8. In the studied material, 105 and 151 different types of adjunction were respectively observed in the laboratory and the multi-media texts. A reduced description of some of the most known adjunction types and their occurrence frequency

[3]enumerations have been added to the conventional adjunctions to cover a large variety of syntactic constructions.

is illustrated in figure 7.10 for a comparison of the laboratory and the multi-media texts.

Figure 7.8: Generic adjunction pattern: M is the governor node, N the governee node, I the introducer.

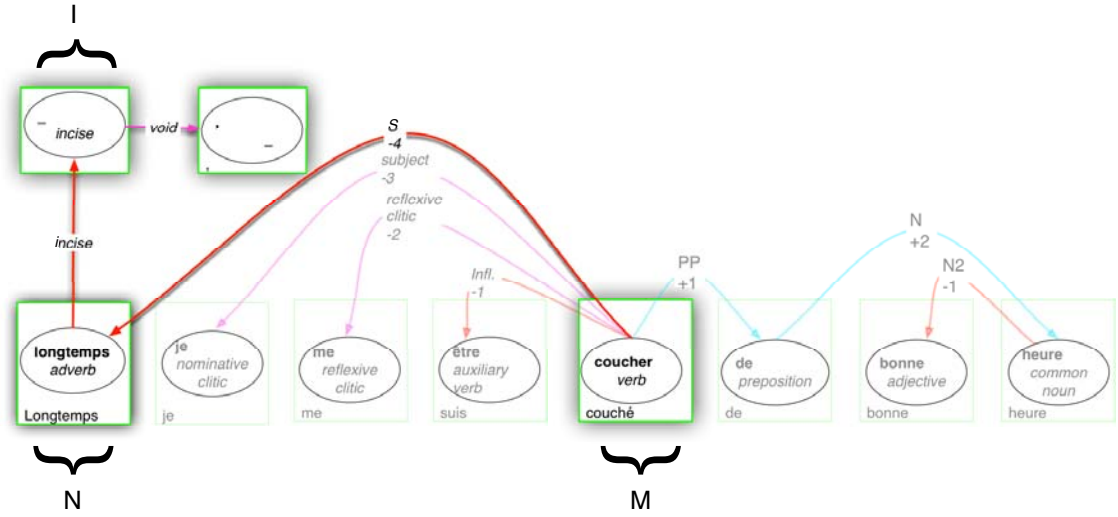


Figure 7.9: Representation of adjunction features for the sentence: "*Longtemps, je me suis couché de bonne heure.*" ("*For a long time I used to go to bed early*") for which "*longtemps*" is a left incise adjunction on the verb of the sentence "*couché*". In such an example, M ("*couché*") is the governor node, N ("*longtemps*") is the governee node, I (empty) is the introducer, both dependency types are *adjunction* (red edges), dependency labels are *incise* and *sentence*, and dependency distance in forms is *-4*.

Additionally, adjunction has the recursive property: a given adjunction can be embedded within another adjunction. Thus, in the case of recursion, only the adjunction with the larger span is extracted.

The following adjunction features are extracted (figure 7.9):

- {governor, introducer, governee} form category;
- *edge type and label* between governor and introducer nodes and between introducer and governee nodes;
- *signed dependency distance* between the adjunction's introducer and the governor node (in forms and in chunks);

A descriptive analysis of adjunction occurrence in the texts used (figure 7.10) confirms evidence for their linguistic complexity: the multi-media text presents a large variety and amount of adjunctions compared the laboratory text.

Syntactic features extracted from the text analysis will further be used for the context-dependent modelling of speech prosody. The extracted syntactic units and syntactic contexts are summarized in tables 7.5 and 7.6.

7.4 Conclusion

In this chapter, an automatic linguistic processing chain was presented and described in order to enrich the linguistic description of a text for the modelling of speech prosody. The linguistic processing chain includes text preprocessing, surface parsing, and deep parsing. A preprocessing is achieved in order to segment a raw text into linguistic units that can be used by a linguistic parser (segmentation into sentences and forms). Surface parsing is used to provide a morpho-syntactic description of a sentence. Then, deep parsing is achieved based on Tree Adjoining Grammar (TAG), and used to represent both dependency graph and constituency structure derived from a sentence. The extraction of syntactic features from the linguistic analysis was presented. The extracted syntactic features are classified into different sets depending on their nature: morpho-syntactic features are extracted from the surface parsing, dependency and constituency features are extracted from deep parsing, and adjunction features are additionally introduced which are retrieved from deep parsing. The linguistic processing chain presented provides an enriched description of the text characteristics that will be further used to refine the context-dependent modelling of speech prosody. In particular, the relevancy of the syntactic characteristics will be compared and discussed for the symbolic and acoustic modelling of speech prosody in chapters 8 and 9.

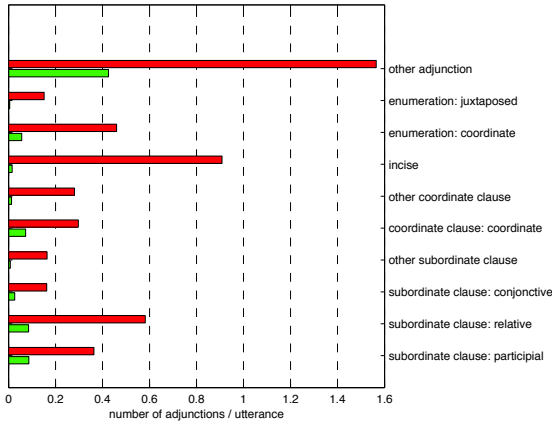


Figure 7.10: Occurrence of different types of adjunction in the laboratory (green) and multi-media (red) speech databases.

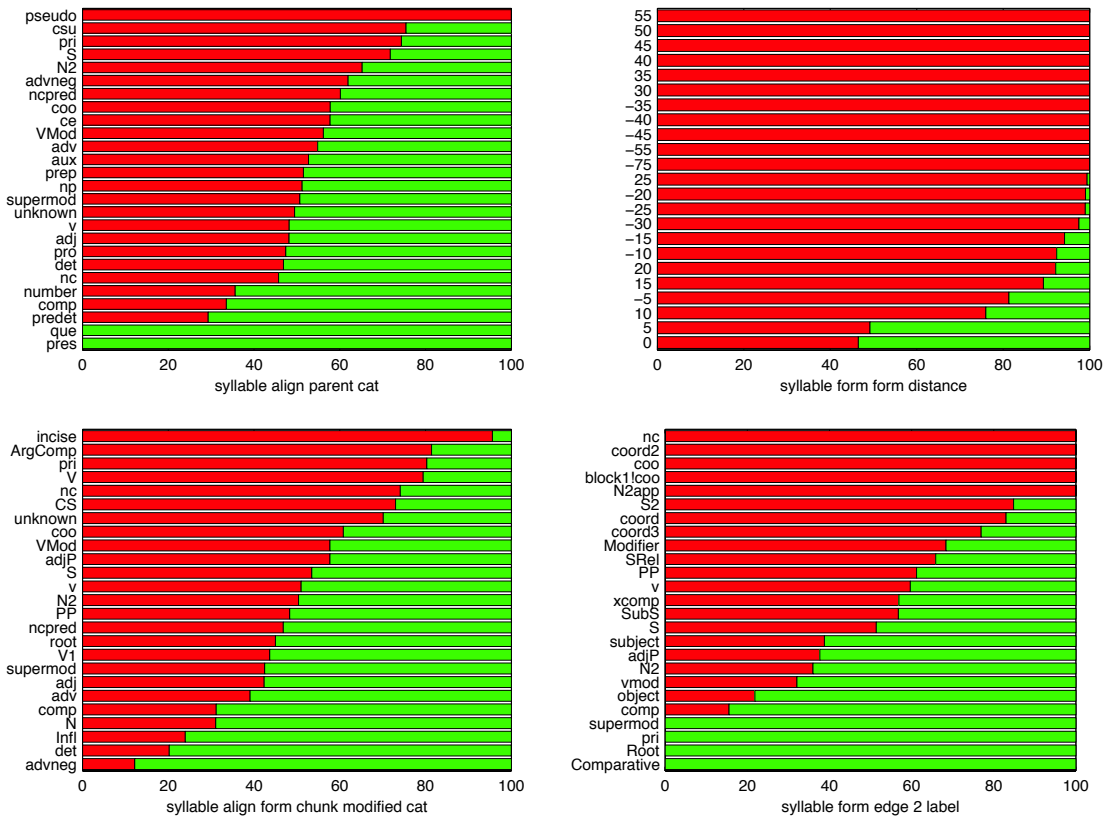


Figure 7.11: Comparative proportion of syntactic characteristics in the laboratory (green) and multi-media (red) speech databases: form lexical category, form dependency distance, governee constituent lexical category, adjunction operation.

	feature set	feature
syntactic contexts	parsing status	parsing status
	form	morpho-syntactic category
		morpho-syntactic class
	dependency	{governor, current, governee} form morpho-syntactic category
		{governor, current, governee} form morpho-syntactic class
		edge type between current form and {governor, governee} form
		edge label between current form and {governor, governee} form
		signed dependency distance between current form and {governor, governee} forms
	constituency	form maximal constituent category
		{governor, current, governee} chunk category
		edge type between the current chunk and {governor, governee} chunks
		edge type between the current chunk and {governor, governee} chunks
		edge label between current form and {governor, governee} form
		chunk depth in the constituent tree
		chunk depth difference between the current chunk and the next chunk
	adjunction	{governor, introducer, governee} form category
		edge type between governor and introducer
		edge label between governor and introducer
		edge type between introducer and governee
		edge label between introducer and governee
		signed dependency distance between introducer and governor

Table 7.5: Description of syntactic features.

	feature set	unit	parent unit
syntactic unit	form	form	sentence
	chunk	form	chunk
		chunk	sentence
	adjunction	form	adjunction
		chunk	adjunction
		adjunction	sentence
prosodic/syntactic unit	form	syllable	form
		form	minor prosodic group
		form	major prosodic group
	chunk	syllable	chunk
		chunk	minor prosodic group
		chunk	major prosodic group
	adjunction	syllable	adjunction
		minor prosodic group	adjunction
		adjunction	major prosodic group

Table 7.6: Description of syntactic / prosodic units.

Chapter 8

Discrete Modelling of Speech Prosody

Contents

8.1 Introduction

The prosodic structure corresponds to the symbolic description of speech prosody, i.e. the description and organization of relevant speech prosody events, in relation to their formal and functional dimensions. In oral language, prosodic markers are produced by speakers and are used by listeners to clarify the meaning and the structure of an utterance or a discourse. In particular, prosodic markers instantiate prosodic domains (e.g., prosodic prominence, prosodic groups) that are associated with various functions (such as focus, segmentation, and hierarchization). Theoretical models have been proposed for the formal representation and the description of abstract prosodic events and their structure: ToBI for English [Silverman et al., 1992]; INTSINT [Hirst et al., 2000], IVTS [Post, 2000], and Prosogram [Mertens, 2004a] for French. Phonological models hypothesize that there exist prosodic markers that correspond to acoustic salience, and associated with specific contours and specific communicative functions. While the representation differs significantly from one model to another, phonological models generally assume that prosodic markers divide into *prosodic prominence* and *prosodic boundaries*, which are basically associated with accentuation and grouping.
From a computational standpoint, the symbolic description of speech prosody provides a high-level control in speech synthesis. Indeed, a variation in the prosodic structure would introduce a significant change in the speech prosody, while the possible acoustic alternatives are highly constrained by a prosodic structure. In particular, a change in the prosodic structure may affect the whole organization of speech prosody, from symbolic to acoustic characteristics. For instance, the insertion of a pause will be perceived as a significant change in speech, especially when this insertion comes with a modification of the prosodic contour that precedes and/or follows the pause ([Martin, 1987, Martin, 2010]: *melodic slope inversion*, or [House, 1995]).

sentence	Longtemps, je me suis couché de bonne heure. *For a long time I used to go to bed early.*
prosodic variation 1	Longtemps / je me suis couché / de bonne heure //
prosodic variation 2	Longtemps // je me suis couché / de bonne heure //
prosodic variation 3	Longtemps // je me suis couché / de bonne heure //

Table 8.1: A study case of prosodic alternatives. / and // denote minor and major prosodic boundaries, respectively.

Two main approaches can be distinguished for the symbolic modelling of speech prosody: on the one hand, *expert approaches* attempt to develop formal models that account for the observed speech prosody characteristics with respect to linguistic, para-linguistic, and extra-linguistic constraints. On the other hand, *statistical methods* attempt to develop a statistical model that accounts for the speech prosody characteristics based on the observation of statistical regularities in large speech databases.

Expert approaches mostly concern the description of the hierarchical organization of speech prosody, and in particular prosodic boundaries ([Cooper and Paccia-Cooper, 1980, Gee and Grosjean, 1983, Selrik, 1984, Ferreira, 1988, Abney, 1992, Watson and Gibson, 2004] for English; [Dell, 1984, Bailly, 1989, Monnin and Grosjean, 1993, Ladd, 1996, Delais-Roussarie, 2000, Mertens, 2004b] for French, [Barbosa, 2006] for some other languages). Expert models assume that a prosodic structure results from the integration of various and potentially conflicting constraints, in particular *syntactic* and *rhythmic* constraints. A prosodic structure is primarily produced by speakers and can be used by listeners to clarify the structure of the utterance, and in particular its syntactic structure. Simultaneously, secondary cognitive constraints tend to produce a prosodic structure with an optimal configuration, in particular with respect to rhythmic regularity [Fraisse, 1974, Dell, 1984]. These constraints conflict in the production of a prosodic structure, and secondary extra-linguistic constraints often override the primary linguistic constraint. Consequently, expert approaches aim at developing a set of formal rules that relates the presence of a prosodic boundary to some specific syntactic cues and with respect to performance constraints. The precise set of formal constraints and their combination are heuristically formulated based on *a priori* expert knowledge and the observation of a limited set of linguistic productions.

Formally, each form of a sentence is associated with the *likelihood* (or the *strength*) that a prosodic boundary exists between this form and the following. The likelihood or the strength of a prosodic boundary is estimated according to separated processing: a syntactic module and an optional rhythmic module.
The linguistic module mostly concerns the extraction of prominent syntactic boundaries from deep syntactic parsing, based on syntactic constituency (*Constituent-Depth* [Cooper and Paccia-Cooper, 1980], *ϕ-phrases* [Gee and Grosjean, 1983, Delais-Roussarie, 2000], *Left-hand-side / Right-hand-side Boundary* [Watson and Gibson, 2004]), syntactic dependency (*Dependency-Grammar-based local markers* [Bailly, 1989, Barbosa, 2006]), or a combination (*Chunks-and-Dependencies* [Abney, 1992]). Some studies attempt to integrate higher-level linguistic constraints, such as syntactic-semantic constraints - defined in terms of the degree of syntactic dependency across successive syntactic constituents [Ferreira, 1988]. A score is associated with each of the syntactic cues considered, and combined to determine the likelihood or the strength of a prosodic boundary conditionally to the observed syntactic cues.
The rhythmic module is used as a regularization process to adjust the produced prosodic structure with respect to the size of the prosodic constituent candidates [Gee and Grosjean, 1983, Bailly, 1989, Delais-Roussarie, 2000, Barbosa, 2006], either in parallel or in cascade with the linguistic module.
Finally, expert approaches provide a valuable formal framework for the symbolic description of speech prosody, in which the constraints that interact in the production of a prosodic structure and the linguistic cues that may be associated with prosodic boundaries are explicitly formulated. However, expert approaches design a *universal* model in which general principles that can be observed across speakers of a language are explicitly formulated. Consequently, expert models do not account for the characteristics associated with a specific speaker or speaking style, and cannot simply be adapted to a specific speaker or a specific speaking style. Nevertheless, recent studies attempt to extend expert models so as to account for the individual strategies of a speaker [Barbosa, 2006]. Such an approach combines expert and statistical approaches: relevant characteristics are formulated based on expert knowledge, and their relative importance is statistically adapted to a specific speaker.

Statistical approaches aim at developing a statistical model that accounts for the statistical dependencies that relate speech prosody characteristics to linguistic cues from the observation of their relative co-occurrence in large speech databases. Formally, statistical models determine the likelihood of a prosodic structure conditionally to the characteristics of a text. In a similar manner as for expert models, statistical approaches mostly concern the modelling of prosodic boundaries[1]. Then, for a given sentence, the statistical model determined the *likelihood* that a

[1]In particular, lexical stress is imposed in stress-based languages such as English and does not require statis-

prosodic boundary exists at the juncture of successive forms.

Statistical approaches mainly divide into *static* (Decision-Tree-Based [Hirschberg, 1991, Black and Taylor, 1994]), *dynamic* (HMM-based [Veilleux et al., 1990, Ross and Ostendorf, 1996, Black and Taylor, 1997a, Sun and Applebaum, 2001, Atterer and Klein, 2002, Schmid and Atterer, 2004, Bonafonte and Agüero, 2004, Bell et al., 2006]), and *hierarchical* (Hierarchical HMM, Weighted Tree Automata [Ostendorf and Veilleux, 1994, Rangarajan Sridhar et al., 2008]) methods.

static methods model the likelihood of a prosodic structure given the observed linguistic information, alone [Hirschberg, 1991, Black and Taylor, 1994].

dynamic methods additionally regularize the likelihood of a prosodic structure given the observed linguistic information with that of the prosodic structure.

Additionally, statistical models differ in their representation of the prosodic structure:

sequential methods assume the prosodic structure as a sequential structure [Veilleux et al., 1990, Ross and Ostendorf, 1996, Black and Taylor, 1997a, Schmid and Atterer, 2004, Bonafonte and Agüero, 2004, Sun and Applebaum, 2001];

hierarchical methods explicitly model the prosodic structure as a hierarchical structure [Ostendorf and Veilleux, 1994, Rangarajan Sridhar et al., 2008].

Linguistic dependencies used to be statistically modelled based on syntactic information extracted from surface parsing, such as lexical category (Part-Of-Speech, or POS), or lexical class (content and function forms) of a form, and punctuation markers. Very few studies exist on the integration of a rich syntactic description for the modelling of speech prosody, without significant improvements [Ingulfen et al., 2005].

Statistical models present various advantages over expert models. Firstly, parametric models can adequately model and adapt to the characteristics of a speaker and/or a speaking style. Secondly, statistical models can accurately model various and complex characteristics to a degree and in a time that would be unattainable by an expert. However, the syntactic characteristics extracted and their complex combinations paradoxically remain relatively low-level information. First, the surface syntactic description that is generally used in statistical modelling is slightly related to the prosodic structure, while theoretical studies have pointed out that a prosodic structure consistently relates to the deep syntactic structure of a sentence. Additionally, experts use *a priori* knowledge to describe relevant syntactic cues while statistical models may fail to retrieve them automatically from the very large amount of possible combinations.

Finally, expert and statistical models are not opposed to each other but benefit from their mutual advances: statistical models are introduced into expert models ([Barbosa, 2006]), and statistical models benefit from expert knowledge. In particular, recent statistical models have been proposed to account explicitly for rhythmic constraints (segmental models [Ostendorf and Veilleux, 1994, Schmid and Atterer, 2004, Bell et al., 2006]), or for the hierarchical modelling of speech prosody [Ostendorf and Veilleux, 1994, Rangarajan Sridhar et al., 2008].

In this chapter, a *linguistic-oriented* approach is proposed to integrate a rich linguistic description into statistical modelling, and to combine theoretical linguistic with statistical methods. A discrete HMM is presented in which the symbolic characteristics of speech prosody are modelled conditionally to the linguistic context in which they are observed. The prosodic grammar used consists of the Rhapsodie representation that was experimented as an alternative to ToBI [Silverman et al., 1992] for the transcription of French prosody [Lacheret et al., 2010] (chapter 3). A context-dependent discrete HMM is used to model the symbolic characteristics of speech

tical modelling. Additionally, residual prosodic prominences such as prosodic focus are not used to be modelled. Consequently, most of the studies focus on major prosodic boundary (or prosodic break) modelling.

prosody in context. During the training, the text is first converted into a sequence of linguistic contexts using the linguistic processing chain described in chapter 7 that includes surface and deep syntactic parsing. Linguistic contexts are clustered using a Decision-Tree so as to minimize the entropy of the prosodic events. Then, a discrete HMM is estimated for each terminal node of the context-dependent tree. During the synthesis, the text is first converted into a sequence of concatenated context-dependent models. Then, the sequence of prosodic events is determined so as to maximize the conditional probability of the sequence of prosodic events given the sequence of linguistic contexts and the models.

This chapter is organized as follows: firstly, the role of the linguistic context in the modelling of speech prosody is assessed using a conventional context-dependent discrete HMM in section 8.2. Secondly, a method that combines linguistic and metric constraints for the modelling of prosodic break is proposed in section 8.3 based on segmental HMMs and Dempster-Shafer fusion, and the relative importance of the linguistic and the metric constraints is assessed depending on the nature of the linguistic information. Finally, a method to vary speech prosody of a speaker based on the *General Viterbi Algorithm* (GVA) is proposed in section 8.4). The proposed methods are either objectively and/or subjectively evaluated.

8.2 Context-Dependent Discrete HMM

8.2.1 Transcription of Speech Prosody

The proposed symbolic modelling of speech prosody is a context-dependent discrete HMM using the Rhapsodie transcription described in chapter 3.

The symbolic grammar is composed of:

major prosodic boundary : F_M, a prosodic boundary which is followed by a pause. A major prosodic boundary is associated with a major prosodic group (MPG);

minor prosodic boundary : F_m, a prosodic boundary which is internal to the prosodic group. A minor prosodic boundary is associated with an minor prosodic group (mPG)

residual prosodic prominence : P, a residual prosodic prominence which mostly relates to semantic and/or discursive focus.

Speech prosody is automatically labelled based on Analor [Avanzi et al., 2008] and ircamProm [Obin et al., 2008c], then converted into a sequence of prosodic events, and represented over the syllable to account for all of the prosodic events simultaneously.

8.2.2 CART Decision-Tree Context-Clustering

As mentioned in section 2, many linguistic or para-linguistic factors affects speech prosody variations. In particular, speech prosody is strongly related to the linguistic structure of an utterance: syntactic structure and prosodic structure affects macro-prosodic variations, while phonemic content affect micro-prosodic variations. Context-dependent models are commonly used to describe the statistical speech characteristics conditionally to a specific linguistic context. However, speech prosody implies various linguistic units (e.g., prosodic, syntactic) each associated with various characteristics. Consequently, an extensive coverage of each of the linguistic context appears unrealistic, especially when providing a rich description of the linguistic characteristics of a text (chapter 7). Additionally, a well-balanced coverage of the observed linguistic contexts cannot be reached. Finally, these problems appear particularly significant in the case of real-world speech databases. A number of methods have been proposed to cluster context-dependent HMM models and share model parameters among linguistic contexts [O'Dell, 1995, Yoshimura et al., 1999, Shinoda and Watanabe, 2000]. In this section, a conventional decision-tree-based context-clustering method based on Classification And Regression Trees

(CART) is presented to cluster linguistic contexts prior to HMM modelling.

In the following, the principles of the Classification and Regression Trees (CART) [Breiman et al., 1984] are shortly described.

Let $\mathbf{s} = [s_1, \ldots, s_N]$ be a sequence of syllables associated with an utterance U. Let $\mathbf{q} = [\mathbf{q}_1, \ldots, \mathbf{q}_N]$ be the sequence of linguistic contexts, where $\mathbf{q}_n = [q_n(1), \ldots, q_n(L)]^\top$ is a (Lx1) linguistic vector which describes the linguistic property associated with syllable s_n. Let $\mathbf{l} = [l_1, \ldots, l_N]$ be the sequence of prosodic events associated with syllable s_n.

Let T be a binary tree with root node S_0 and leaf nodes $\mathbf{S} = (S_1, \ldots, S_M)$, where M is the number of leaf nodes.

Let $\mathrm{E}(S_m)$ denotes the information entropy of the node S_m given the prosodic events L_m associated with the linguistic contexts corresponding to the node S_m.

The information entropy of the tree T is given by:

$$\mathrm{E}(S) = \sum_{m=1}^{M} \mathrm{E}(S_m) \tag{8.1}$$

where $\mathrm{E}(S_m)$ is the information entropy that corresponds to the node S_m

$$\mathrm{E}(S_m) = -\sum_{n=1}^{N} \mathrm{p}(q_n) \log_2 \mathrm{p}(q_n) \tag{8.2}$$

The change in information entropy $\mathrm{E}(S')$ by splitting leaf node S_m through question q into nodes $S_{m,q+}$ and $S_{m,q-}$ is given by:

$$\Delta_{\mathrm{E}}^{q}(S') \quad = \quad \mathrm{E}(S_{m,q+}) + \mathrm{E}(S_{m,q-}) - \mathrm{E}(S_m) \tag{8.3}$$

The question $\widehat{q}_{\mathrm{E}}$ that maximizes the increase of the information gain at node S_m is given by:

$$\widehat{q}_{\mathrm{E}} = \underset{\mathbf{q}}{\operatorname{argmax}} \; -\Delta_{\mathrm{E}}^{q}(S) \tag{8.4}$$

The context-dependent tree is then derived as follows:

1. **tree initialization**

$$\begin{aligned} \mathrm{T}^{(0)} &= \mathrm{T}_0 \\ S^{(0)} &= S_0 \end{aligned} \tag{8.5}$$

2. **tree recursion**

   for each leaf node S_m of the context-tree $\mathrm{T}^{(i)}$

   tree selection

   (a) information gain calculation:$\Delta\ {}_{\mathrm{E}}^{q}(S)$, $q \in [1, Q]$

   (b) optimal splitting context: $\widehat{q}_{\mathrm{E}} = \underset{\mathbf{q}}{\operatorname{argmax}} \; -\Delta_{\mathrm{E}}^{q}(S)$

   tree derivation

$$S'_m \quad \leftarrow (S_{m,\widehat{q}-}, S_{m,\widehat{q}+}) \tag{8.6}$$

tree update

$$\begin{aligned} \mathrm{T}^{(i+1)} &= \mathrm{T}' \\ S^{(i+1)} &= S' \end{aligned} \tag{8.7}$$

3. **tree termination**

$$\begin{aligned} \widehat{\mathrm{T}} &= \mathrm{T}^{(i)} \\ \widehat{S} &= S^{(i)} \end{aligned} \tag{8.8}$$

8.2.3 Parameters Estimation

During the training, linguistic contexts are first clustered so as to derive a context-dependent tree.

Then, a context-dependent HMM model $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_{S_1}, \ldots, \boldsymbol{\lambda}_{S_M})$ is constructed from the set of terminal contexts $S = (S_1, \ldots, S_M)$ of the decision-tree, where $\boldsymbol{\lambda}_{S_m} = (\boldsymbol{\Pi}_{S_m}, \mathbf{A}_{S_m}, \mathbf{B}_{S_m})$ denotes the estimated HMM parameters associated with the context S_m.

There is a certain inconsistency in the proposed approach since the criterion used for the context-clustering differs from the criterion used for the modelling. Nevertheless, the proposed model will be used as a first approximation in this study. In further studies, HMM modelling prior to the context-dependent modelling, and adequate context-clustering methods based on Maximum-Likelihood [O'Dell, 1995, Yoshimura et al., 1999, Shinoda and Watanabe, 2000] will be used to consistently derive the context-dependent model.

8.2.4 Parameters Inference

During the inference, the text is first converted into a linguistic context sequence $\mathbf{q} = [\mathbf{q}_1, ..., \mathbf{q}_N]$, where $\mathbf{q}_n = [q_n(1), \ldots, q_n(L)]^\top$ is a (Lx1) context vector which describes the linguistic properties associated with syllable s_n.

Then, the optimal sequence of prosodic events is determined so as to maximize the probability of the sequence of prosodic events $\mathbf{l} = [l_1, ..., l_N]$, conditionally to the linguistic context sequence $\mathbf{q}$ and the model $\boldsymbol{\lambda}$:

$$\widehat{\mathbf{l}} = \underset{\mathbf{l}}{\operatorname{argmax}} \left(\mathrm{p}(\mathbf{l}|\mathbf{q}, \boldsymbol{\lambda})\right) \tag{8.9}$$

$$= \underset{\mathbf{l}}{\operatorname{argmax}}\, \mathrm{p}(l_1)\mathrm{p}(\mathbf{q}_1|l_1, \boldsymbol{\lambda}) \times \prod_{n=2}^{N} \mathrm{p}(\mathbf{q}_n|l_n, \boldsymbol{\lambda})\mathrm{p}(l_n|l_{n-1}) \tag{8.10}$$

Using Bayes' theorem,

$$\mathrm{p}(l_n|\mathbf{q}_n, \boldsymbol{\lambda}) = \frac{\mathrm{p}(\mathbf{q}_n|l_n, \boldsymbol{\lambda})\ \mathrm{p}(\mathbf{q}_n|\boldsymbol{\lambda})}{\mathrm{p}(l_n|\boldsymbol{\lambda})} \tag{8.11}$$

Hence, assuming that the probability of the linguistic context sequence is constant during the maximization:

$$\widehat{\mathbf{l}} = \underset{\mathbf{l}}{\operatorname{argmax}}\, \mathrm{p}(l_1|\mathbf{q}_1, \boldsymbol{\lambda}) \times \prod_{n=2}^{N} \frac{\mathrm{p}(l_n|\mathbf{q}_n, \boldsymbol{\lambda})}{\mathrm{p}(l_n)} \mathrm{p}(l_n|l_{n-1}) \tag{8.12}$$

On the right hand of the equation, the first term denotes the observation probability of the prosodic event l_n conditionally to the linguistic context $\mathbf{q}_n$ at time n, and the second term denotes the probability associated with the sequence of prosodic events $\mathbf{l}$ regardless to the linguistic context.

The solution to this problem is achieved by using the conventional *Viterbi Algorithm* (VA) [Forney, 1973].

sentence	Longtemps	,	je	me	suis	couché	de	bonne	heure	.
prosodic structure										
F_M	*								*	
F_m	*						*		*	
P	*	*					*		*	
syllable	Long-	temps	##	je	me	suis	cou-	ché	de	bonne	heure	##

Table 8.2: Determination of the sequence of symbolic characteristics.

8.2.5 Evaluation

8.2.5.1 How to Evaluate Speech Prosody? Objective Evaluation and Prosodic Variability

The objective evaluation of prosodic models is a major problem since a speaker has various alternatives to realize a speech prosody depending on his strategies and the context of the speech communication. Consequently, any sentence is potentially associated with various and equally likely speech prosody realizations. However, conventional evaluation procedures only consist in the comparison of the inferred prosodic sequence with the observed prosodic realization. Thus, a correct or plausible speech prosody may actually be considered as incorrect while not strictly corresponding to the actual realization.

A solution to this problem consists of designing a speech database in which each sentence is associated with several realizations [Ostendorf and Veilleux, 1994] that would be considered as possible alternatives. Then, the inferred sequence could be adequately compared with the set of possible realizations with respect to the specific strategies of a speaker. In particular, one can evaluate whether the inferred sequence matches one of the alternatives. However, such an idealistic case appears unrealistic for the evaluation of statistical generative models that are usually based on large speech databases: the recording of several realizations for each sentence would be extensive and time-consuming.

A realistic solution is to define a reasonable distance across different alternatives that would account for the possible prosodic variations. For instance, a major prosodic group may simply be merged or divided by merging two consecutive major prosodic groups while preserving an internal prosodic group at their boundary (pause omission), or by splitting a major prosodic group at the boundary of one of the constitutive internal prosodic groups (pause insertion). This operation is actually equivalent to transform a minor prosodic boundary (F_m) into a major prosodic boundary (F_M) and vice versa. Thus, most of the prosodic alternatives rather relate to a change in the precise *nature* of a prosodic marker (prosodic boundary, prosodic prominence) than a change in the *presence* of a prosodic marker. An example of observed prosodic variations is presented in table 8.3. Naturally, more complex prosodic variations may occur, but a simple distance among prosodic markers may fairly account for prosodic alternatives as a first approximation.
In order to provide a realistic performance measure that partially accounts for potential prosodic alternatives, a performance measure in which errors are penalized depending on the precise nature of prosodic markers is proposed, based on the *Weighted Cohen's Kappa*.

8.2.5.2 Evaluation Scheme

Context-dependent discrete HMMs trained with different sets of linguistic contexts were compared. Evaluation was conducted according to a 10-fold cross-validation [Devijver and Kittler, 1982]. K-fold cross-validation consists of partitioning a set of observations into K complementary subsets of equal size. For each possible partition, models are trained on a combination of (K-1) subsets

sentence	Longtemps, je me suis couché de bonne heure. *For a long time, I used to go to bed early.*
variation #1	(Longtemps) (je me suis couché) (de bonne heure) //
variation #2	(Longtemps) // (je me suis couché) (de bonne heure) //
variation #3	(Longtemps) // (je me suis couché) // (de bonne heure) //

Table 8.3: Study case of prosodic alternatives in which an original sentence is gradually segmented into an increase number of prosodic groups. Parentheses denote internal prosodic boundaries and double bars denote major prosodic boundaries.

and evaluated on the remaining subset. In particular, K-fold cross-validation is a statistical method commonly used to assess the generalization ability of a model, i.e. the performance regardless to some specific training and evaluation sets.

Linguistic Contexts

Linguistic information were extracted from text using the linguistic processing chain described in chapter 7. Models were compared with respect to the following linguistic contexts: morpho-syntactic, dependency, constituency, and adjunction syntactic features. The linguistic units used were: syllable, and the syntactic units. Linguistic features were converted into linguistic contexts over the syllable by computing locational and weight contexts, and representing 1-order left-to-right contexts and 1-order child-to-parent contexts in the case of the dependency contexts.

The different sets of linguistic contexts that were compared are defined as:

$$
\begin{array}{lrcl}
\text{morpho-syntactic:} & Q_{\text{morpho}}^{(syllable)} & = & \mathrm{Q}_{\text{segment}} \cup Q_{\text{morpho}}; \\
\text{dependency:} & Q_{\text{dep}}^{(syllable)} & = & \mathrm{Q}_{\text{segment}} \cup Q_{\text{morpho}} \cup Q_{\text{dep}}; \\
\text{constituency:} & Q_{\text{chunk}}^{(syllable)} & = & \mathrm{Q}_{\text{segment}} \cup Q_{\text{morpho}} \cup Q_{\text{dep}} \cup Q_{\text{chunk}}; \\
\text{adjunction:} & Q_{\text{adj}}^{(syllable)} & = & \mathrm{Q}_{\text{segment}} \cup Q_{\text{morpho}} \cup Q_{\text{dep}} \cup Q_{\text{chunk}} \cup Q_{\text{adj}}.
\end{array}
$$

Each set of linguistic contexts was derived by adding a richer syntactic description to the previous set. The morpho-syntactic context set will be referred as the baseline set for the evaluation.

Evaluation Corpus

Speaker-dependent models were trained and evaluated on the laboratory and multi-media speech databases.

8.2.5.3 Evaluation Metrics

The F_1-*measure* (F-measure) is the most commonly used performance metric in information retrieval [Van Rijsbergen, 1979].

$$F_1(r,p) = \frac{2rp}{r+p} \tag{8.13}$$

where r and p denotes recall and precision measures, respectively.

$$r = \frac{N_{tp}}{N_{tp} + N_{fn}} \tag{8.14}$$

$$p = \frac{N_{tp}}{N_{tp} + N_{fp}} \tag{8.15}$$

where N_{tp}, N_{tn}, N_{fp}, and N_{fn} denotes true positive, true negative, false positive, and false negative respectively, that are directly computed from the observed confusion matrix. In the case

stream	prosodic structure
corpus	
training corpus	(K-1)/K laboratory speech database (8h) (K-1)/K multi-media speech database (4h30)
evaluation corpus	1/K laboratory speech database (1h) 1/K multi-media speech database (30mn)
feature extraction	
window	syllable
frame rate	syllable
feature	hierarchical prosodic structure F_M, F_m, P
feature transform	
transform	linearization
context	**M**: morpho-syntactic context $Q^{(syllable)}_{morpho}$ **D**: dependency context $Q^{(syllable)}_{dep}$ **C**: constituency context $Q^{(syllable)}_{chunk}$ **A**: adjunction context $Q^{(syllable)}_{adj}$
clustering	DT CART
model	
topology	discrete HMM ergodic

Table 8.4: Evaluation of the *discrete HMMs* with *rich linguistic context*: model setup

of a N classes classification problem, it is generally assumed that random performance[2] is equal to $\frac{1}{N}$. However, there is no confidence interval measure available for such a performance floor, in particular depending on the total number of observations, the number of classes, and the relative of observations per class.

The Kappa statistic [Cohen, 1960] was chosen as an alternative to the F_1-measure to measure performance. Cohen's Kappa statistic [Cohen, 1960] measures the proportion of agreement between two independent sources with correction for random agreement:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{8.16}$$

where p_o and p_e are the observed agreement and the agreement expected by chance, respectively. On the right hand side of the equation, the numerator denotes the observed agreement corrected by that expected by chance, and the denominator the maximum corrected-agreement that can be observed. The measure varies from -1 to 1: -1 is perfect disagreement; 0 is chance; 1 is perfect agreement. In particular, Cohen's Kappa manages agreement in the case where both sources are associated with uncertainty (i.e., reference is not available), which is the case in the present since the reference has been derived from automatic transcription of speech prosody. Additionally, *Weighted Cohen's Kappa* [Cohen, 1968] is a refinement of Cohen's Kappa in which errors are more or less penalized according to a distance metric defined between the class labels.

Bayesian Information Criterion (BIC) [Schwarz, 1978] is additionally used to compare models trained with different linguistic context sets. Bayesian Information Criterion is a criterion for model selection in which the likelihood of a model is regularized by the complexity of the model.

$$BIC = -2L + k \log N \tag{8.17}$$

where L is the log-likelihood of the model given the observations from which the parameters of the model have been estimated, k is the number of free parameters of the model, and N is the

[2] random performance corresponds with the performance of a random classifier.

number of observations used to estimate the parameters of the model.

Additionally, the *Relative Error Reduction* (RER) is introduced to account for the relative gain in performance with comparison to the baseline model. Relative error reduction is the difference in performance of a model compared to a reference performance, which is normalized with respect to the difference to the maximum performance that can be obtained.

$$RER(\lambda,\lambda_{ref}) = \frac{s(\lambda) - s(\lambda_{ref})}{1 - s(\lambda_{ref})} \tag{8.18}$$

where s(.) denotes a performance score normalized in the [0,1] interval.

Relative error reduction provides a meaningful measure when a baseline performance is available, since a given absolute gain in performance may change in significance depending on the gap remaining to the maximal performance.

Finally, performance measures were defined as:

Average $\mathbf{F}_1$-measure along the prosodic classes;

Weighted Cohen's Kappa used with a linear penalization along the prosodic scale (table 8.5);

Cohen's Kappa measured for each prosodic class;

Bayesian Information Criterion measured on the training set;

Relative error reduction of the overall *Weighted Cohen's Kappa* and the *Cohen's Kappa* per prosodic class. The baseline performance is defined as being the performance obtained with the morpho-syntactic model;

distance	F_M	F_m	P	NP
F_M	0	1/3	2/3	3/3
F_m	1/3	0	1/3	2/3
P	2/3	1/3	0	1/3
NP	3/3	2/3	1/3	0

Table 8.5: Distance matrix used for the *Linear Cohen's Kappa*.

8.2.6 Results & Discussion

Table 8.6 summarizes the mean performance obtained for the laboratory and multi-media speech databases depending on the linguistic context. A description of the mean performance and 95% confidence interval for both speakers is provided in figure 8.1. The mean performance obtained for the different prosodic events is presented on figure 8.2 and 8.3 for the laboratory and the multi-media speech databases.

	laboratory					multi-media				
linguistic context	$\mathbf{F}_1$	$\mathbf{w-\kappa}$	$\boldsymbol{\kappa_{F_M}}$	$\boldsymbol{\kappa_{F_m}}$	$\boldsymbol{\kappa_P}$	$\mathbf{F}_1$	$\mathbf{w-\kappa}$	$\boldsymbol{\kappa_{F_M}}$	$\boldsymbol{\kappa_{F_m}}$	$\boldsymbol{\kappa_P}$
morpho	61.7	62.9	74.1	41.1	**31.2**	49.2	44.5	54.6	**27.5**	14.4
dependency	61.7	63.1	75.0	40.4	31.1	49.4	44.7	55.8	25.1	15.9
constituency	**64.2**	**65.9**	**84.1**	**42.9**	*30.6*	**52.1**	**47.1**	**65.6**	25.0	**18.0**
adjunction	**67.4**	**70.5**	**94.3**	**46.3**	*30.5*	**53.6**	**49.4**	**72.2**	25.0	**17.4**

Table 8.6: Performance of the context-dependent discrete HMM depending on the linguistic context.

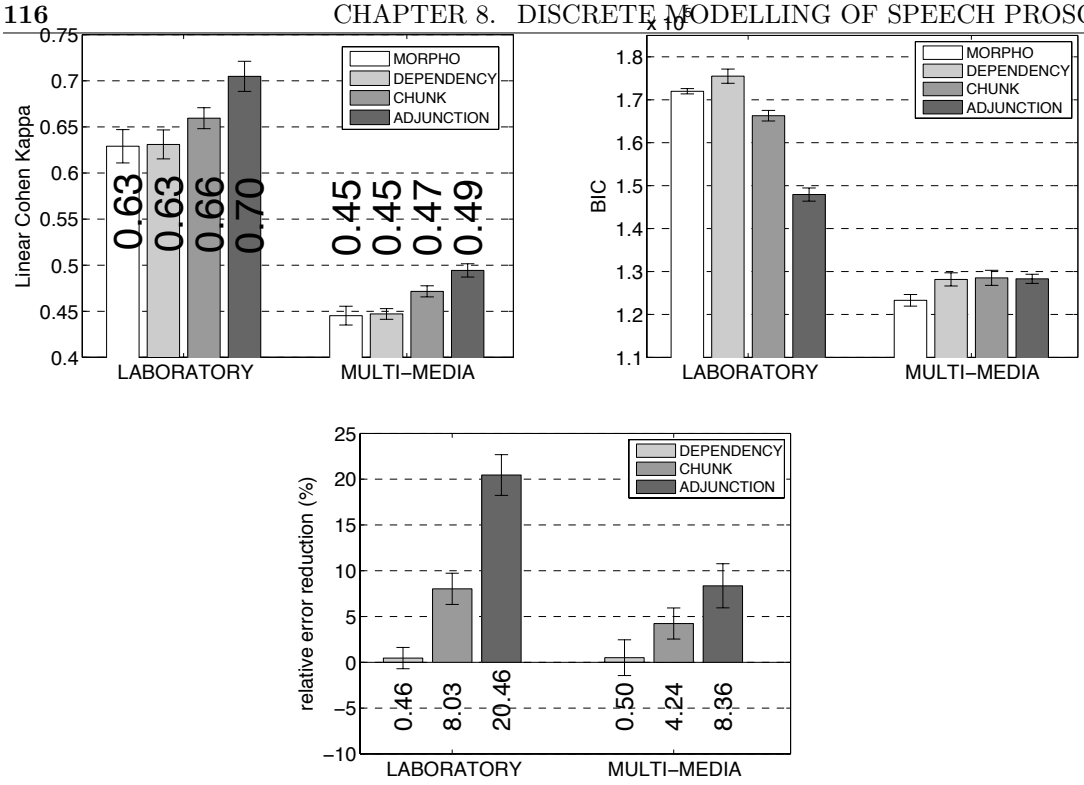
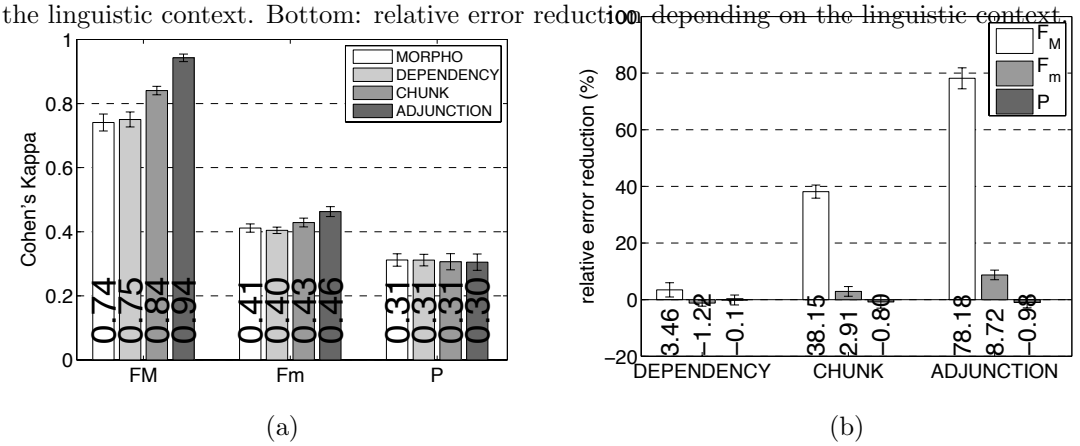


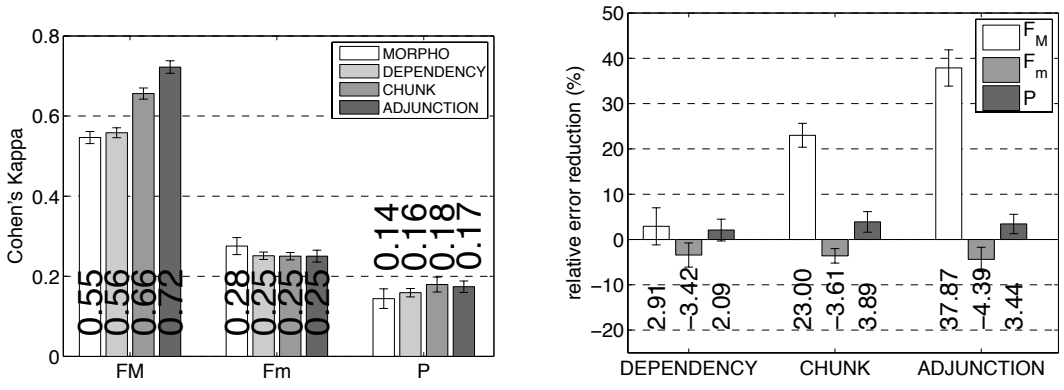
Figure 8.1: Mean performance and 95% confidence interval. Top: overall Linear Cohen's Kappa depending on the linguistic context. Middle: Bayesian Information Criteria value depending on the linguistic context. Bottom: relative error reduction depending on the linguistic context.



(a) (b)

Figure 8.2: Mean performance and 95% confidence interval for the laboratory corpus. (a) Cohen's Kappa of the prosodic event depending on the linguistic context. (b) Relative error reduction for each prosodic event depending on the linguistic context.

The overall performance increases as the linguistic description is enriched (figure 8.1). The increase is particularly significant for the constituency and adjunction contexts. In particular, an overall performance of 70.5% and 49.4% in Kappa is obtained on the different speech databases with the adjunction context, while only 62.9% and 44.5% with the conventional morpho-syntactic context. This constitutes a relative error reduction of 21% and 11% in Kappa, which additionally comes with a relative reduction of 19% and 7.5% in BIC. Conversely, there is no significant difference between the form-based contexts (morpho-syntactic and local dependencies). Thus, the rich syntactic information that are extracted from the deep syntactic parsing clearly improve the symbolic modelling of speech prosody while reducing the complexity of the model. In particular, this indicates that the prosodic structure more closely relates to global syntactic cues (associated with large syntactic units) rather than on local syntactic cues only (associated with small syntactic



(a) (b)

Figure 8.3: Mean performance and 95% confidence interval for the multi-media corpus. (a) Cohen's Kappa of the prosodic event depending on the linguistic context. (b) Relative error reduction for each prosodic event depending on the linguistic context.

units). Finally, adjunctions are proved to provide a strongly relevant syntactic information for the symbolic modelling of speech prosody.

The performance is clearly dependent on the nature of the prosodic event (figures 8.2 and 8.3). Firstly, prosodic boundaries present substantial (F_M) and moderate (F_m) performance while residual prosodic prominence (P) slight performance only. This is consistent with performances reported in literature for other speech prosody transcription systems. Secondly, the enrichment of the syntactic description does not uniformly affect the different prosodic labels.

One the one hand, the improvement is particularly significant for the modelling of prosodic boundaries. The improvement is drastically significant for major prosodic boundary modelling: For the laboratory corpus, κ_{F_M}=74.1% and 94.3% with the morpho-syntactic and adjunction sets respectively which represents a relative error reduction of 78% in the prediction of major prosodic boundaries. For the multi-media corpus, κ_{F_M}=54.6% and 72.2% with the morpho-syntactic and adjunction sets respectively which represents a relative error reduction of 38% in the prediction of major prosodic boundaries.
To a lesser extent, the improvement is fairly significant for minor prosodic boundary modelling: For the laboratory corpus, κ_{F_m}=41.1% and 46.3% with the morpho-syntactic and adjunction sets respectively which represents a a relative error reduction of 9% in the prediction of minor prosodic boundaries. For the multi-media corpus, κ_{F_m}=27.5% and 25.0% with the morpho-syntactic and adjunction sets respectively and no significant difference is observed.
Thus, adjunctions constitute a highly relevant syntactic cue of the prosodic structure, especially of major prosodic boundaries. Nevertheless, adjunction is a generic term that covers a large range of very different syntactic phenomena that are not all necessarily relevant for the modelling of speech prosody. Thus, one needs to refine the description of adjunctions so as to distinguish more precisely those that are relevant for the symbolic modelling of speech prosody.
On the other hand, the prediction of residual prominences is relatively poor, and no significant improvement is observed while increasing the richness of the syntactic description. For the laboratory corpus, κ_P=31.2% and 30.5% with the morpho-syntactic and adjunction sets. For the multi-media corpus, κ_P=14.4% and 17.4% with the morpho-syntactic and adjunction sets.

This confirms evidence that prosodic grouping significantly relates to the syntactic structure while residual prosodic prominences only poorly. Residual prominences are generally assumed to encode semantic and discursive information, thus hardly predictable from a syntactic description only. A higher-level linguistic description is needed to accurately model the location of such prosodic prominences.

Finally, a similar tendency in performance is observed for both speech databases. However, the overall performance for the multi-media corpus is clearly lower, and the increase in performance with the gradual enrichment of the syntactic description is less pronounced than for the laboratory corpus. Such differences may be simply interpreted in terms of the reliability of the syntactic analysis, and eventually by the difference in prosodic complexity of the speakers. Firstly, the syntactic parsing is less robust thus less reliable on complex syntactic structures. Secondly, the professional speaker provides more varied and complex prosodic strategies that increase the complexity of the modelling. On the one hand, the multi-media corpus is linguistically complex, and syntactic parsing achieved complete analysis only on 51% of the utterances. As the performance of the syntactic analysis significantly drops when used for partial parsing, the resulting syntactic analysis is less reliable. On the other hand, the laboratory corpus is linguistically simple, and syntactic parsing achieved complete analysis for 69% of the utterances, thus providing a more robust syntactic analysis.

8.2.7 Conclusion

In this section, a *linguistic-oriented* approach was proposed to integrate a rich linguistic description into the statistical modelling of speech prosody. A discrete HMM was presented in which the symbolic variations of speech prosody are modelled conditionally to the linguistic context. During the training, the text is converted into a sequence of linguistic contexts using the linguistic processing chain described in chapter 7 that includes surface and deep syntactic parsing. Then, a context-dependent discrete HMM is used to model the symbolic variations of speech prosody depending on the linguistic context. During the synthesis, the text is first converted into a sequence of concatenated context-dependent models. Then, the sequence of prosodic events is determined so as to maximize the probability of the sequence of prosodic events conditionally to the sequence of context-dependent models.

The proposed model was objectively evaluated with respect to different sets of linguistic contexts. The rich syntactic description has been shown to drastically improve the performance of the model, and the performance to increase when the linguistic description is enriched. In particular, adjunctions were proved to be highly relevant for the symbolic modelling of speech prosody, especially for major prosodic boundaries. Nevertheless, adjunction is a generic term that covers a large range of very different syntactic phenomena that are not all necessarily relevant for the modelling of speech prosody. Thus, the description of adjunctions needs to be refined so as to distinguish more precisely those that are relevant for the symbolic modelling of speech prosody. However, the syntactic description failed to accurately model residual prosodic prominences. In section 8.3, a context-dependent segmental-HMM will be presented to combine the linguistic and the metric constraints into a single statistical framework. Finally, the description of higher linguistic levels is needed to model residual prosodic prominences accurately.

8.3 Reformulating Prosodic Break Model into Segmental HMMs and Information Fusion

Linguistic studies generally assume that the production of a prosodic punctuation marker - a *prosodic break* - results from the integration of various potentially conflicting constraints, in particular *syntactic* and *metric* constraints [Selrik, 1984, Dell, 1984, Bailly, 1989, Delais-Roussarie, 2000]. A prosodic break is primarily produced by speakers and can be used by listeners to clarify the structure of the utterance. Simultaneously, secondary cognitive constraints (performance constraints) tend to produce a segmentation into prosodic breaks with an optimal configuration [Gee and Grosjean, 1983], in particular with respect to metric regularity [Fraisse, 1974, Dell, 1984]. These constraints conflict in the production of a prosodic structure, and secondary extra-linguistic constraints often override the primary linguistic constraint.

In speech synthesis, the adequate insertion of prosodic breaks guarantees the intelligibility, the naturalness, and the variety of the synthesized speech. Statistical methods have been proposed to combine linguistic and metric constraints based on segmental models [Ostendorf and Veilleux, 1994, Schmid and Atterer, 2004, Bell et al., 2006]) in the modelling and adaptation of prosodic breaks. However, the proposed methods generally remain based on surface syntactic information (POS) solely, while deep syntactic information is ignored. Additionally, the relative importance of linguistic and metric constraints is not considered, or is inadequately formulated.

In this section, a statistical method that combines linguistic and metric constraints in the modelling of prosodic breaks is proposed based on segmental HMMs and Dempster-Shafer fusion, and the relative importance of linguistic and metric constraints is assessed depending on the nature of the linguistic information. A discrete segmental HMM is used in which prosodic breaks are modelled conditionally to the linguistic context in which they are observed, and the distance between successive prosodic breaks (length of a prosodic group) is explicitly modelled. Dempster-Shafer fusion is used to balance the relative importance of the linguistic and the metric constraints into the segmental HMM.

During the training, the text is first converted into a sequence of linguistic contexts using the linguistic processing chain described in chapter 7 that includes surface and deep syntactic parsing. Then, a context-dependent segmental HMM is estimated in which the observation probabilities and the segment probabilities are estimated separately. The observation probabilities are estimated using the context-dependent discrete HMM presented in section 8.1, and the segment duration probabilities are estimated with a normal distribution. During the synthesis, the text is first converted into a sequence of concatenated context-dependent segmental models. Then, the sequence of prosodic breaks is determined so as to maximize the conditional probability of the prosodic break sequence given the linguistic observations and the segmental models. Additionally, Dempster-Shafer fusion is used so as to optimally combine the linguistic and the metric constraints into segmental HMMs. Segmental HMMs are objectively evaluated with respect to different sets of linguistic contexts, and the relative importance of linguistic and metric constraints is assessed.

This section is organized as follows: segmental HMMs and their application to prosodic break modelling are presented in section 8.3.1, and Dempster-Shafer fusion is presented in section 8.3.4. The evaluation is described and discussed in sections 8.3.5 and 8.3.6.

8.3.1 Segmental HMMs

Segmental HMMs [Russel and Moore, 1985, Levinson, 1986, Gales and Young, 1993, Ostendorf et al., 1996] were introduced in speech recognition in which state sequences are explicitly represented as *segments* with an explicit modelling of the segment state-occupancy duration. Segmental HMM is a generalization of hidden Markov model (HMM) that addresses two principal limitations of the conventional hidden Markov model: 1) state duration modelling,

and 2) assumption of conditional independence of the observations given the state sequence.

Let define $\mathbf{o} = [o_1, \ldots, o_T]$ an observation sequence of length T, $\mathbf{q} = [q_1, \ldots, q_T]$ the associated state sequence, $\mathbf{s} = [s_1, \ldots, s_K]$ the associated segment sequence of length K, and $\mathbf{d} = [d_1, \ldots, d_K]$ the corresponding segment durations.

A segmental hidden Markov model $\boldsymbol{\lambda}$ is defined in a similar manner to the hidden Markov model with a reformulation of the state sequence into segment sequence and the add of an explicit state duration:

$$\boldsymbol{\lambda} = (\mathbf{\Pi}, \mathbf{A}, \mathbf{B}, \mathbf{D}) \tag{8.19}$$

where:

- Π is the initial state probability distribution: $\mathbf{\Pi} = \{\pi_i\}_{i=1}^N$

$$\pi_i = \mathrm{p}(s_1 = i) \qquad i \in [1, N] \tag{8.20}$$

- $\mathbf{A}$ is the segment transition probability distribution: $\mathbf{A} = \{a_{i,j}\}_{i,j=1}^N$

$$a_{i,j} = \mathrm{p}(s_k = j | s_{k-1} = i) \qquad \begin{matrix} k \in [1, K] \\ i, j \in [1, N] \end{matrix} \tag{8.21}$$

- $\mathbf{B}$ is the output probability distribution: $\mathbf{B} = \{b_{i,d}(s)\}_{i=1}^N$

$$b_{i,d}(o_{[t+1:t+d]}) = \mathrm{p}(o_{[t+1:t+d]} | s = i) \qquad \begin{matrix} t \in [1, T] \\ i \in [1, N] \end{matrix} \tag{8.22}$$

- $\mathbf{D}$ is the segment duration probability distribution: $\mathbf{D} = \{d_i(s)\}_{i=1}^N$

$$d_i(s) = \mathrm{p}(d | s = i) \qquad i \in [1, N] \tag{8.23}$$

and N is the number of states.

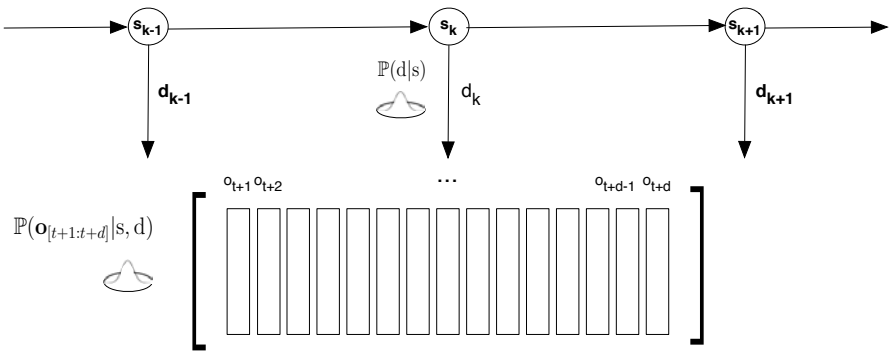


Figure 8.4: Schematic illustration of segmental HMMs

In a similar manner as for hidden Markov models, there are three common problems related to segmental HMMs (1) the evaluation of the probability $p(\mathbf{o}|\lambda)$ of an observation sequence $\mathbf{o} = [o_1, \ldots, o_T]$ given the model λ; (2) the determination of the optimal state sequence $\widehat{\mathbf{q}} = [\widehat{q}_1, \ldots, \widehat{q}_T]$ given an observation sequence $\mathbf{o} = [o_1, \ldots, o_T]$ and the model λ; (3) the estimation of the model parameters λ which optimize a given objective function of an observation sequence $\mathbf{o} = [o_1, \ldots, o_T]$ given the model λ.

In the present study, only the determination of the optimal sequence $\widehat{\mathbf{q}} = [\widehat{q}_1, \ldots, \widehat{q}_T]$ given an observation sequence $\mathbf{o} = [o_1, \ldots, o_T]$ and the model λ will be described and reformulated for prosodic break modelling. The solution to the other problems is similar to that of a HMM, and detailed formulations can be found in literature [Russel and Moore, 1985, Levinson, 1986, Gales and Young, 1993, Ostendorf et al., 1996].

The reformulation of prosodic break modelling into a segment model requires to reformulate prosodic breaks as segments. Actually, a *prosodic break* instantiates a prosodic segment (*prosodic phrase*) that is defined as the segment left/right bounded by a prosodic break. Thus, the modelling of prosodic breaks can reformulated in terms of prosodic segments.

Let define $\mathbf{q} = [\mathbf{q}_1, \ldots, \mathbf{q}_N]$ the sequence of linguistic contexts of length N, where $\mathbf{q}_n = [q_n(1), \ldots, q_n(L)]^\top$ is the (Lx1) linguistic context vector which describes the linguistic characteristics associated with the n-th syllable, $\mathbf{l} = [l_1, \ldots, l_N]$ the corresponding sequence of prosodic events, where l_n denotes the prosodic event associated with the n-th syllable, $\mathbf{s} = [s_1, \ldots, s_K]$ the associated sequence of prosodic phrases of length K, and $\mathbf{d} = [d_1, \ldots, d_K]$ the corresponding segment state-durations, where d_k denotes the length of prosodic phrase s_k.

In prosodic break modelling, the segment model can be simplified as follows:

1. one segment: $s_k = [\, l_{[t_{k-1}+1:t_k-1]} = \bar{b},\ l_{t_k} = b \,]$

2. segment transition = 1

where: $\mathbf{t} = [t_1, \ldots, t_K]$ denotes the sequence of segment boundaries, and b denotes a prosodic break and $\bar{b}$ the absence of a prosodic break.

8.3.2 Parameters Estimation

During the training, the parameters of the linguistic and segment duration models are estimated separately.

$$\boldsymbol{\lambda} = \left(\boldsymbol{\lambda}^{(\text{linguistic})}, \boldsymbol{\lambda}^{(\text{metric})}\right) \tag{8.24}$$

The linguistic model $\boldsymbol{\lambda}^{(\text{linguistic})}$ is estimated using the context-dependent discrete HMM described in section 8.1 . First, linguistic contexts are first clustered so as to derive a context-dependent tree. Then, a context-dependent HMM $\boldsymbol{\lambda}^{(\text{linguistic})} = (\lambda_{S_1}^{(\text{linguistic})}, \ldots, \lambda_{S_M}^{(\text{linguistic})})$ is constructed from the set of terminal contexts $S = (S_1, \ldots, S_M)$ of the decision-tree, where $\lambda_{S_m} = (\boldsymbol{\Pi}_{S_m}, \mathbf{A}_{S_m}, \mathbf{B}_{S_m})$ denotes the estimated HMM parameters associated with the context S_m.

The segment duration model $\boldsymbol{\lambda}^{(\text{metric})} = (\mathbf{D})$ is estimated with a normal distribution.

8.3.3 Parameters Inference

During the synthesis, the segment sequence $\widehat{(\mathbf{s}, \mathbf{d})}$ is determined so as to maximize the conditional probability of the segment sequence $\mathbf{s}$ and the segment duration sequence $\mathbf{d}$ given the linguistic context sequence $\mathbf{q}$:

$$\widehat{(\mathbf{s},\mathbf{d})} = \underset{\mathbf{s},\mathbf{d}}{\text{argmax}}\, \mathrm{p}(\mathbf{s},\mathbf{d}|\mathbf{q}) \tag{8.25}$$

$$= \underset{\mathbf{s}_{[1:K]}}{\text{argmax}}\Big(\max_{\mathbf{d}_{[1:K]}} \underbrace{\mathrm{p}(\mathbf{q}_{[1:T]}|\mathbf{s}_{[1:K]},\mathbf{d}_{[1:K]})}_{\substack{\text{observation}\\ \text{probability}}} \times \underbrace{\mathrm{p}(\mathbf{d}_{[1:K]}|\mathbf{s}_{[1:K]})}_{\substack{\text{segment}\\ \text{probability}}} \times \underbrace{\mathrm{p}(\mathbf{s}_{[1:K]})}_{\substack{\text{segment transition}\\ \text{probability}}}\Big) \tag{8.26}$$

Since only one type of segment is being considered, the optimal segment sequence $\widehat{(\mathbf{s},\mathbf{d})}$ is equivalent to the determination of the optimal segment duration sequence $\widehat{\mathbf{d}}$:

$$\widehat{\mathbf{d}} = \underset{\mathbf{d}_{[1:K]}}{\text{argmax}}\, \underbrace{\mathrm{p}(\mathbf{q}_{[1:T]}|\mathbf{d}_{[1:K]})}_{\substack{\text{observation}\\ \text{probability}}} \times \underbrace{\mathrm{p}(\mathbf{d}_{[1:K]})}_{\substack{\text{segment}\\ \text{probability}}} \tag{8.27}$$

Additionally, assuming conditional independence of the observations given the state sequence:

$$\mathrm{p}(\mathbf{q}_{[t_{k-1}+1:t_k]}|\mathrm{s}_k,\mathrm{d}_k) = \left(\prod_{t=t_{k-1}+1}^{t_k-1} \mathrm{p}(q_t=\bar{b}|l_t)\right)\mathrm{p}(q_{t_k}=b|l_{t_k}) \tag{8.28}$$

where $\mathbf{t}=[t_1,\ldots,t_K]$ denote the sequence of segment boundaries that corresponds to the sequence of segment duration $\mathbf{d}=[d_1,\ldots,d_K]$.

Using Bayes' theorem:

$$\mathrm{p}(\mathbf{q}_{[t_{k-1}+1:t_k]}|\mathrm{d}_k) = \left(\prod_{t=t_{k-1}+1}^{t_k-1} \frac{\mathrm{p}(l_t=\bar{b}|q_t)\mathrm{p}(q_t)}{\mathrm{p}(l_t=\bar{b})}\right) \times \frac{\mathrm{p}(l_{t_k}=b|q_{t_k})\mathrm{p}(q_t)}{\mathrm{p}(l_{t_k}=b)} \tag{8.29}$$

Then, assuming that the probability of the linguistic context sequence is constant during the maximization:

$$\widehat{\mathbf{d}} = \underset{\mathbf{d}_{[1:K]}}{\text{argmax}} \prod_{k=1}^{K} \left(\left(\prod_{t=t_{k-1}+1}^{t_k-1} \frac{\mathrm{p}(l_t=\bar{b}|q_t)}{\mathrm{p}(l_t=\bar{b})}\right) \times \frac{\mathrm{p}(l_{t_k}=b|q_{t_k})}{\mathrm{p}(l_{t_k}=b)}\right) \times \mathrm{p}(d_k|\mathbf{l}_{[t_{k-1}+1:t_k-1]}=\bar{b}, l_{t_k}=b) \tag{8.30}$$

Finally, the optimal segment sequence can be reformulated in the conventional state sequence manner:

$$\widehat{\mathbf{l}} = \underset{\mathbf{l}}{\text{argmax}} \prod_{k=1}^{K} \frac{\mathrm{p}(\mathbf{l}_{[t-d_k+1:t-1]}=\bar{b}, l_t=b\,|\mathbf{q}_{[t-d_k+1:t]})}{\mathrm{p}(\mathbf{l}_{[t-d_k+1:t-1]}=\bar{b}, l_t=b)} \times \mathrm{p}(d_k|\mathbf{l}_{[t_{k-1}+1:t_k-1]}=\bar{b}, l_{t_k}=b) \tag{8.31}$$

$$= \underset{\mathbf{l}}{\text{argmax}} \prod_{k=1}^{K} \underbrace{\mathrm{p}_o(l_{t_k})}_{\substack{\text{observation}\\ \text{probability}}} \underbrace{\mathrm{p}_s(l_{t_k})}_{\substack{\text{segment}\\ \text{probability}}} \tag{8.32}$$

where $\mathrm{p}_s(l_{t_k}) = \mathrm{p}(\mathbf{l}_{[t_{k-1}+1:t_k-1]}=\bar{b}, l_{t_k}=b\,|d_k)$ denotes the partial probability that the k-th segment with duration d_k ends at time t, and $\mathrm{p}_o(l_{t_k}) \propto \mathrm{p}(\mathbf{l}_{[t_{k-1}+1:t_k-1]}=\bar{b}, l_{t_k}=b\,|\mathbf{q}_{[t_{k-1}+1:t_k]})$ the partial observation probability over the k-th segment with duration d_k.

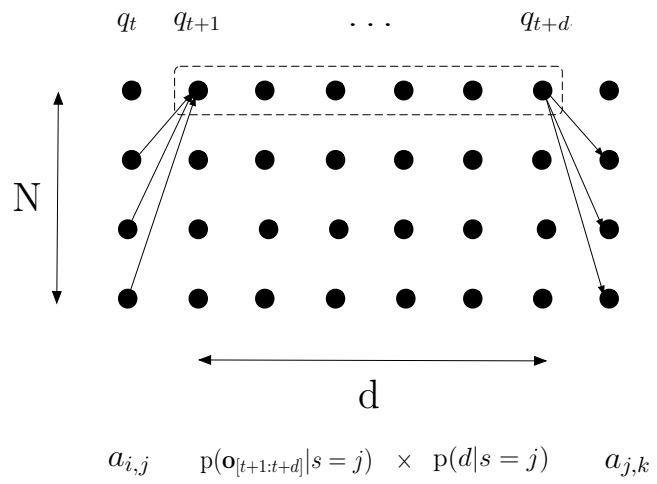


Figure 8.5: Schematic illustration of segmental HMMs decoding (N=4). $a_{i,j}$ is the transition probability from segment i to segment j at time t, $\mathrm{p}(\mathbf{o}_{t+1:t+d}|s=j) \times \mathrm{p}(d|s=j)$ is the probability that the segment is in state j from time $t+1$ to time $t+d$ with duration d, and $a_{j,k}$ is the transition probability from segment j to segment k at time $t+d$.

In the preceding equation, the observation probability accounts for the linguistic constraint, and the segment probability for the metric constraint.

The solution to this problem is obtained by a reformulation of the conventional *Viterbi Algorithm* (VA) to account for the segment duration probabilities [Ostendorf et al., 1996]. The determination of the optimal sequence requires the computation of the partial probabilities for each possible segment duration $d \in [1, D]$, where D is the maximum expected duration of a segment.

8.3.4 Segmental HMMs & Dempster-Shafer Fusion

In the conventional formulation of segmental HMMs, the segment probability and the observation probability are equally considered. However, linguistic studies pointed out that the linguistic and the metric constraints are not of equal importance in the production of a prosodic break. In particular, the metric constraint is generally assumed to be secondary compared to the linguistic constraint, and the integration of the constraints results from a complex process. Consequently, a proper integration of the observation and the segment probabilities into the segmental HMMs would be convenient to balance adequately the linguistic and metric constraints. Dempster-Shafer fusion will be used to optimally balance the observation probability and the segment probability into the segmental HMMs.

Dempster-Shafer theory [Shafer, 1976] is a mathematical theory commonly used for information fusion in statistical processing. In particular, Dempster-Shafer theory provides a proper probabilistic formulation for information fusion, in which the *reliability* that can be conferred to different sources of information can be explicitly formulated. In the Dempster-Shafer fusion, PDFs can be reformulated into mass functions (MFs) to account for the reliability that can be conferred to each PDF, and then combined with the Dempster-Shafer fusion rule. The principle of the Dempster-Shafer combination is shortly described and its integration into segmental HMMs and prosodic break modelling is presented.

8.3.4.1 Mass Function

An elementary mass function m is a function of $\mathcal{P}(\mathcal{C})$ in $\mathbb{R}_+$ that verify:

$$\begin{cases} \mathrm{m}(\emptyset) = 0 \\ \sum_{\mathcal{A}\in\mathcal{P}(\mathcal{C})} = 1 \end{cases} \tag{8.33}$$

where $\mathcal{C}$ is the state alphabet, and $\mathcal{P}(\mathcal{C})$ is the power set of $\mathcal{C}$.

Mass functions present the advantage over conventional probabilities that a mass can be assigned to composite classes rather than singletons only, thus can be used to account for the *reliability* that can be conferred to different sources of information during the fusion.

8.3.4.2 Dempster-Shafer Fusion

The Dempster-Shafer fusion of two masses is given by:

$$m(A) = (m_1 \oplus m_2)(A) \tag{8.34}$$
$$\propto \sum_{B_1\cap B_2=A} m_1(B_1) \times m_2(B_2) \tag{8.35}$$

Hence, the Dempster-Shafer fusion of a mass m and a probability p is a probability given by:

$$(m \oplus p)(x) = \frac{\sum_{x\in u} m(u)p(x)}{\sum_{l'\in\mathcal{C}} \sum_{x'\in u'} m(u')p(x')} \tag{8.36}$$

where m denotes the mass associated with a source of information for which the reliability may vary and p the probability associated with another source of information.

In order to balance the relative importance of the linguistic constraint $\mathrm{p}_o(l_t)$ and the metric constraint $\mathrm{p}_s(l_t)$ into the segmental HMM, one of the PDFs is alternatively replaced by a mass function (MF), while the other remains a PDF:

$$m_o(l_t) = \alpha\ \mathrm{p}_o(l_t) \qquad m_o(\mathcal{C}) = 1-\alpha \tag{8.37}$$
$$m_s(l_t) = \beta\ \mathrm{p}_s(l_t) \qquad m_s(\mathcal{C}) = 1-\beta \tag{8.38}$$

where α and β denote the reliability that is associated with the observation probability $\mathrm{p}_o(l_t)$ and the segment probability $\mathrm{p}_s(l_t)$ respectively, and $m_o(\mathcal{C})$ and $m_s(\mathcal{C})$ the corresponding model *ignorance*.

The Dempster-Shafer fusion of m_o and m_s is then given by:

$$(m_o \oplus m_s)(l) \propto \alpha(1-\beta)\mathrm{p}_o(l_t) + \alpha\beta\ \mathrm{p}_o(l_t)\mathrm{p}_s(l_t) + \beta(1-\alpha)\mathrm{p}_s(l_t) \tag{8.39}$$

Hence,

$$(m_1 \oplus m_2)(l_t) \propto \begin{cases} \mathrm{p}_o(l_t), & \alpha=1,\ \beta=0 & ① \\ \mathrm{p}_s(l_t), & \alpha=0,\ \beta=1 & ② \\ \mathrm{p}_o(l_t)\ \mathrm{p}_s(l_t), & \alpha=1,\ \beta=1 & ③ \end{cases} \tag{8.40}$$

① denotes that only the segment probability is considered, ② denotes that only observation probability is considered (conventional HMM), and ③ denotes that the segment and observation

probabilities are equally considered (conventional segmental HMM). In the latter case, the expression is equivalent to the conventional Bayes combination rule.

Finally, the relative confidence α and β are rewritten into a single weight (α,β) so that the relative importance of the linguistic and the segment probabilities is linearly interpolated from the metric constraint solely to the linguistic constraint solely. Thus: $(\alpha,\beta) = -1$ will refer to $\alpha = 0$ and $\beta = 1$, $(\alpha,\beta) = 0$ to $\alpha = 1$ and $\beta = 1$, and $(\alpha,\beta) = +1$ to $\alpha = 1$ and $\beta = 0$.

8.3.5 Evaluation

The evaluation was conducted to assess the relative importance of the linguistic and the metric constraints, and their combination in prosodic break modelling. In particular, a large range of combination of linguistic contexts was used to estimate context-dependent segmental HMMs, and various combinations of the linguistic and metric constraints - from the individual performance of the metric constraint to the individual performance of linguistic constraint - were compared. The evaluation of the segmental HMM was similar to that used in section 8.2, with the exception that only the prosodic break modelling (F_M) was evaluated. Two baseline models were used for the comparison: the conventional punctuation rule-based model (P) in which a prosodic break is inserted after each punctuation marker, and the segmental HMM estimated with the conventional morpho-syntactic linguistic context (M).

stream	prosodic structure
corpus	
training corpus	(K-1)/K laboratory speech database (8h) (K-1)/K multi-media speech database (4h30)
evaluation corpus	1/K laboratory speech database (1h) 1/K multi-media speech database (30mn)
feature extraction	
window	syllable
frame rate	syllable
feature	prosodic break F_M
feature transform	
transform	linearization
context	combinations of: **M**: morpho-syntactic context $Q^{(syllable)}_{morpho}$; **D**: dependency context $Q^{(syllable)}_{dep}$; **C**: constituency context $Q^{(syllable)}_{chunk}$; **A**: adjunction context $Q^{(syllable)}_{adj}$
clustering	DT CART
model	
topology	discrete segmental HMM segment: normal distribution observation: discrete HMM

Table 8.7: Evaluation of the *segmental HMMs*: model setup

8.3.5.1 Evaluation Scheme

The comparison of context-dependent segmental HMMs was conducted using different set of linguistic contexts and different combination of the linguistic and the metric constraints. Evaluation was conducted according to a 10-fold cross-validation.

Linguistic Contexts

Linguistic information were extracted from text using the linguistic processing chain described in chapter 7. Models were compared with respect to any combination of the following linguistic feature sets: morpho-syntactic (M), dependency (D), constituency (C), and adjunction (A). The used linguistic units were: syllable, and the syntactic units. Linguistic features were converted into linguistic contexts over the syllable by computing locational and weight contexts, and representing 1-order left-to-right contexts and 1-order child-to-parent contexts in the case of the dependency contexts.

Evaluation Corpus

Speaker-dependent models were trained and evaluated on the laboratory and multi-media speech databases. During the training, the segmental and the linguistic probabilities are estimated separately: the context-dependent model is estimated using a conventional context-dependent HMM, and the segment duration probability is estimated with a normal distribution.

8.3.5.2 Evaluation Metrics

The evaluation metrics are the same at those used in section 8.2. However, only the F_1 measure is presented for clarity. A paired Student t-test [Box et al., 1978] was employed to assess whether a significant difference exists between the models being compared.

8.3.6 Results & Discussion

Table 8.8 summarizes the mean performance obtained for the laboratory and multi-media speech databases depending on the linguistic context, and the comparison of the metric model only, the conventional segmental-HMM, the linguistic model only, and the optimal configuration of linguistic and metric constraints into the segmental-HMM. For concision, the optimal combinations of the linguistic contexts and the baseline models are presented in the table only. Overall performances depending on the linguistic context and the configuration of the linguistic and metric constraints are presented in figures 8.6 and 8.8. Precision and recall measures of the optimal configuration depending on the linguistic context are presented in figures 8.7 and 8.9.

The optimal configuration significantly outperforms the conventional segmental-HMM and the linguistic model for all of the linguistic contexts, and corresponds to a prior importance of the linguistic constraint over the metric constraint. Additionally, a significant correlation exists between the optimal balance of the linguistic and metric constraints and the performance of the linguistic model ($\rho(\mathrm{p}_o)$=+0.75), the conventional segmental-HMM ($\rho(\mathrm{p}_s\mathrm{p}_o)$=+0.64), and their differential ($\rho(\mathrm{p}_o - \mathrm{p}_s\mathrm{p}_o)$=+0.80), but not with the metric model ($\rho(\mathrm{p}_s)$=+0.29). Thus, the balance of the linguistic and the metric constraints varies depending on the relevancy of the linguistic or/and the metric constraint. In particular, the optimal configuration gradually tends to the linguistic constraint when the linguistic information increase in reliability (the balance (α,β) varies from +0.23 to +0.56 for the laboratory corpus in correlation with a linguistic performance which varies from 78.3% to 95%), and is very close to the linguistic constraint when the metric constraint is not reliable (the balance (α,β) varies from +0.58 to +0.73 in correlation with a segment performance of 39%).

The conventional segmental-HMM outperforms the linguistic model only when the linguistic information is slightly relevant. In particular, the conventional segmental-HMM significantly outperforms the linguistic model based on the conventional morpho-syntactic information (78.3% and 74.2% for the laboratory corpus, and 58.7% and 59.2% for the multi-media corpus), which is consistent with the performance obtained in the literature [Schmid and Atterer, 2004]. However, the conventional segmental-HMM is outperformed when the linguistic description is enriched.

The relevancy of the different linguistic features in prosodic break modelling confirms and refines observations reported in section 8.1. Adjunction (A) and to a lesser extent constituency (C) are the most relevant single linguistic contexts (91.7% and 83.8% for the laboratory corpus, 63.2% and

context	$p_{optimal}$	α,β	p_s	$p_s p_o$	t-test	p_o	t-test
laboratory							
M/D/C/A	**96.3**	+0.44	65.4	92.1	*<0.001*	95.0	*<0.001*
M/C/A	**96.0**	+0.48	65.4	92.1	*<0.001*	94.7	*<0.001*
C/A	**96.0**	+0.54	65.4	92.0	*<0.001*	94.6	*<0.001*
D/C/A	**95.8**	+0.56	65.4	91.7	*<0.001*	94.6	*<0.001*
D/A	**94.1**	+0.41	65.4	89.1	*<0.001*	92.6	*<0.001*
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
M	**78.3**	+0.23	65.4	75.5	*<0.001*	74.2	*<0.001*
P	**66.3**	-	-	-	-	-	-
multi-media							
M/D/C/A	**75.3**	+0.70	39.0	70.0	*<0.001*	74.0	*0.03*
M/C/A	**75.2**	+0.58	39.0	69.6	*<0.001*	73.6	*0.04*
D/C/A	**74.2**	+0.68	39.0	68.6	*<0.001*	72.8	*0.02*
C/A	**73.7**	+0.73	39.0	67.4	*<0.001*	72.6	*0.2*
M/C	**69.6**	+0.65	39.0	65.5	*<0.001*	68.0	*0.1*
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
M	**59.2**	+0.62	39.0	56.7	*0.03*	58.7	*0.5*
P	**55.1**	-	-	-	-	-	-

Table 8.8: Ranking of the F_1 measure for the optimal configuration, the segmental model only p_s, the conventional segmental-HMM $p_s p_o$, and the linguistic model only p_o. Significance test for the comparison of the optimal configuration, the conventional segmental-HMM and the linguistic model. M and P denote the conventional context-dependent model with the morpho-syntactic context and the rule-based punctuation model.

65.3% for the multi-media corpus), and their combination (CA) is strongly relevant (96.0% and 73.7% for the laboratory and the multi-media speech databases respectively). Morpho-syntactic (M) and dependency (D) are slightly relevant linguistic contexts (78.3% and 73.8% for the laboratory corpus, 59.2% and 52.0% for the multi-media corpus). Nevertheless, the optimal performance is obtained for the combination of all of the linguistic contexts (MDCA) (96.3% and 75.3% for the laboratory and multi-media speech databases respectively).

Additionally, recall and precision mutually increase when the linguistic description is enriched (figures 8.7 and 8.9). However, the increase in recall is large (from 62% to 92% for the laboratory corpus, and from 42% to 67% for the multi-media corpus) compared to the increase in precision (from 92% to 97% for the laboratory corpus, and from 76% to 82% for the multi-media corpus). Thus, the enrichment of the linguistic description significantly decreases the omission of prosodic breaks, while the false insertion of prosodic breaks remains globally marginal regardless to the linguistic description.

The increase in performance obtained with the enrichment of the linguistic description is significantly larger compared to that obtained with the integration of the metric constraint. For the linguistic constraint, the increase in performance is of 18% by comparison of the conventional morpho-syntactic context (78.3%) and the optimal linguistic context (96.3%). For the combination of the linguistic and the metric constraints, the increase in performance does not exceed 4% (74.2% and 78.3% for the conventional morpho-syntactic context), and 2% with a rich linguistic description (95% and 93.3% for the optimal linguistic context).

Finally, the performance significantly varies depending on the speech database. The overall performance, and the increase in performance due to the enrichment of the linguistic description and the combination of the linguistic and the metric constraints are significant larger for the laboratory corpus compared to the multi-media corpus. The difference may be simply interpreted in terms of the reliability of the syntactic analysis, and eventually by the difference in prosodic

complexity of the speakers, which confirms the observations reported in section 8.1.

Qualitatively, the combination of the linguistic and the metric constraints tends to produce a more natural prosodic segmentation compared to the linguistic constraint solely. In particular, insertion or omission of a prosodic break is generally consistent and produces a plausible prosodic alternative, while the actual segmentation is not exactly matched. Thus, the inherent prosodic variability may partially explain the slight objective improvement compared to the observed qualitative improvement. A study case of the linguistic and metric combination is presented in table 8.9. The prevalence of the metric constraint ((α,β)= -1.0 and -0.6) produces a well-balanced configuration of prosodic breaks, but is linguistically unlikely. The introduction of the linguistic constraint gradually provide a set of plausible prosodic alternatives (from $(\alpha,\beta) = +0.0$ to +1.0), while the metric constraint is gradually being omitted ($(\alpha,\beta) = +0.6$ and +1).

8.3.7 Conclusion

In this section, a statistical method that combines linguistic and metric constraints in the modelling of prosodic breaks was proposed based on segmental HMMs and Dempster-Shafer fusion, and the relative importance of the linguistic and the metric constraints was assessed depending on the nature of the linguistic information. A discrete segmental HMM was used in which prosodic breaks are modelled conditionally to the linguistic context in which they are observed, and the distance across successive prosodic breaks (length of a prosodic group) is explicitly modelled. Dempster-Shafer fusion was additionally employed to balance the relative importance of the linguistic constraint and the metric constraint into segmental HMMs.

During the training, a context-dependent segmental HMM is estimated in which the observation probabilities and the segment probabilities are estimated independently. During the synthesis, the text is first converted into a sequence of concatenated context-dependent segmental models. Then, the sequence of prosodic breaks is determined so as to maximize the conditional probability of the prosodic break sequence given the sequence of linguistic observations. Additionally, Dempster-Shafer fusion is used so as to optimally combine the linguistic constraint and the segment constraint into segmental HMMs. Segmental HMMs were objectively evaluated with respect to different sets of linguistic contexts, and the relative importance of the linguistic and the metric constraints was assessed.

The optimal combination of the linguistic and the metric constraints in segmental HMM was proved to significantly outperform the conventional segmental-HMM and the linguistic model only. The linguistic constraint was shown to be prior to the metric constraint, and the optimal configuration to gradually tend to the linguistic constraint when the linguistic description is enriched, or when the segment constraint is slightly reliable. The conventional segmental HMM was found to outperform the linguistic model only when the linguistic information used is slightly relevant. Finally, the increase in performance obtained by the integration of the metric constraint and its combination with the linguistic constraint remains slight compared to that obtained with the enrichment of the linguistic description. In further studies, the metric model will be refined to improve the modelling of the metric constraint and its combination with the linguistic constraint, and will be evaluated on the modelling of various speaking styles.

sentence	Et,	une	demi-heure	après,	la	pensée	qu'il	était	temps	de	chercher	le	sommeil	m'éveillait.
(α,β)														
-1.0								*						*
-0.6						*					*			*
0.0				*										*
+0.6	*			*										*
+1.0	*			*									*	*
syllable	Et	une	de-mi-heure	a-près	la	pen-sée	qu'il	é-tait	temps	de	cher-cher	le	som-meil	m'é-vei-llait.

Table 8.9: Evolution of the prosodic break configuration depending on the combination of the metric and the linguistic constraints for the sentence: "*Et, une demi-heure après, la pensée qu'il était temps de chercher le sommeil m'éveillait.*" ("*And half an hour later the thought that it was time to go to sleep would awaken me.*"). (α,β) = -1.0 denotes the metric constraint solely, (α,β) = 0.0 denotes the balanced combination of the metric and the linguistic constraints, (α,β) = +1.0 denotes the linguistic constraint solely.

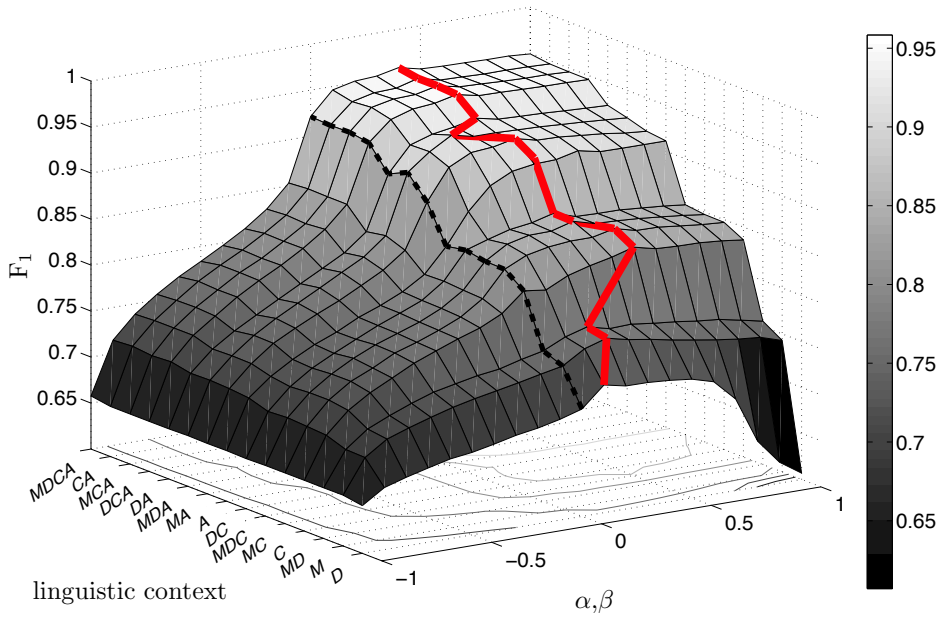


Figure 8.6: Laboratory corpus: F_1 measure depending on the linguistic context and the balance (α,β) of metric and linguistic constraints.

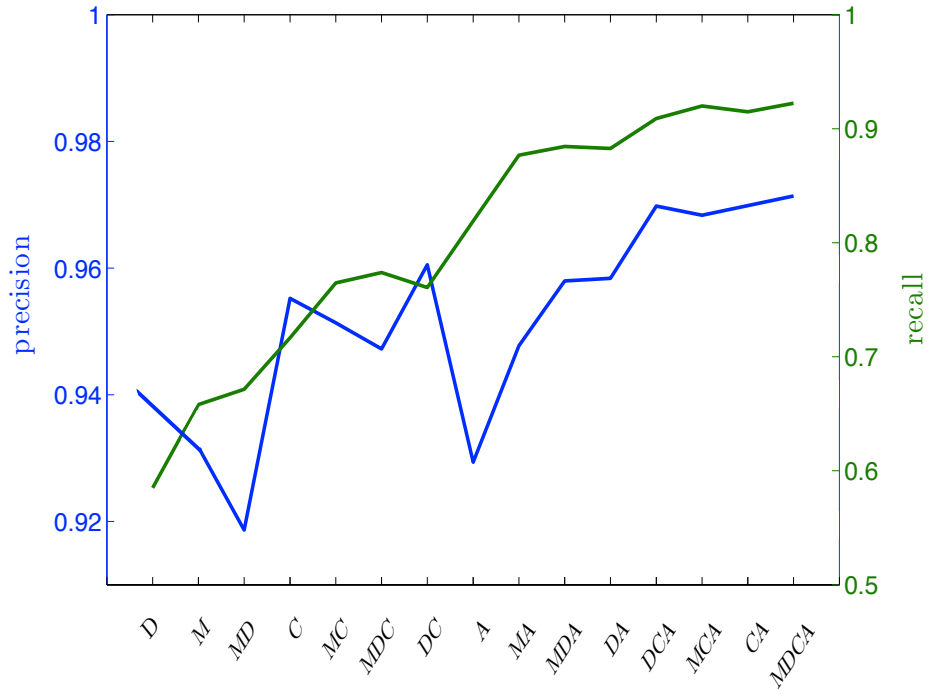


Figure 8.7: Laboratory corpus: precision and recall of the optimal model depending on the linguistic context.

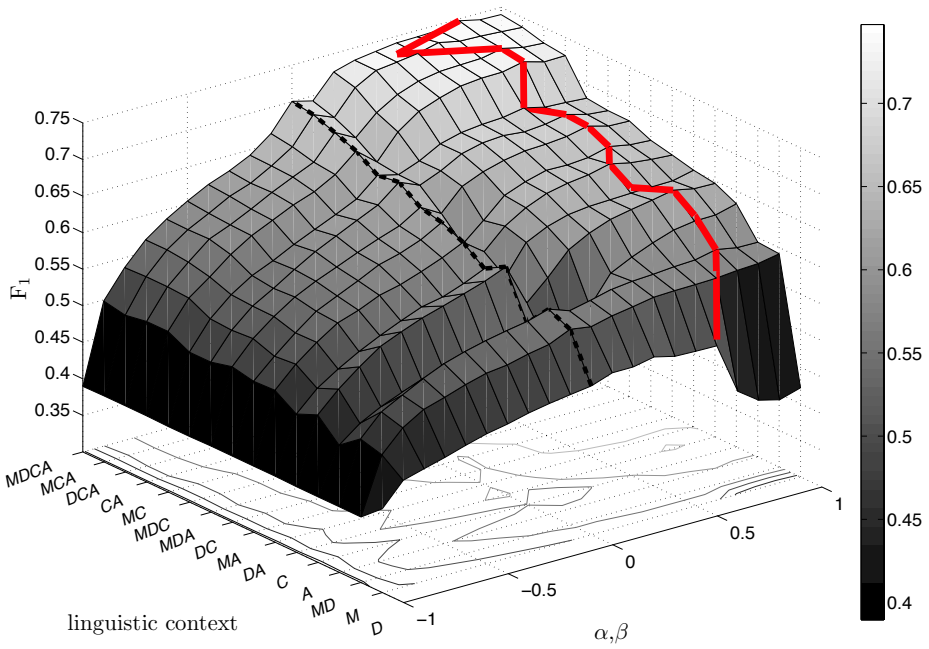


Figure 8.8: Multi-media corpus: F_1 measure depending on the linguistic context and the balance (α,β) of metric and linguistic constraints.

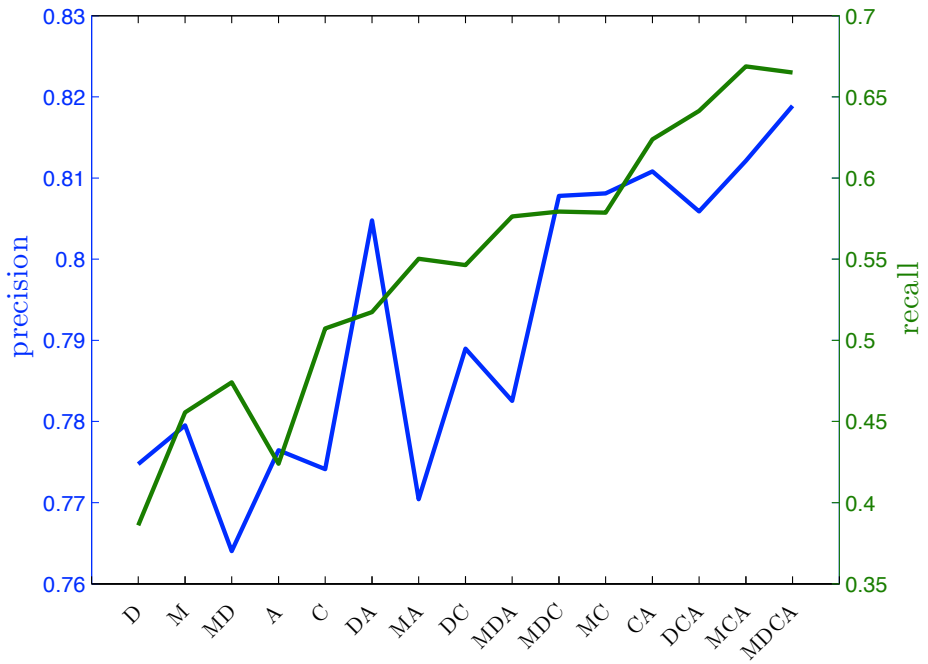


Figure 8.9: Multi-media corpus: precision and recall of the optimal model depending on the linguistic context.

8.4 Modelling Alternatives to Vary Speech Prosody

Speech synthesis systems based on statistical modelling [Zen et al., 2009] suffer from a certain inconsistency: the model is based on the estimation of the speech parameters associated with stochastic processes, while the synthesis of the speech parameters remains deterministic. Indeed, for a given sentence, the sequence of speech parameters to be inferred is entirely determined by the sequence of linguistic contexts associated with the sentence: the inferred sequence corresponds to the most-likely sequence conditionally to the sequence of linguistic contexts and the model. Thus, to each sentence corresponds one and only one prosodic realization. Consequently, conventional methods model the characteristics of a specific speaker (speaker-dependency) solely, without accounting for variability (speaker-variability).

A speaker has various ways to realize an utterance, depending on his *speaking style*, his *prosodic strategy*, and the context of the speech communication. A speaker does not have a single strategy, but rather a *strategic space*, i.e. a variety of alternative strategies that can be all potentially realized. For instance, a professional speaker (e.g., story telling, theatre) continuously varies his speech from one interpretation to the other, and this variation contributes to a large extent to render his speech natural and expressive. The variations of a speaker are usually motivated by the pragmatic and/or discursive context, but may simply result from the arbitrary choice by the speaker within his strategic space. This variability can be observed either in terms of symbolic (prosodic structure) or acoustic variations (prosodic phrasing). Linguistic studies and theoretical models have formally account for prosodic variability in relation to the syntactic-semantic structure of a sentence, and any sentence can be associated with a variety of likely alternatives (for instance, see [Delais-Roussarie, 2000] for French).

In real speech synthesis applications, the worst-case scenario can be observed in announcement systems, in which a single sentence needs to be synthesized and repeated a large number of times to human listeners. The synthesized speech prosody is often perceived as unnatural due to the absence of variations in speech prosody, regardless of its quality. Thus, modelling and exploiting the prosodic variability of a speaker would significantly improve the naturalness of a speech synthesis system by modelling its variety. Finally, both speech prosody *quality* and *variety* shall be distinguished in the modelling of speech prosody. Firstly, the conventional approach aims to optimize the quality of the synthesized speech prosody regardless of its variety. Secondly, the proposed approach aims to optimize the variety of alternatives that can be synthesized while preserving the quality of the synthesized speech prosody. Finally, the quality and the variety of a speech prosody contribute to the perception of the *naturalness* of a speech synthesis system.

In this section, a method to vary the prosody of a speaker in speech synthesis based on the *Generalized Viterbi Algorithm* (GVA) is proposed. The symbolic description of speech prosody is used to model the characteristics of a speaker. The proposed approach is based on the modelling and the synthesis of various alternatives of speech prosody for a given text. During the training, a context-dependent discrete HMM is used to model the prosodic strategies of a speaker (section 8.2). During the synthesis, the speech prosody parameters are usually determined using the conventional *Viterbi Algorithm* (VA) ([Forney, 1973]) so as to determine the most likely speech prosody given the linguistic context sequence ([Veilleux et al., 1990, Black and Taylor, 1994, Ross and Ostendorf, 1996]) (or extension of the Viterbi Algorithm in the case of hierarchical [Ostendorf and Veilleux, 1994] and segmental [Black and Taylor, 1997a, Schmid and Atterer, 2004] HMMs). In the proposed approach, the *Generalized Viterbi Algorithm* (GVA) ([Hashimoto, 1987]) is introduced to provide various alternatives that can be used to vary the prosody of a speaker in speech synthesis. The proposed method is validated with objective and subjective evaluations.

8.4.1 The *Generalized Viterbi Algorithm* (GVA)

In conventional HMM-based applications, it is only of interest to find the single optimal state sequence, i.e. to maximize the conditional probability $p(\mathbf{q}|\mathbf{o},\boldsymbol{\lambda})$ of the sequence $\mathbf{q}$ given observations

o and model $\boldsymbol{\lambda}$:

$$\widehat{\mathbf{q}} = \underset{\mathbf{q}}{\text{argmax}}\ \text{p}(\mathbf{q}|\mathbf{o}, \boldsymbol{\lambda}) \tag{8.41}$$

This problem is solved using dynamic programming and Viterbi Algorithm. As the *Viterbi algorithm* [Forney, 1973] defines a trellis in which at each time t corresponds N lists (one per state) of a single candidate, it is called a N-list 1-survivor (N,1) algorithm . The *Generalized Viterbi Algorithm* [Hashimoto, 1987] is a generalization of the Viterbi algorithm to the (N,K) case in which the constraint on the number of survivors is relaxed. The implementation is straightforward from the Viterbi algorithm, substituting in the selection step the N most likely candidates to the single most likely candidate for each state.

Let define:
$\max^k$ the k-th maximum value of a vector, argmax^k its index in the vector, $\delta_t^k(i) = \max^k \text{p}(q_1, ..., q_t = i, o_1, ..., o_t|\boldsymbol{\lambda})$ the probability corresponding to the k-th optimal path which accounts for the partial observation sequence $[o_1, ...o_t]$ and ends in state i at time t;
$\Gamma_t^k(i,j) = \text{p}(q_1, ..., q_{t-1} = i, q_t = j, o_1, ..., o_t|\boldsymbol{\lambda})$ the probability corresponding to the k-th optimal path which accounts for the partial observation sequence $[o_1, ...o_{t-1}]$ and ends in state i at time t-1 and in state j at time t.

The structure of the generalized Viterbi algorithm can be written as follows:

- **initialization**:

$$\delta_1^k(i) = \pi_i b_i(o_1) \qquad k \in [1, K],\ i \in [1, N] \tag{8.42a}$$

$$\psi_1^k(i) = 0 \qquad k \in [1, K],\ i \in [1, N] \tag{8.42b}$$

- **recursion**:
  - induction:

$$\Gamma_t(i,j) = \delta_{t-1}^k(i) a_{i,j} \qquad p \in [1, K] \tag{8.43a}$$

$$i, j \in [1, N] \tag{8.43b}$$

  - selection:

$$\delta_t^p(j) = \left[\max_{\mathbf{i},\mathbf{k}}{}^p\ \Gamma_t(i,j)\right]\ b_j(o_t) \qquad p \in [1, K],\ j \in [1, N] \tag{8.44a}$$

$$\psi_t^p(j) = \underset{\mathbf{i},\mathbf{k}}{\text{argmax}}{}^p\ \Gamma_t(i,j) \qquad p \in [1, K],\ j \in [1, N] \tag{8.44b}$$

$$\tag{8.44c}$$

- **termination**:

$$\widehat{P}^p = \max_{\mathbf{i},\mathbf{k}}{}^p \delta_T^k(i) \qquad p \in [1, K] \tag{8.45a}$$

$$\widehat{q}_T^p = \underset{\mathbf{i},\mathbf{k}}{\text{argmax}}{}^k \delta_T(i) \qquad p \in [1, K] \tag{8.45b}$$

- **sequence backtracking**:

$$\widehat{q_t}^{\,k} = \phi_{t+1}^{k}(\widehat{q}_{t+1}^{k}) \qquad k \in [1, K] \tag{8.46a}$$

An illustration of the Generalized Viterbi Algorithm is presented in figure 8.10 and in table 8.10 with 3 survivors for the sentence: "*Longtemps, je me suis couché de bonne heure.*" ("*For a long time I used to go to bed early.*").

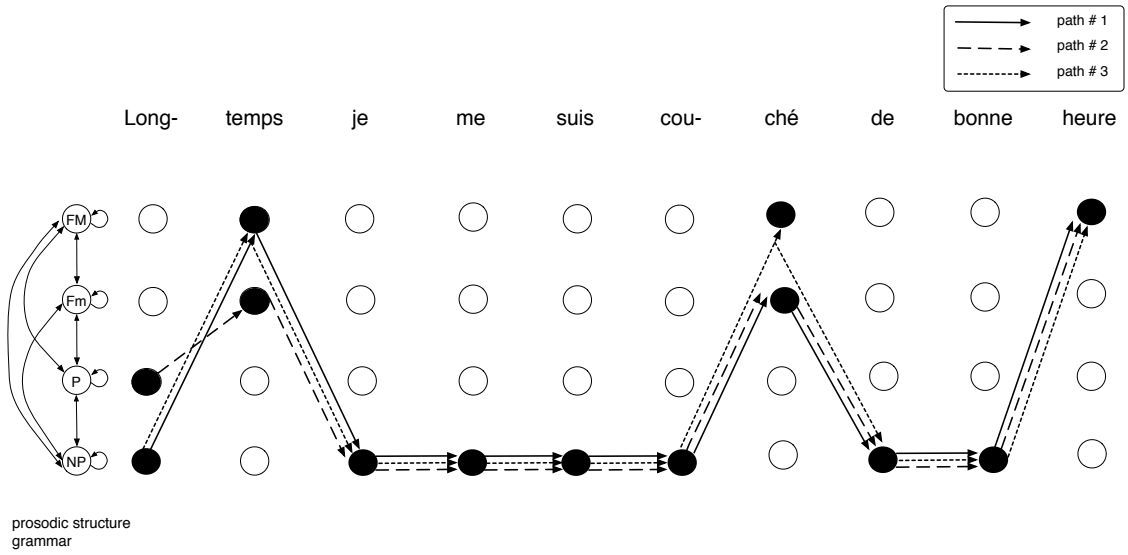


Figure 8.10: Illustration of the *Generalized Viterbi Algorithm* used for the inference of prosodic alternatives for the sentence: "*Longtemps, je me suis couché de bonne heure.*" ("*For a long time I used to go to bed early.*").

sentence	Longtemps,	je	me	suis	couché	de	bonne	heure.		
prosodic structure										
F_M	*							*		
F_m	*					*		*		
P	*					*		*		
syllable	Long-	temps	je	me	suis	cou-	ché	de	bonne	heure

sentence	Longtemps,	je	me	suis	couché	de	bonne	heure.		
prosodic structure										
F_M								*		
F_m	*					*		*		
P	*	*				*		*		
syllable	Long-	temps	je	me	suis	cou-	ché	de	bonne	heure

sentence	Longtemps,	je	me	suis	couché	de	bonne	heure.		
prosodic structure										
F_M	*					*		*		
F_m	*					*		*		
P	*					*		*		
syllable	Long-	temps	je	me	suis	cou-	ché	de	bonne	heure

Table 8.10: Different alternatives as determined by the *Generalized Viterbi Algorithm* for the utterance: "*Longtemps, je me suis couché de bonne heure.*" ("*For a long time I used to go to bed early.*"). (a) 1st most likely state sequence; (b) 2nd most likely state sequence ; (c) 3rd most likely state sequence.

8.4.2 Evaluation

8.4.2.1 Objective evaluation

No method exists to properly integrate variability into conventional objective evaluation schemes, and there is currently no known solution to this problem. Conventional objective evaluation

methods are based on the comparison of a single observed sequence and a single inferred sequence (1-to-1 comparison). In particular, the set of possible alternatives of the observed sequence remains unknown to the observation. Consequently, the performance measure is inadequate since the inferred sequence may match one of the possible alternatives, while actually not the observed prosodic realization. Thus, no method exists to determine automatically and accurately how likely is the inferred sequence. A solution to this problem has been proposed by designing a speech database in which several realizations of each utterance are observed [Ostendorf and Veilleux, 1994]. Then, the single inferred sequence can be compared to the set of possible alternatives (1-to-N comparison). In particular, one can evaluate whether the inferred sequence matches one the observed alternatives.

The actual problem is the complete extension of the variability problem, since a set of inferred candidates is available for the comparison. The ideal case would be to compare the set of inferred candidates with a set of observed alternatives (M-to-N comparison). However, such a case is unrealistic for the evaluation of statistical models that are usually based on large speech databases: the recording of several realizations for each utterance would be very extensive and time consuming. Thus, the proposed evaluation method will be limited to the comparison of a set of inferred candidates with the single observed sequence (M-to-1 comparison). The M-to-1 comparison is the inverse problem of the 1-to-N comparison mentioned above: the evaluation is used to estimate whether one of the possible inferred candidates matches the observed sequence.

The proposed evaluation scheme has been defined as follows:

- each of the inferred sequences is individually compared to the observed sequence;
- the observed sequence is compared to the inferred sequence which best matches with it among the set of inferred sequence candidates. Such a sequence candidate will be referred as the *optimal sequence*.

Evaluation was conducted using the evaluation scheme described in section 8.2.5.

Linguistic Contexts

Linguistic information were extracted from text using the linguistic processing chain described in chapter 7 that includes surface and deep syntactic parsing. chain described in chapter 7. Models were compared with respect to the following linguistic feature sets: morpho-syntactic (M), dependency (D), constituency (C), and adjunction (A) syntactic features. The used linguistic units were: syllable, and the syntactic units. Linguistic features were converted into linguistic contexts over the syllable by computing locational and weight contexts, and representing 1-order left-to-right contexts and 1-order child-to-parent contexts in the case of the dependency contexts.

Each set of linguistic contexts was derived by adding a richer syntactic description to the previous set. The morpho-syntactic context set will be referred as the baseline set for the evaluation.

Evaluation Corpus

Speaker-dependent models were trained and evaluated on the laboratory and multi-media speech databases.

stream	prosodic structure	
corpus		
training corpus	(K-1)/K laboratory corpus (8h) (K-1)/K multi-media corpus (4h30)	
evaluation corpus	1/K laboratory corpus (1h) 1/K multi-media corpus (30mn)	
feature extraction		
feature	hierarchical prosodic structure F_M, F_m, P	
window	syllable	
frame rate	syllable	
feature transform		
transform	linearization	
unit	syllable	
model		
context	**M**: morpho-syntactic context	$Q^{(syllable)}_{morpho}$
	D: dependency context	$Q^{(syllable)}_{dep}$
	C: constituency context	$Q^{(syllable)}_{chunk}$
	A: adjunction context	$Q^{(syllable)}_{adj}$
clustering	DT CART	
topology	discrete HMM ergodic	

Table 8.11: Evaluation of the *symbolic model* using *Generalized Viterbi Algorithm*: model setup

Evaluation Metrics

The evaluation metrics were the same at those used in section 8.2. For each sentence of the evaluation set, the 15-th most likely prosodic sequence candidates were inferred according to the models and compared to the observed prosodic sequence. The optimal sequence was defined as being the sequence that maximizes the *Linear Cohen's Kappa* of the inferred sequence candidates. Argument of the optimal sequence was additionally determined.

Results and Discussion

Overall performance measures are presented in table 8.12 for the full linguistic context.
Performance of each of the inferred sequence candidates is represented in figure 8.12. As expected, the performance is decreasing as a function of the argument of the inferred sequence candidate. This confirms that the likelihood is consistent to the performance measure, thus a reliable criterion for the modelling of speech prosody.

A comparison of the optimal sequence and each of the inferred sequence candidates for the full linguistic context is presented in figure 8.12 and a comparison of the optimal sequence and the conventional most-likely sequence is presented in figure 8.11. The optimal sequence dramatically outperforms any of the inferred sequence candidate regardless to the linguistic context, and in particular the conventional most-likely sequence (89.3% and 58.5% compared to 70.5% and 49.4% with the full linguistic context for the laboratory and the multi-media speech database, respectively). This shows evidence that the proposed method is able to produce reliable prosodic

corpus	laboratory					multi-media				
	$\mathbf{F_1}$	$\mathbf{w-\kappa}$	$\boldsymbol{\kappa_{F_M}}$	$\boldsymbol{\kappa_{F_m}}$	$\boldsymbol{\kappa_P}$	$\mathbf{F_1}$	$\mathbf{w-\kappa}$	$\boldsymbol{\kappa_{F_M}}$	$\boldsymbol{\kappa_{F_m}}$	$\boldsymbol{\kappa_P}$
optimal	**85.8**	**89.3**	**96.3**	**84.0**	**63.3**	**59.3**	**58.5**	**75.8**	**38.1**	**23.4**
path #1	67.3	70.5	94.3	46.3	30.5	53.6	49.4	72.2	25.0	17.4
path #2	67.1	69.6	94.0	45.1	30.9	53.5	49.5	71.8	25.3	17.3
path #3	66.1	67.8	93.7	42.6	30.0	53.4	49.3	71.8	25.1	17.1
path #4	65.8	67.1	93.3	42.6	29.3	53.5	49.4	71.6	25.4	17.3
path #5	65.4	66.2	93.1	40.8	29.6	53.3	49.1	71.7	24.4	17.2
path #6	65.1	65.9	92.7	41.1	28.9	53.4	49.0	71.6	24.4	17.6
path #7	64.9	65.3	92.6	40.2	29.1	53.3	49.0	71.7	25.0	16.7
path #8	64.9	65.3	92.2	41.5	28.2	53.4	49.1	71.4	25.2	17.1
path #9	64.3	64.3	92.0	39.2	28.1	53.1	48.9	71.5	24.8	16.6
path #10	64.3	64.2	91.7	39.5	28.4	53.4	49.0	71.7	25.1	17.0
path #11	64.0	63.6	91.6	38.9	28.1	53.2	48.9	71.3	24.9	17.1
path #12	64.0	63.5	91.3	39.5	27.6	53.5	49.0	71.5	25.0	17.7
path #13	63.5	63.1	91.3	38.1	27.1	53.2	48.8	71.4	25.1	16.5
path #14	63.8	63.2	91.0	39.4	27.3	53.4	48.7	71.3	24.9	17.6
path #15	63.5	63.0	91.1	38.7	26.9	53.2	48.8	71.1	25.2	16.9

Table 8.12: Performances obtained with the full linguistic context set for the optimal sequence and each of the inferred sequence candidates. Performance measures are respectively F_1-*measure*, *linear Cohen's Kappa*, and *Cohen's Kappa* for each of the prosodic events.

corpus	laboratory					multi-media				
	$\mathbf{F_1}$	$\mathbf{w-\kappa}$	$\boldsymbol{\kappa_{F_M}}$	$\boldsymbol{\kappa_{F_m}}$	$\boldsymbol{\kappa_P}$	$\mathbf{F_1}$	$\mathbf{w-\kappa}$	$\boldsymbol{\kappa_{F_M}}$	$\boldsymbol{\kappa_{F_m}}$	$\boldsymbol{\kappa_P}$
morpho										
optimal	**76.1**	**81.2**	**83.1**	**70.3**	**51.4**	**58.5**	**58.2**	**63.4**	**44.2**	**23.3**
most-likely	61.7	62.9	74.1	41.1	31.2	49.2	44.5	54.6	27.5	14.4
dependency										
optimal	**76.2**	**81.3**	**83.4**	**69.4**	**52.2**	**55.3**	**54.2**	**62.5**	**37.4**	**21.0**
most-likely	61.7	63.1	75.0	40.4	31.1	49.4	44.7	55.8	25.1	15.9
constituency										
optimal	**79.3**	**83.3**	**89.9**	**73.5**	**53.8**	**58.3**	**57.5**	**70.1**	**38.4**	**24.1**
most-likely	64.2	65.9	84.1	42.9	30.6	52.1	47.1	65.6	25.0	18.0
adjunction										
optimal	**85.8**	**89.3**	**96.3**	**84.0**	**63.3**	**59.2**	**59.3**	**76.4**	**38.0**	**23.1**
most-likely	67.4	70.5	94.3	46.3	30.5	53.6	49.4	72.2	25.0	17.4

Table 8.13: Comparison of the performance obtained for the optimal sequence and the most-likely sequence depending on the linguistic context. Performance measures are respectively F_1-*measure*, *linear Cohen's Kappa*, and *Cohen's Kappa* for each of the prosodic events.

alternatives since one of these corresponds more closely to the actual realization of the speaker. The most-likely sequence may be simply interpreted as another possible prosodic alternative. Secondly, this observation does not depend on the linguistic context since the relative improvement is comparable for any of the linguistic contexts (figure 8.11).

Argument of the optimal sequence is presented in figure 8.14. The optimal sequence corresponds to the $5 < k < 6$ and $7 < k < 8$ sequence for the laboratory speech and the multi-media speech databases, respectively. There is no significant difference depending on the linguistic context. This indicates that the optimal sequence remains relatively closed to the most-likely sequence, thus that a small set of candidates suffices to produce consistent prosodic alternatives.

Finally, the reduction error obtained for the optimal sequence by comparison with the most-likely sequence for the full linguistic context is presented in figure 8.13 depending on the nature of the

prosodic event. Inferred prosodic alternatives do not affect the prosodic structure homogeneously: alternatives drastically affect minor prosodic boundaries F_m (84.0% and 38.1% compared to 46.3% and 25.0% for the laboratory and the multi-media speech databases respectively), and to a very less extent major prosodic boundaries F_M (96.3% and 75.8% compared to 94.3% and 72.2% for the laboratory and the multi-media speech databases respectively), and prosodic prominences P (63.3% and 23.4% compared to 30.5% and 17.4% for the laboratory and the multi-media speech databases respectively). Not surprisingly, prosodic prominence remains the most difficult prosodic event to model with syntactic information only.

A comparison of the performance obtained for the different speech databases confirms evidence for the previously reported observations: linguistic complexity of the speech database strongly affects the modelling of speech prosody and the ability to reproduce consistent prosodic alternatives. Indeed, the performance obtained for the most-likely sequence and the optimal sequence (figure 8.11), and for the different prosodic events (figure 8.13) is significantly lower for the multi-media corpus compared to the laboratory corpus. Additionally, the increase in performance obtained with the optimal sequence is less pronounced and the argument of the optimal sequence is higher for the multi-media corpus compared to the laboratory corpus. This clearly indicates that the alternatives associated with the different inferred sequence candidates is less consistent (figure 8.12). As mentioned in section 8.2, this may be due to the syntactic complexity and the prosodic complexity of the multi-media speech database. Thus, this may explain that the proposed method partially fails to model the large variety of prosodic strategies used by the professional speaker.

The objective evaluation can be summarized into two main conclusions: 1) the proposed method succeeds in producing consistent prosodic alternatives, 2) that can be inferred from a limited amount of prosodic candidates.

The objective evaluation indicates that the observed prosodic realization can be significantly approached by one of the prosodic alternatives inferred from a limited amount of prosodic sequence candidates. However, this does not explicitly indicate how natural and varied are the prosodic alternatives. To complete the evaluation of the proposed method, a perceptual experiment was conducted to assess whether natural and distinctive are the different prosodic alternatives.

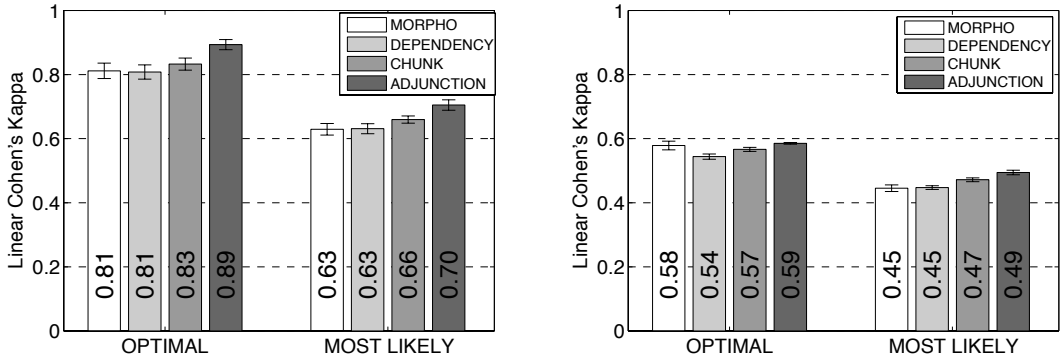


Figure 8.11: Comparison of the mean performance and 95% confidence interval obtained for the optimal sequence and the most likely sequence depending on the linguistic context. On top: laboratory corpus, on bottom: multi-media corpus.

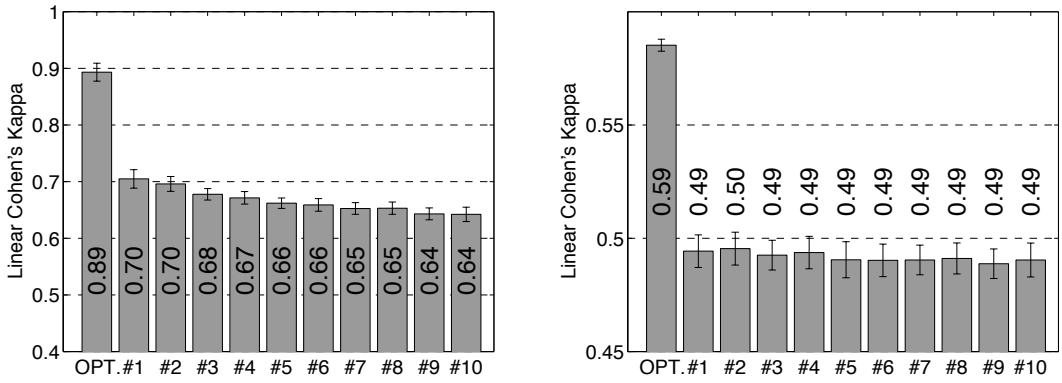


Figure 8.12: Comparison of the mean performance and 95% confidence interval obtained for the optimal sequence and each of the k-th most likely sequences ($k = 10$) with the full linguistic context. On top: laboratory corpus, on bottom: multi-media corpus.

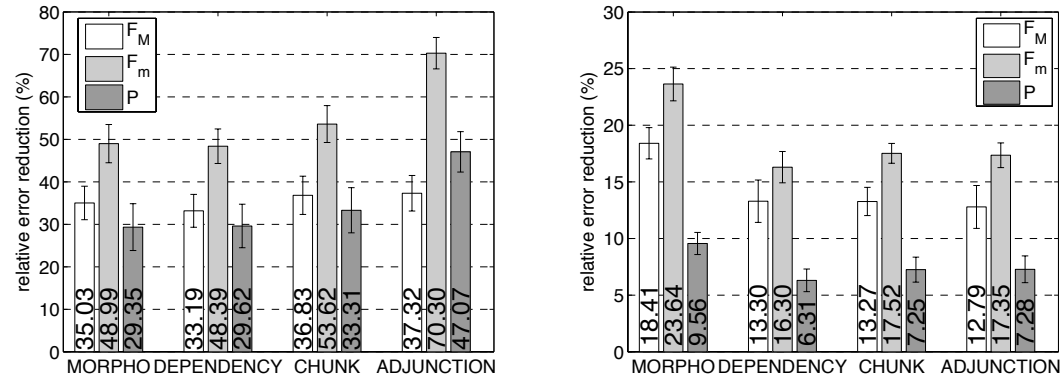


Figure 8.13: Mean performance and 95% confidence interval obtained depending on the prosodic event (F_M,F_m,P) with the full linguistic context. On top: laboratory corpus, on bottom: multi-media corpus.

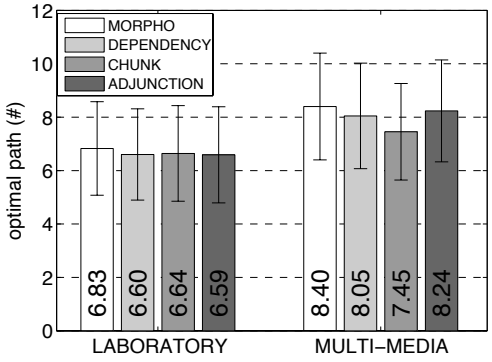


Figure 8.14: Mean argument and 95% confidence interval of the optimal sequence depending on the linguistic context.

8.4.2.2 Subjective evaluation

A subjective evaluation was conducted to assess the *naturalness* and the *distinctiveness* of the inferred prosodic alternatives for the multi-media speech database.

8.4.2.3 Stimuli

The proposed method was integrated into the HTS speech synthesizer [Zen et al., 2009]. Discrete/continuous HMMs were estimated on the multi-media speech corpus with the full linguistic contexts. The evaluation corpus consisted of a set of 8 sentences randomly extracted form the fairy tale: *Le Petit Poucet* (*"Litlle Tom Thum"*), by French writer Charles Perrault [Perrault, 1697]. For each sentence, 3 prosodic alternatives were determined using to the proposed method: the 1st, 3rd, and 5th most-likely prosodic sequences were selected. This was done in order to limit the number of prosodic alternatives to be compared and to emphasize prosodic differences between each of the selected prosodic alternatives. Then, speech was synthesized for each of the prosodic alternatives using the speech synthesizer. This results into 24 utterances to be evaluated: 8 utterances with 3 variations each.

Table 8.14 presents a study case of prosodic alternatives as inferred for one of the utterance used in the evaluation set.

sentence	La chose réussit comme il l'avait pensée. *Things fell out just as he had anticipated.*
alternative #1	(La chose réussit) / (comme il l'avait pensée) //
alternative #2	(La chose réussit) // (comme il l'avait pensée) //
alternative #3	(La **CHOSE** réussit) / (comme il l'avait pensée) //

Table 8.14: Study case of prosodic alternatives determined with the Generalized Viterbi Algorithm. Simple bars denote minor prosodic boundaries, double bars major prosodic boundaries, and bold font prosodic prominences.

stream	source/filter	duration	f_0
corpus			
training corpus	multi-media corpus (5h)		
evaluation corpus	C-TALE text corpus (24 sentences)		
feature extraction			
feature	5-order aperiodicity 39-order MFCC	state-duration	f_0
window	50-ms blackmann		
frame rate	5ms		
feature transform			
transform	-	log	log
dynamic	1-orderΔ ,Δ^2	-	1-orderΔ ,Δ^2
model			
topology	5-state HMM normal distribution semi-tied covariance	5-state HMM normal distribution	5-state MSD-HMM normal distribution semi-tied covariance
context	baseline linguistic context, $\mathrm{Q}_{\mathrm{rich}}^{(phone)} = Q_{\mathrm{adj}}^{\prime(phone)} \cup Q_{\mathrm{proso}}^{(phone)}$		
clustering	DT ML-MDL		

Table 8.15: Evaluation of the *symbolic model* using *Generalized Viterbi Algorithm*: speech synthesizer model setup

8.4.2.4 Participants

20 subjects participated in this experiment: 20 native French speakers; 13 expert participants, 7 naïve participants. Expert participants had a variety of backgrounds: speech and audio technologies, linguistic, musicians.

8.4.2.5 Procedure

The experiment consisted of a subjective evaluation of speech prosody *naturalness* and *distinctiveness* using a *Mean Opinion Scale* (MOS)[3], and was conducted using crowd-sourcing technique on web social networks[4].

[3]the experiment is available at the following link: `http://recherche.ircam.fr/equipes/analyse-synthese/obin/pmwiki/pmwiki.php/Main/HTSMultipleProsoStructure`

[4]*Ircam Analysis and Synthesis Perceptual Experiments* on Facebook : `http://www.facebook.com/group.php?gid=150354679034&ref=ts`

The experiment was divided into two distinct parts as follows:

In the first part, participants were asked to rate the *speech prosody naturalness* of each of the synthesized speech utterances. Sentences to be synthesized were randomly selected. For each synthesized speech utterance, each of the speech alternatives was randomly presented to the participants. Participants were asked to rate the *prosodic naturalness* of each of the speech utterances on a Mean Opinion Scale.

Speech prosody naturalness was referred as:

- a "correct" prosody: the utterance is pronounced as it could be expected from a native speaker.
- a "lively" prosody. The opposite of a lively prosody is a monotone prosody.

Participants were additionally asked to ignore speech synthesis artefacts.

In the second part, participants were asked to rate the *speech prosody distinctiveness* of each pair of alternatives corresponding to a given speech utterance. For each speech utterance, each pair of alternatives was randomly presented to the participants. Participants were asked to rate the *speech prosodic distinctiveness* of each pair of speech utterance alternatives on a Mean Opinion Scale.

Speech prosody distinctiveness was referred as:

- how different are the two speech utterances according to their prosody?

Participants were additionally asked to ignore speech synthesis artefacts.

Finally, additional information were gleaned from the participants : speech expertise (expert, naïve), language (native French speaker, non-native French speaker, non-French speaker), age, and listening condition (headphones or not). Participants were encouraged to use headphones.

Score	Quality	Impairment
5	excellent	imperceptible
4	good	perceptible but not annoying
3	fair	slightly annoying
2	poor	annoying
1	bad	very annoying

Table 8.16: MOS scale for the evaluation of speech prosody naturalness.

Score	Difference
5	perfectly different
4	significantly different
3	fairly different
2	slightly different
1	no difference

Table 8.17: MOS scale for the evaluation of speech prosody distinctiveness.

8.4.2.6 Results and Discussion

Results of the subjective evaluation are presented in figure 8.17.

On the one hand, synthesized prosodic alternatives have been all perceived as being fairly natural (overall: MOS $= 3.32 \pm 0.22$; and MOS $= 3.18 \pm 0.28$, MOS $= 3.44 \pm 0.22$, MOS $= 3.33 \pm 0.22$ for the different alternatives). There is no significant difference across the prosodic alternatives according to their naturalness ($F(2, 57) = 1.08$, $p = 0.35$); in particular, there is no significant difference between the standard most-likely sequence and the other prosodic alternatives. This suggests that the synthesized prosodic alternatives have been perceived as being equally natural.

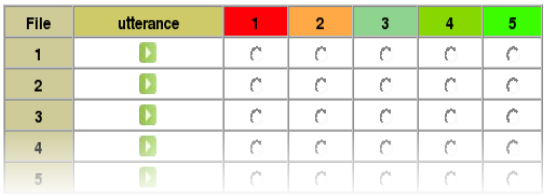


Figure 8.15: Illustration of the web interface used for the evaluation of speech prosody naturalness.

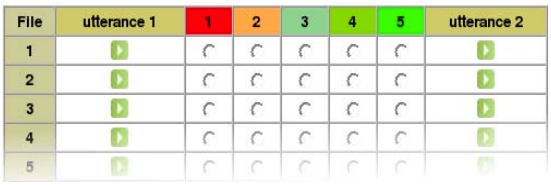
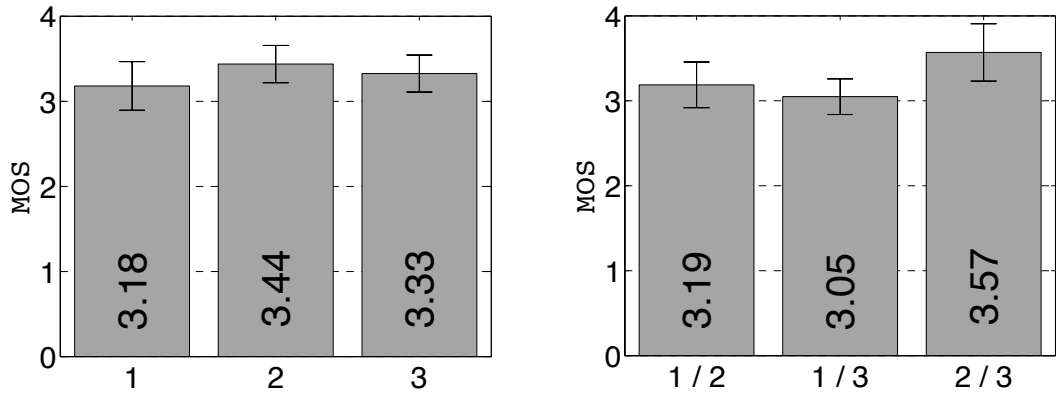


Figure 8.16: Illustration of the web interface used for the evaluation of speech prosody distinctiveness.

On the other hand, the different prosodic alternatives have been perceived as fairly different (MOS $= 3.27 \pm 0.26$). In a same manner as previously, there is no significant difference between each pairwise comparison ($F(2, 57) = 1.53$, $p = 0.23$). However, there is a significant difference between the third alternatives and the others ($F(2, 57) = 4.40$, $p = 0.04$), which suggests that increasing the number of alternatives (K) also increases the difference across the different inferred alternatives. Interestingly, expert linguists perceived the synthesized prosodic alternatives as substantially different (MOS $= 4.00 \pm 0.26$); other experts as fairly different (MOS $= 3.34 \pm 0.16$); naive listeners only slightly different (MOS $= 2.8 \pm 0.56$). As expert linguists are supposed to be more sensitive in the perception of variations in speech prosody, this indicates evidence for substantial differences across the prosodic alternatives.

Finally, the different synthesized speech alternatives have been perceived as equally natural and significantly different from each other. This constitutes a subjective validation of the proposed method for the inference of various prosodic alternatives in speech synthesis.



(a) prosodic naturalness

(b) prosodic distinctiveness

Figure 8.17: MOS/CMOS and 95% confidence interval for the evaluation of (a) speech prosody naturalness, and (b) speech prosody distinctiveness.

8.4.3 Conclusion

In this section, a method to vary the prosody of a speaker in speech synthesis based on the *Generalized Viterbi Algorithm* (GVA) was proposed. The symbolic description of speech prosody was used to model the characteristics of a speaker. The proposed approach was based on the modelling and the synthesis of various alternatives of speech prosody for a given text. During the training, a context-dependent discrete HMM is used to model the prosodic strategies of a speaker. During the synthesis, the *Generalized Viterbi Algorithm* is used to determine various alternatives that can be used to vary the prosody of a speaker in speech synthesis. The proposed method was

validated with objective and subjective evaluations.

In the objective evaluation, the proposed method was shown to generate consistent alternatives that can be inferred from a limited amount of prosodic candidates. Additionally, the optimal alternative was shown to drastically improve the symbolic modelling of speech prosody compared to the conventional Viterbi algorithm. In the subjective evaluation in speech synthesis, the determined alternatives were perceived as equally natural and significantly different from each other. In particular, expert listeners perceived the alternatives as substantially distinct. Objective and subjective evaluations provide evidence that the proposed method succeeds in modelling the speech prosody strategies of a speaker, and speech prosody alternatives in speech synthesis.

In further studies, the proposed method will be used to improve either the variety of the speech prosody or the quality of the synthesized speech. First, the *Generalized Viterbi Algorithm* will be combined to the segmental HMM to improve the discrete modelling of speech prosody, and the variety of the synthesized alternatives. Second, the symbolic modelling of speech prosody will be combined to the acoustic modelling of speech to improve the quality of speech synthesis, either based on HMMs or on unit-selection [Bulyko and Ostendorf, 2001]. Finally, a unified model will be proposed to integrate the symbolic and acoustic modelling of speech prosody and the modelling of the acoustic speech characteristics into a single modelling/decoding framework so as to improve simultaneously the quality and the variety of the synthesized speech.

8.5 Conclusion

In this chapter, the symbolic modelling of speech prosody, with a particular focus on the combination of statistical modelling methods and linguistic theory, has been presented. The RHAPSODIE transcription system was developed and used to represent the symbolic characteristics of speech prosody based on the perception of prosodic prominences and prosodic boundaries. A context-dependent discrete HMM was used to model the symbolic characteristics of speech prosody in context. During the training, the text is first converted into a sequence of linguistic contexts using the linguistic processing chain that includes surface and deep syntactic parsing. Linguistic contexts are clustered using a Decision-Tree so as to minimize the entropy of the prosodic events. Then, a discrete HMM is estimated for each terminal node of the context-dependent tree. During the synthesis, the text is first converted into a sequence of concatenated context-dependent models. Then, the sequence of prosodic events is determined so as to maximize the conditional probability of the sequence of prosodic events given the sequence of linguistic contexts and the models.

Firstly, the role of the linguistic context in the modelling of speech prosody was assessed using the conventional context-dependent discrete HMM. Secondly, a method that combines linguistic and metric constraints for prosodic break modelling was proposed based on segmental HMMs and Dempster-Shafer fusion, and the relative importance of linguistic and metric constraints was assessed depending on the nature of the linguistic information. Finally, a method to vary the speech prosody of a speaker was proposed based on the *General Viterbi Algorithm* (GVA). The proposed methods were either objectively and/or subjectively evaluated with two speaker-dependent speech databases.

The rich linguistic description has been demonstrated to dramatically improve the modelling of speech prosody, and has shown that the performance gradually increases with the enrichment of the linguistic description. In particular, adjunctions (e.g., relative clauses, incises) and to a lesser extent constituency proved to be highly-reliable syntactic cues for the symbolic modelling of speech prosody, especially for major prosodic boundaries.

The combination of linguistic and metric constraints into segmental HMM was shown to significantly outperform the conventional segmental HMM and the linguistic model only in the modelling of prosodic breaks. The linguistic constraint was shown to be prior to the metric constraint, and the optimal configuration to gradually tend to the linguistic constraint when the linguistic description is enriched, or when the segment constraint is slightly reliable. The conventional segmental-HMM was found to outperform the linguistic model only when the linguistic information used is slightly relevant. Finally, the increase in performance obtained by the integration of the metric constraint and its combination with the linguistic constraint remains relatively slight compared to that obtained with the enrichment of the linguistic description.

The modelling of speech prosody strategies was proved objectively to produce consistent speech prosody alternatives that can be inferred from a limited amount of speech prosody candidates. In particular, the optimal speech prosody alternative was shown to drastically improve speech prosody modelling compared to the conventional Viterbi algorithm. Speech prosody alternatives were proved to be perceived as equally natural and consistently different in a subjective evaluation in speech synthesis.

Chapter 9

Continuous Modelling of Speech Prosody

Contents

9.1 Introduction

Alongside the development of high-quality speech synthesis systems [Zen et al., 2009], the modelling of speech prosody has emerged as a major concern to improve the naturalness, the liveliness, and the variety of the synthetic speech. Speech prosody is generally described as the co-occurrence of acoustic gestures occurring simultaneously over different temporal domains [Fujisaki, 1983, Van Santen and Moebius, 1999] and associated to different communicative functions (linguistic, expressive). High-quality modelling of speech prosody is desirable for natural and expressive speech synthesis and adequate modelling of speaking style, and a prerequisite in real multi-media applications (e.g., avatars, story telling, dialogue systems, digital arts).

A variety of methods have been proposed to model speech prosody (f_0 [Yoshimura et al., 1999], temporal structure [Zen et al., 2004]), and local and global variations (*Global Variance* (GV) [Toda and Tokuda, 2007, Toda and Young, 2009], *Minimum Generation Error* (MGE) [Qin et al., 2009], or *Rich Context Modelling* [Yan et al., 2009]). However, conventional methods usually model the short-term variations of speech prosody (*frame-based*, or *instantaneous variations*), while the long-term variations of speech prosody are not explicitly considered. Historically, a number of methods have been proposed to model the long-term variations of speech prosody, in particular for the modelling of f_0 variations (for French, [Aubergé, 1991, Morlec, 1997, Holm, 2003]). Recent studies have proposed to integrate long-term variations into HMM modelling, either for the modelling of f_0 variations [Latorre and Akamine, 2008, Qian et al., 2009], or with extension to state-duration modelling [Gao et al., 2008]. However, the proposed methods remain *mixed* models, i.e. the conventional model is used to model the *instantaneous* variations of f_0, while stylization of long-term variations are used as trajectory constraints only. In particular, *instantaneous* variations remain the minimal and target temporal domain for the modelling of speech prosody.

The combination of short and long-term variations into context-dependent HMM causes a number of problems for the analysis and statistical modelling of speech prosody. Each temporal domain supports a specific set of linguistic units and their characteristics. Thus, the modelling of speech prosody requires the identification of temporal domains over which relevant speech prosody variations are observed, and the determination of the contexts (i.e., linguistic units and the associated characteristics) that can be used to model the observed variations. However, prosodic domains remain linguistically ill-defined and there are no known methods to decompose the observed speech prosody variations accurately over the different prosodic units. Moreover, the formal relationship that exists between prosodic units and syntactic units, and the identification of the linguistic contexts that are relevant for the description of speech prosody, are extremely complex.

In this chapter, a *unified* trajectory model based on the stylization and the simultaneous modelling of f_0 variations over various temporal domains is presented. In the proposed approach, the syllable is used as the minimal temporal domain for the description of speech prosody, and f_0 variations are stylized and modelled simultaneously over various temporal domains which cover short-term and long-term variations. During the training, a context-dependent model is estimated according to the joint stylized f_0 contours over the syllable and a set of long-term temporal domains, and the clustering of context-dependent models is driven by long-term trajectories. During the synthesis, f_0 variations are determined using the long-term variations as trajectory constraints.

To overcome the problems associated with long term modelling, the proposed method models simultaneously speech prosody variations on a set of arbitrary and/or prosodically-motivated temporal domains. The syllable is defined as the minimal prosodic unit for the description of speech prosody, and the stylization of prosodic contours over the different temporal domains and the linguistic description of the various linguistic units are shared over the syllable.

The proposed method presents several differences compared to the conventional HMM-based model:

temporal domain the syllable is defined as the minimal temporal domain for the description and

the modelling of speech prosody, while the phoneme and phoneme partition are used in the conventional HMM-based model (*hidden semi-Markov model* [Zen et al., 2004]).

signal model short-term and long-term variations are simultaneously modelled over relevant temporal domains, and prosodic contours are explicitly stylized using a *Discrete Cosine Transform* (DCT), while short-term variations are modelled solely in the conventional HMM-based model (frame-based [Yoshimura et al., 1999]). In particular, speech prosody is entirely modelled in the stylized domain (contours).

trajectory model long-term trajectories are used as trajectory constraints, while local trajectory constraints are used in the conventional HMM-based model (*Trajectory Model* [Tokuda et al., 2003], partial derivatives of the instantaneous f_0).

The chapter is organized as follows: stylization of speech prosody and conventional context-dependent continuous modelling are described in section 9.2. Trajectory modelling of short and long term speech prosody variations is presented in section 9.3. The role of the linguistic context and the trajectory model are evaluated and discussed in section 9.4.

9.2 Context-Dependent Continuous HMM

9.2.1 Stylization of Speech Prosody

The *Discrete Cosine Transform* (DCT) is used to stylize the f_0 variations over various temporal domains [Teutenberg et al., 2008] (chapter 3). The principle of the DCT is to decompose speech prosody contours on a basis of slowly time-varying functions defined by zero-phase cosine functions $\boldsymbol{\phi} = (\cos(\omega_1), \ldots, \cos(\omega_K))$ at discrete frequencies $\omega_k = \frac{\pi}{2T}(2k+1)$.

Two classes of temporal domains are defined for the stylization of f_0 variations:

Fixed-order syllable context accounts for f_0 variations occurring on the syllable and its immediate context (0-order represents the f_0 variations over the syllable, 1-order the f_0 variations over the 1-left-to-right syllable context, ...);

Linguistic units accounts for f_0 variations occurring over long-term temporal domains (e.g., minor/major prosodic groups).

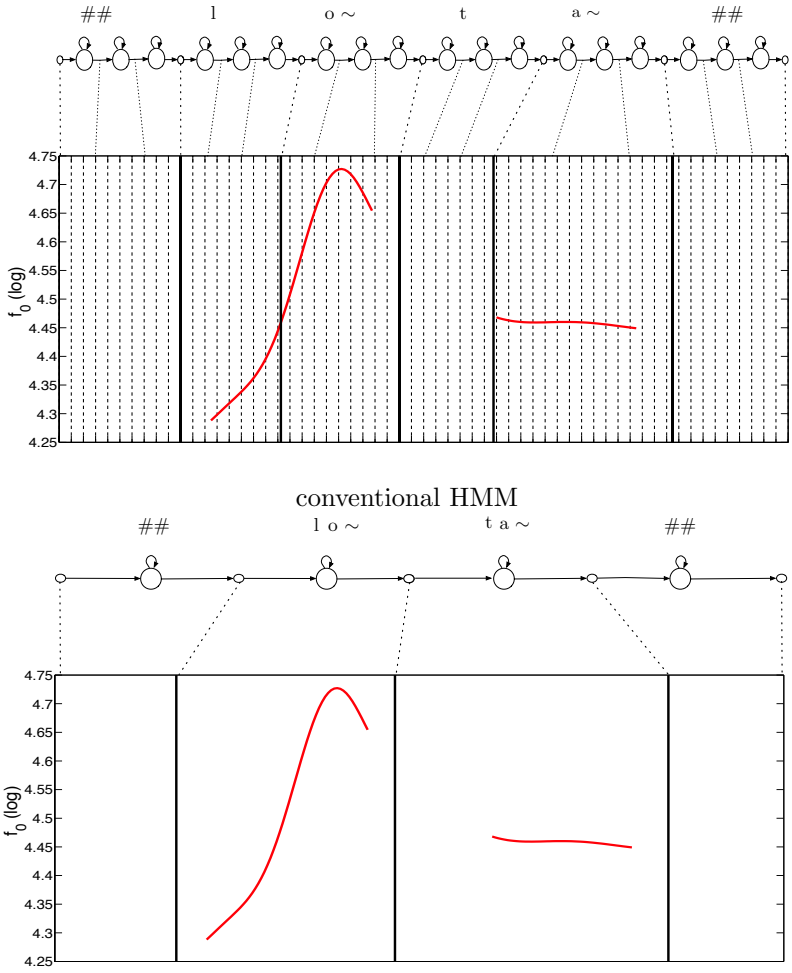
For each of temporal domain, f_0 variations are stylized using a 5-order *Discrete Cosine Transform*. F_0 is linearly interpolated in the logarithmic domain prior to the stylization. The stylization over various temporal domains aims at representing f_0 variations with more or less details, and to model short and long term dependencies. An illustration of f_0 stylization is presented in figures 9.2 and **??**.

9.2.2 Parameters Estimation

Let $\mathbf{q} = [\mathbf{q}_1, \ldots, \mathbf{q}_T]$ be the sequence of linguistic contexts, where $\mathbf{q}_t = [q_t(1), \ldots, q_t(L)]^\top$ is a (Lx1) linguistic vector which describes the linguistic characteristics associated with the t-th syllable.
Let $\mathbf{o} = [\mathbf{o}_1, \ldots, \mathbf{o}_T]$ be the sequence of stylized speech prosody contours over the syllable, where $\mathbf{o}_t = [o_t(1), \ldots, o_T(D)]^\top$ is a (Dx1) observation vector which describes the acoustic characteristics associated with a given prosodic dimension and the t-th syllable.

A HMM $\boldsymbol{\lambda}_\mathbf{q}$ is estimated for each of the linguistic contexts. Each of the context-dependent HMMs is assumed to be a single-state HMM over the syllable with single normal distribution and diagonal covariance matrix.

Continuous HMMs for duration are estimated according to syllable duration in the logarithmic domain. Continuous HMMs for f_0 are estimated according to the Discrete Cosine Transform of linearly interpolated f_0 in the logarithmic domain over the syllable and a set of long-term temporal domains.



Context-dependent HMM with stylization of f_0 contours over syllable.

Figure 9.1: Schematic comparison of *frame-based* and *syllable-based* modelling of f_0 variations.

9.2.3 Decision-Tree-Based Context-Clustering

A number of methods have been proposed to cluster HMM models and share model parameters among linguistic contexts [O'Dell, 1995, Yoshimura et al., 1999, Shinoda and Watanabe, 2000]. In this section, a decision-tree-based context-clustering method based on *Maximum-Likelihood Minimum-Description-Length* (ML-MDL) is described.

9.2.3.1 *Maximum-Likelihood Minimum-Description-Length* Decision-Tree Context-Clustering (ML-MDL)

Maximum-Likelihood Minimum-Description-Length is one of the most commonly used criterion to derive a decision-tree-based context-clustering in speech recognition [Shinoda and Watanabe, 2000] and speech synthesis [Yoshimura et al., 1999]. First, the maximum-likelihood objective function used to derive the model tree typology is consistent with HMM-based methods. Second, the minimum-description-length criterion is a reliable criterion for model selection [Rissanen, 1984], which has been formally described in the case of normal distributions [Schwarz, 1978] and decision-tree maximum-likelihood context-clustering [Shinoda and Watanabe, 2000]. In particular, the minimum-description-length is a criterion for model selection in which the likelihood of a model is additionally regularized by the complexity of the model.

Let T be a binary tree with root node S_0 and leaf nodes $\mathbf{S} = (S_1, \ldots, S_M)$, and $\boldsymbol{\lambda}_{\boldsymbol{S}} = (\lambda_{S_1}, \ldots, \lambda_{S_M})$ the model associated to the set of leaf nodes S, where M is the number of leaf nodes.

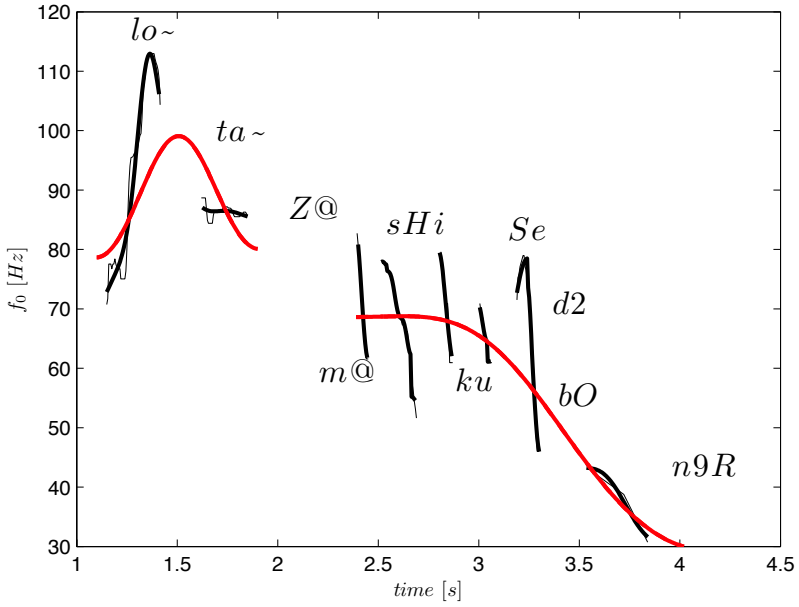


Figure 9.2: Stylization of melodic contours over various temporal domains with a 5-order Discrete Cosine Transform for the utterance: "*Longtemps, je me suis couché de bonne heure.*" ("For a long time I used to go to bed early."). Thin black line represents the observed f_0 variations. Bold black line represents syllable contours. Bold red line represents phrase contours.

Let L(S_m) denote the log-likelihood of model λ_{S_m} given the observation sequences O_m associated with the contexts corresponding to the node S_m.

The log-likelihood of the model $\boldsymbol{\lambda_S}$ is given by:

$$L(S) = \sum_{m=1}^{M} L(S_m) \tag{9.1}$$

The description length of the model $\boldsymbol{\lambda_S}$ is given by:

$$D(S) \quad = \quad -L(S) + \frac{1}{2} k \log\Gamma(\,S) + \log I \tag{9.2}$$

where $\Gamma = \sum_{m=1}^{M} \Gamma_m$ is the total state occupancy probability,$\Gamma_{\ m} = \sum_{t=1}^{T} \gamma_t(m)$ is the total state occupancy probability at node S_m, $\gamma_t(m)$ is the state occupancy probability at node S_m (equation 6.30), k is the number of free parameters, and I is the number of possible models.

In equation (12.4), the first term represents the negative of the model log-likelihood, the second term represents the model complexity, and the third term represents the code length required to encode the model $\boldsymbol{\lambda_S}$. This last term will be assumed constant in the following.

Assuming that the covariance of each gaussian probability density function is diagonal, the description length of the model can be rewritten as:

$$D(S) \quad = \quad -L(S) + dM \log\Gamma(S) + \log I \tag{9.3}$$

where d is the dimensionality of the observation feature vector.

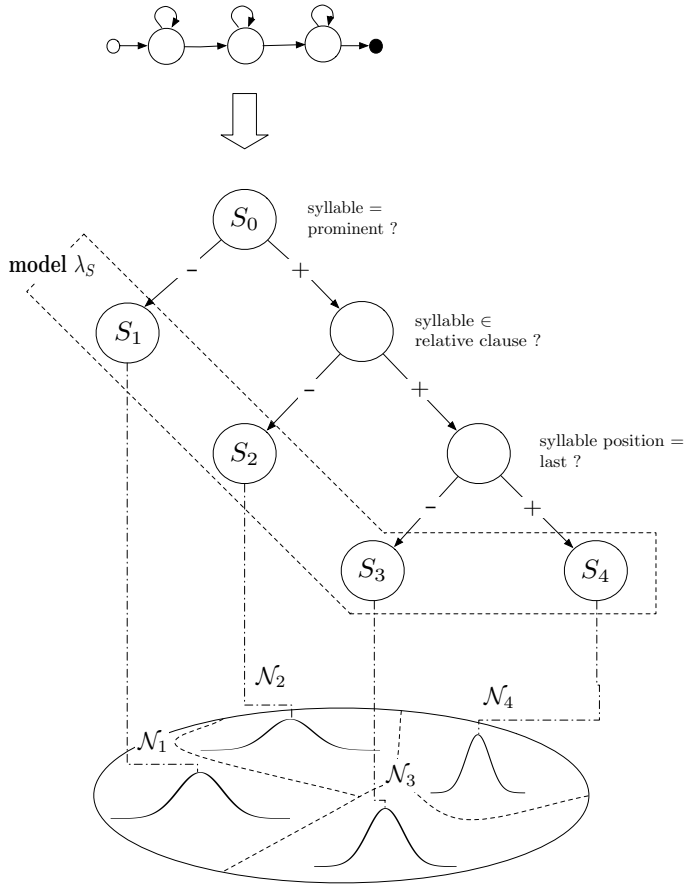


Figure 9.3: Schematic example of decision-tree-based context-clustering of continuous models of speech prosody.

The increase of model log-likelihood $L(S')$ by splitting leaf node S_m through question q into nodes $S_{m,q+}$ and $S_{m,q-}$ is given by:

$$\begin{aligned}\Delta_L^q(S') &= L(S_{m,q+}) + L(S_{m,q-}) - L(S_m) \\ &= -\frac{1}{2}\Big(\Gamma(S_{m,q+})\log|\Sigma(S_{m,q+})| + \Gamma(S_{m,q-})\log|\Sigma(S_{m,q-})| - \Gamma(S_{m,q})\log|\Sigma(S_{m,q})|\Big) \end{aligned} \tag{9.4}$$

where$\Gamma(\ .)$ and $\Sigma(.)$ denote the total state occupancy probability and the covariance matrix in tree node, respectively.

The change in model description length $D(S')$ by splitting leaf node S_m through question q into nodes $S_{m,q+}$ and $S_{m,q-}$ is given by:

$$\Delta_{MDL}^q(S') = -\Delta_L^q(S') + dM(\log\Gamma(S_0)) \tag{9.5}$$

where$\Gamma(\ .)$ denotes the total state occupancy probability in tree node, and d the dimensionality of the observation feature.

The question $\widehat{q}_{MDL}$ which minimizes the increase of model description length at node S_m is given by:

$$\widehat{q}_{MDL} = \underset{\mathbf{q}}{\text{argmax}} \ -\Delta_{MDL}^q(S) \tag{9.6}$$

The context-dependent tree is then derived as follows:

1. **tree initialization**

$$\begin{aligned} T^{(0)} &= T_0 \\ S^{(0)} &= S_0 \\ \boldsymbol{\lambda_S}^{(0)} &= \lambda_{S_0} \end{aligned}$$

2. **tree recursion**

   for each leaf node S_m of the context-tree $T^{(i)}$

   tree selection

   (a) description length calculation:$\Delta\ ^{q}_{MDL}(S)$, $q \in [1, Q]$
   (b) optimal splitting context: $\widehat{q}_{MDL} = \underset{\mathbf{q}}{\text{argmax}}\ -\Delta^{q}_{MDL}(S)$

   tree derivation

   if$\Delta\ ^{\widehat{q}}_{MDL}(S) < 0$, split node S_m and model parameters λ_{S_m}:

$$\begin{aligned} S'_m &\leftarrow (S_{m,\widehat{q}-}, S_{m,\widehat{q}+}) \\ \lambda S'_m &\leftarrow (\lambda_{S_{m,\widehat{q}-}}, \lambda_{S_{m,\widehat{q}+}}) \end{aligned}$$

   tree update

$$\begin{aligned} T^{(i+1)} &= T' \\ S^{(i+1)} &= S' \\ \boldsymbol{\lambda_S}^{(i+1)} &= \boldsymbol{\lambda_{S'}} \end{aligned}$$

3. **tree termination**

$$\begin{aligned} \widehat{T} &= T^{(i)} \\ \widehat{S} &= S^{(i)} \\ \widehat{\lambda_S} &= \lambda_S^{(i)} \end{aligned}$$

An example of decision-tree-based minimum-description-length context-clustering is presented in figure 9.3.

9.2.4 Parameters Inference

9.2.4.1 Formulation of the problem

During the synthesis, the text is first converted into a sequence of concatenated context-dependent HMMs $\boldsymbol{\lambda}$. Then, the prosodic sequence $\mathbf{o} = [\mathbf{o}_1, \ldots, \mathbf{o}_T]$ is determined so as to maximize the probability of the observation sequence $\mathbf{o}$ conditionally to the sequence of context-dependent HMMs $\boldsymbol{\lambda}$ and the sequence length T [Tokuda et al., 2000].

$$\widehat{\mathbf{o}} = \underset{\mathbf{o}}{\text{argmax}}\ \mathrm{p}(\mathbf{o}|\boldsymbol{\lambda}, T) \tag{9.7}$$

$$= \underset{\mathbf{o}}{\text{argmax}} \sum_{\mathbf{q}} \mathrm{p}(\mathbf{o}, \mathbf{q}|\boldsymbol{\lambda}, T) \tag{9.8}$$

$$\widehat{\mathbf{o}} \simeq \underset{\mathbf{o}}{\text{argmax}} \max_{\mathbf{q}} \mathrm{p}(\mathbf{o}, \mathbf{q}|\boldsymbol{\lambda}, T) \tag{9.9}$$

Using Bayes' theorem,

$$\widehat{\mathbf{o}} = \underset{\mathbf{o}}{\text{argmax}} \max_{\mathbf{q}} \mathrm{p}(\mathbf{o}|\mathbf{q},\boldsymbol{\lambda}, T)\mathrm{p}(\mathbf{q},\boldsymbol{\lambda}, T) \tag{9.10}$$

Thus, the determination of the optimal observation sequence $\mathbf{o}$ given the model $\boldsymbol{\lambda}$ and the sequence length T divides into the following two problems:

$$\widehat{\mathbf{q}} = \underset{\mathbf{q}}{\operatorname{argmax}}\ \mathrm{p}(\mathbf{q}|\boldsymbol{\lambda}, T) \tag{9.11}$$

$$\widehat{\mathbf{o}} = \underset{\mathbf{o}}{\operatorname{argmax}}\ \mathrm{p}(\mathbf{o}|\widehat{\mathbf{q}},\boldsymbol{\lambda}) \tag{9.12}$$

The first problem consists in the determination of the optimal state sequence $\mathbf{q}$ given the model $\boldsymbol{\lambda}$ and the sequence length T. The second problem consists in the determination of the optimal observation sequence $\mathbf{o}$ given the state sequence $\widehat{\mathbf{q}}$, and the sequence of context-dependent models $\boldsymbol{\lambda}$.

9.2.4.2 Determination of the optimal state sequence

The first problem is to determine the optimal state sequence $\widehat{\mathbf{q}} = [\widehat{q_1}, \ldots, \widehat{q_T}]$ given the sequence of context-dependent models $\boldsymbol{\lambda}$, and the sequence length T. Thus, the state sequence $\widehat{\mathbf{q}}$ is determined so as to maximize the probability of the state-sequence $\mathbf{q}$ conditionally to the sequence of context-dependent models $\boldsymbol{\lambda}$ and the sequence length T:

$$\widehat{\mathbf{q}} = \underset{\mathbf{q}}{\operatorname{argmax}}\ \mathrm{p}(\mathbf{q}|\lambda, T) \tag{9.13}$$

In the conventional HMM-based speech synthesis, the state sequence $\widehat{\mathbf{q}}$ is given by the mean sequence of the state-duration probabilities that corresponds to the sequence of context-dependent models $\boldsymbol{\lambda}$ [Zen et al., 2004]:

$$\widehat{\mathbf{q}} = [\mathbf{q}_{1,[1:\widehat{d_1}]}, \ldots, \mathbf{q}_{N,[T-\widehat{d_N}+1:T]}] \tag{9.14}$$

where $\widehat{\mathbf{d}}$ denotes the mean sequence of state-duration probabilities that corresponds to the sequence of context-dependent models $\boldsymbol{\lambda}$:

$$\widehat{\mathbf{d}} = [\widehat{\mathbf{d}_1}, \ldots, \widehat{\mathbf{d}_N}] \tag{9.15}$$

$\widehat{\mathbf{d}_n}$ denotes the mean state-duration that corresponds to the n-th context-dependent model, and N the length of the sequence of context-dependent models $\boldsymbol{\lambda}$.

9.2.4.3 Determination of the optimal observation sequence

The second problem is to determine the optimal observation sequence $\widehat{\mathbf{o}} = [\widehat{\mathbf{o}}_1, \ldots, \widehat{\mathbf{o}}_T]$ given the state sequence $\widehat{\mathbf{q}} = [\widehat{q_1}, \ldots, \widehat{q_T}]$, and the sequence of context-dependent models $\boldsymbol{\lambda}$. Thus, the observation sequence $\widehat{\mathbf{o}}$ is determined so as to maximize the probability of the observation sequence $\mathbf{o}$ conditionally to the state-sequence $\widehat{\mathbf{q}}$ and the sequence of context-dependent models $\boldsymbol{\lambda}$:

$$\widehat{\mathbf{o}} = \underset{\mathbf{o}}{\operatorname{argmax}}\ \mathrm{p}(\mathbf{o}|\widehat{\mathbf{q}}, \boldsymbol{\lambda}) \tag{9.16}$$

According to the conditional independence assumption:

$$\mathrm{p}(\mathbf{o}|\widehat{\mathbf{q}}, \boldsymbol{\lambda}) = \prod_{t=1}^{T} \mathrm{p}(\mathbf{o}_t|\widehat{\mathbf{q}}_t, \boldsymbol{\lambda}) \tag{9.17}$$

Then,

$$\max(\mathrm{p}(\mathbf{o}|\widehat{\mathbf{q}}, \boldsymbol{\lambda})) = \prod_{t=1}^{T} \max \mathrm{p}(\mathbf{o}_t|\widehat{\mathbf{q}}_t, \boldsymbol{\lambda}) \tag{9.18}$$

Thus, assuming that the observation probability density is a single normal distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the observation sequence $\mathbf{o}$ is given by the mean sequence of the corresponding state sequence $\widehat{\mathbf{q}}$ and the context-dependent models $\boldsymbol{\lambda}$:

$$\mathbf{o} = [\boldsymbol{\mu}_{q_1}, \ldots, \boldsymbol{\mu}_{q_T}] \tag{9.19}$$

9.3 Trajectory Modelling of Short and Long Term Variations

In the absence of any additional constraint, the optimal observation sequence simply corresponds to the mean sequence that is associated with the state sequence and the context-dependent models. Consequently, the optimal observation sequence would present discontinuities that would result in a degradation of the synthesized speech. Additionally, the determination of the optimal acoustic sequence does not benefit of a global maximization due to the conditional independence assumption. To alleviate this problem, the *Trajectory Model* [Tokuda et al., 2003] has been proposed in which the relationship between static and dynamic observations is explicitly formulated, and the sequence of dynamic observations is used as a *trajectory constraint* to maximize the conditional probability of the static observation sequence.

9.3.1 Trajectory Model

The principles of the *Trajectory Model* is here shortly reminded. In the *Trajectory Model*, an augmented observation sequence is defined as the concatenation of the static observation sequence and a set of dynamic observation sequences. The static observation sequence denotes the conventional observation sequence, and the dynamic observation sequence is composed of the k-th discrete time derivatives of the static observation sequence.

Let $\mathbf{c} = [\mathbf{c}_1, \ldots, \mathbf{c}_T]^\top$ be the static observation sequence, where $\mathbf{c}_t = [c_t(1), \ldots, c_t(D)]^\top$ be the (Dx1) static observation vector at time t.
Let $\mathbf{\Delta}^{(k)}\mathbf{c} = [\mathbf{\Delta}^{(k)}\mathbf{c}_1, \ldots, \mathbf{\Delta}^{(k)}\mathbf{c}_T]^\top$ be the k-th dynamic observation sequence, where $\mathbf{\Delta}^{(k)}\mathbf{c}_t = [\ \Delta^{(k)}c_t(1), \ldots, \Delta^{(k)}c_t(D)]^\top$ be the (Dx1) dynamic observation vector at time t.

The k-th dynamic vector $\mathbf{\Delta}^{(k)}\mathbf{c}_t$ is defined as k-th discrete time derivative of $\mathbf{c}$ at time t:

$$\mathbf{\Delta}^{(k)}\mathbf{c}_t = \sum_{\tau=-L}^{L} w^{(k)}(\tau)\mathbf{c}_t(t+\tau) \tag{9.20}$$

where w denotes the derivative window.

Then, the augmented observation sequence $\mathbf{o} = [\mathbf{o}_1^\top, \ldots, \mathbf{o}_T^\top]$ is defined as the concatenation of the static and dynamic observation sequences, where the observation vector $\mathbf{o}_t$ is formulated as a function of the observation vector $\mathbf{c}_t$ and its k-th discrete time derivatives vectors $\mathbf{\Delta}^{(k)}\mathbf{c}_t$:

$$\mathbf{o}_t = [\mathbf{c}_t^\top, \mathbf{\Delta}^{(1)}\mathbf{c}_t^\top, \ldots, \mathbf{\Delta}^{(K)}\mathbf{c}_t^\top]^\top \tag{9.21}$$

Since each of the k-th discrete derivatives consists in a finite linear combination of the static observation sequence, the augmented observation sequence can be factorized as follows:

$$\mathbf{o} = \mathbf{W}\mathbf{c} \tag{9.22}$$

where $\mathbf{W}$ is a (DTxKDT) sparse matrix composed of the derivative windows $w^{(k)}$.

The optimal observation sequence $\widehat{\mathbf{o}} = [\widehat{\mathbf{o}_1}^\top, \ldots, \widehat{\mathbf{o}_T}^\top]$ is determined so as to maximize the probability of the observation sequence $\mathbf{o} = [\mathbf{o}_1^\top, \ldots, \mathbf{o}_T^\top]$ conditionally to the model $\boldsymbol{\lambda}$ and the sequence length T:

$$\widehat{\mathbf{o}} = \underset{\mathbf{o}}{\operatorname{argmax}} \max_{\mathbf{q}} \mathrm{p}(\mathbf{o}|\mathbf{q}\boldsymbol{\lambda})\ \mathrm{p}(\mathbf{q}|\boldsymbol{\lambda}, T) \tag{9.23}$$

In the same manner as described previously, the determination of the optimal observation sequence $\mathbf{o}$ divides into the following sub-problems:

$$\widehat{\mathbf{q}} = \underset{\mathbf{q}}{\operatorname{argmax}}\ \mathrm{p}(\mathbf{q}|\lambda, T) \tag{9.24}$$

$$\widehat{\mathbf{o}} = \underset{\mathbf{o}}{\operatorname{argmax}}\ \mathrm{p}(\mathbf{o}|\widehat{\mathbf{q}}, \boldsymbol{\lambda}) \tag{9.25}$$

The optimal state sequence is determined according to the state duration model. The optimal augmented observation sequence is determine so as to maximize the conditional probability of the observation sequence $\mathbf{o}$ under the dynamic constraint $\mathbf{o} = \mathbf{W}\mathbf{c}$.

Under the dynamic constraint $\mathbf{o} = \mathbf{W}\mathbf{c}$, the maximization of $\mathrm{p}(\mathbf{o}|\widehat{\mathbf{q}}, \boldsymbol{\lambda})$ with respect to $\mathbf{o}$ is equivalent to that with respect to $\mathbf{c}$:

$$\widehat{\mathbf{o}} = \underset{\mathbf{o}}{\operatorname{argmax}}\ \mathrm{p}(\mathbf{o}|\mathbf{q}, \boldsymbol{\lambda}) \quad \Leftrightarrow \quad \widehat{\mathbf{c}} = \underset{\mathbf{c}}{\operatorname{argmax}}\ \mathrm{p}(\mathbf{W}\mathbf{c}|\mathbf{q}, \boldsymbol{\lambda}) \tag{9.26}$$

with $\widehat{\mathbf{o}} = \mathbf{W}\widehat{\mathbf{c}}$.

The maximization of $\mathrm{p}(\mathbf{W}\mathbf{c}|\widehat{\mathbf{q}}, \boldsymbol{\lambda})$ with respect to $\mathbf{c}$ is reached at the critical point $\widehat{\mathbf{c}}$:

$$\widehat{\mathbf{c}} \quad | \quad \frac{\partial \log \mathrm{p}(\mathbf{W}\widehat{\mathbf{c}}|\mathbf{q}, \boldsymbol{\lambda})}{\partial \mathbf{c}} = 0 \tag{9.27}$$

Assuming that each state probability density function is a single normal distribution $\mathcal{N}(\mu_q, \Sigma_q)$, the above equation can be reformulated as a set of linear equations which can be solved efficiently:

$$\mathbf{R}_q \widehat{\mathbf{c}} = \mathbf{r}_q \tag{9.28}$$

where:

$$\mathbf{R}_q = \mathbf{W}^\top \Sigma_q^{-1} \tag{9.29}$$

$$\mathbf{r}_q = \mathbf{W}^\top \Sigma_q^{-1} \mu_q \tag{9.30}$$

In this section, a formulation of the *Trajectory Model* based on the stylization of the f_0 variations over various temporal domains is presented. The syllable is assumed as the minimal temporal domain for the description of speech prosody, and f_0 variations are stylized and modelled simultaneously over different temporal domains: short-term variations correspond to the stylization of f_0 contours over the syllable, and long-term variations correspond to the stylization of f_0 contours over long-term temporal domains. During the training, a context-dependent HMM is estimated from the joint short-term and long-term variations. During the synthesis, the short-term variations are determined so as to maximize the conditional probability of the short-term variations under the constraint of the long-term trajectories.

The proposed method is evaluated in a subjective evaluation with different long-term trajectories, and compared to the conventional HMM-based speech synthesis.

9.3.2 Parameters Estimation

A context-dependent duration model is estimated according to syllable duration in the logarithmic domain. A context-dependent f_0 model is estimated in which the f_0 variations are stylized using the Discrete Cosine Transform of linearly interpolated f_0 in the logarithmic domain over syllable and a set of high-level units.

Let $\mathbf{q} = [\mathbf{q}_1, \ldots, \mathbf{q}_N]$ be the sequence of linguistic contexts, where $\mathbf{q}_n = [q_n(1), \ldots, q_n(L)]^\top$ is a (Lx1) linguistic vector which describes the linguistic characteristics associated with the n-th syllable.
Let $\mathbf{c} = [\mathbf{c}_1, \ldots, \mathbf{c}_N]$ be the static observation sequence of stylized f_0 contours over the syllable-level unit, where $\mathbf{c}_n = [c_n(1), \ldots, c_n(D)]^\top$ is a (Dx1) observation vector which describes the short-term f_0 characteristics associated with the n-th syllable.
Let $\boldsymbol{\Delta}^{(k)}\mathbf{c} = [\boldsymbol{\Delta}^{(k)}\mathbf{c}_1, \ldots, \boldsymbol{\Delta}^{(k)}\mathbf{c}_N]$ be the dynamic observation sequence of stylized f_0 contours over the k-th long-term temporal domain, where $\boldsymbol{\Delta}^{(k)}\mathbf{c}_n = [\ \Delta^{(k)}c_n(1), \ldots, \Delta^{(k)}c_n(D)]^\top$ is a (Dx1)

observation vector which describes the long-term f_0 characteristics associated with the n-th syllable.

Let $\mathbf{o} = [\mathbf{o}_1, \ldots, \mathbf{o}_N]$ be the augmented observation sequence, where $\mathbf{o}_n = [\mathbf{c}_n^\top, \mathbf{\Delta}^{(1)}\mathbf{c}_n^\top, \ldots, \mathbf{\Delta}^{(K)}\mathbf{c}_n^\top]^\top$ is a (KDx1) observation vector which describes the short-term and long term f_0 characteristics associated with the n-th syllable, and K the total number of long-term temporal domains being modelled.

The static observation sequence $\mathbf{c}$ denotes the short-term f_0 stylization over the syllable, and the dynamic observation sequence $\mathbf{\Delta}^{(k)}\mathbf{c}$ denotes the long-term variations that will be used as trajectory constraints.

A HMM $\boldsymbol{\lambda}_\mathbf{q}$ is estimated for each of the linguistic contexts. Each of the context-dependent HMMs is assumed to be a single-state HMM with single normal distribution and diagonal covariance matrix. Then, a context-dependent HMM $\boldsymbol{\lambda}$ is derived based on Maximum-Likelihood Minimum-Description-Length (ML-MDL). The long-term variations are used as additional trajectory constraints to refine the clustering of the models. A conventional context-dependent HMM is used to model syllable durations.

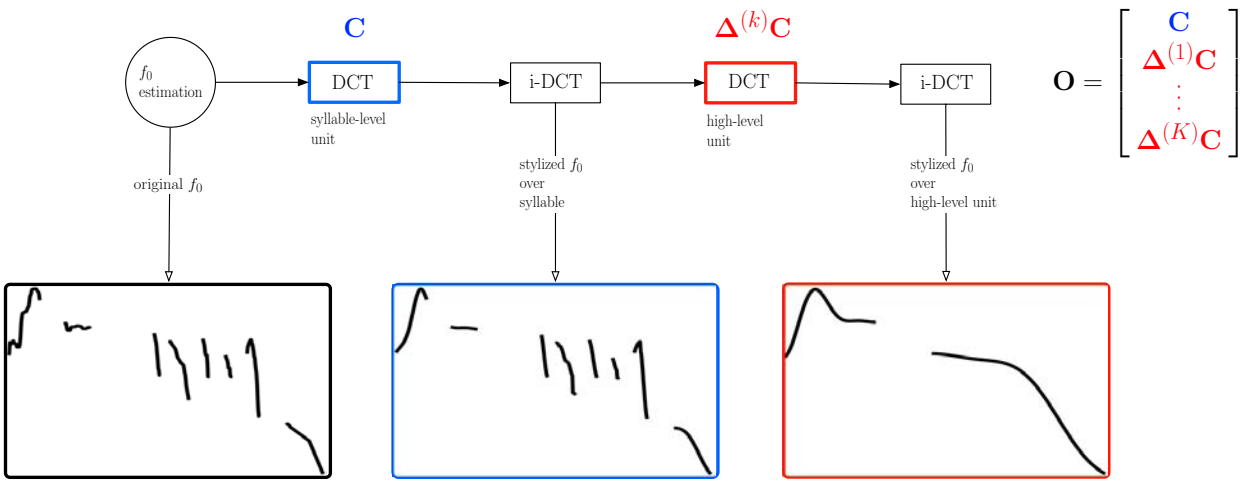


Figure 9.4: Computation of the long-term trajectories. The augmented sequence $\mathbf{o}$ is composed of the static sequence $\mathbf{c}$ and the dynamic sequences $\mathbf{\Delta}^{(k)}\mathbf{c}$ are directly computed from the static sequence.

9.3.3 Parameters Inference

The inference of the sequence of f_0 parameters is similar to that described in the *Trajectory Model* with the exception that the frame-based static observation is reformulated into the stylized f_0 contour over the syllable, and the frame-based dynamic observation (partial derivative) is reformulated into the stylized long-term f_0 contours. The sequence of syllable durations is determined with the conventional static method as the sequence of mean durations.

The optimal static observation sequence $\mathbf{c}$ is determined so as to maximize the log-likelihood of the short-term observation sequence $\mathbf{o}$, under the constraint of the long-term trajectories $\mathbf{\Delta}^{(\mathbf{k})}\mathbf{c}$.

9.3.3.1 Maximization of Joint-Likelihood

The optimal observation sequence $\widehat{\mathbf{o}} = [\widehat{\mathbf{o}_1}^\top, \ldots, \widehat{\mathbf{o}_T}^\top]$ is determined so as to maximize the probability of the observation sequence $\mathbf{o}$ conditionally to the model $\boldsymbol{\lambda}$.

$$\widehat{\mathbf{o}} = \underset{\mathbf{o}}{\operatorname{argmax}} \max_{\mathbf{q}} \mathrm{p}(\mathbf{o}|\mathbf{q}, \boldsymbol{\lambda})\ \mathrm{p}(\mathbf{q}|\boldsymbol{\lambda}) \tag{9.31}$$

The determination of the optimal observation sequence $\mathbf{o}$ divides into the following sub-problems:

$$\widehat{\mathbf{q}} = \underset{\mathbf{q}}{\operatorname{argmax}}\ \mathrm{p}(\mathbf{q}|\boldsymbol{\lambda}) \tag{9.32}$$

$$\widehat{\mathbf{o}} = \underset{\mathbf{o}}{\operatorname{argmax}}\ \mathrm{p}(\mathbf{o}|\widehat{\mathbf{q}}, \boldsymbol{\lambda}) \tag{9.33}$$

Assuming that each syllable is modelled by a single-state HMM, the optimal state sequence $\widehat{\mathbf{q}}$ simply corresponds to the concatenated sequence of context-dependent models associated with each syllable of the syllable sequence:

$$\widehat{\mathbf{q}} = [\mathbf{q}_1, \ldots, \mathbf{q}_N] \tag{9.34}$$

where N denotes is the total number of syllables in the syllable sequence, and $\mathbf{q}_n$ the state which corresponds to the context-dependent model associated with the n-th syllable.

The maximization of $\mathrm{p}(\mathbf{o}|\widehat{\mathbf{q}}, \boldsymbol{\lambda})$ with respect to $\mathbf{o}$ is equivalent to the maximization of $\mathrm{p}(\mathbf{c}|\widehat{\mathbf{q}}, \boldsymbol{\lambda})$ with respect to $\mathbf{c}$ under the dynamic constraints $\boldsymbol{\Delta}^{(k)}\mathbf{c}$:

$$\widehat{\mathbf{o}} = \underset{\mathbf{o}}{\operatorname{argmax}}\ \mathrm{p}(\mathbf{o}|\widehat{\mathbf{q}}, \boldsymbol{\lambda}) \ \Leftrightarrow\ \widehat{\mathbf{c}} = \underset{\mathbf{c}}{\operatorname{argmax}}\ \mathrm{p}(\mathcal{F}(\mathbf{c})|\widehat{\mathbf{q}}, \boldsymbol{\lambda}) \tag{9.35}$$

under the constraint:

$$\mathbf{o} = \mathcal{F}(\mathbf{c}) = \left[\mathbf{c}^\top, \boldsymbol{\Delta}^{(\mathbf{k})}\mathbf{c}^\top, \ldots, \boldsymbol{\Delta}^{(\mathbf{K})}\mathbf{c}^\top\right]^\top \tag{9.36}$$

A local solution to this problem is determined recursively using a quasi-Newton method.

9.3.3.2 Local Optimization Using a Quasi-Newton Method

Due to the stylization processing of the f_0 contours, the relationship between short and long term trajectories can not be simply formulated, thus there is no close-form solution to this problem. Additionally, there is no known method to determine the short-term observation sequence which globally maximizes the conditional probability of the observation sequence under the long-term constraints. Consequently, a recursive estimation of the locally optimal observation sequence is achieved using a quasi-Newton method.

In this study, both gradient and Hessian of the objective function are assumed to remain analytically unknown. Thus, a quasi-Newton method which assumes numerical calculation of the gradient and the Hessian is used to estimate the locally optimal static observation sequence under the dynamic constraints. Actually, the stylization of f_0 contours considerably reduces the complexity of the optimization compared to the conventional HMM, thus reasonably support the use of a quasi-Newton optimization without an explicit formulation of the gradient. Nevertheless, some approximations [Latorre and Akamine, 2008, Qian et al., 2009] on the parametrization of the long-term domains f_0 contours would improve the determination of the optimal observation sequence, and reduce computational cost.

For this purpose, an objective function $\mathcal{O}$ is defined as the log-likelihood of the static observation sequence given the state sequence and the model under the dynamic constraints:

$$\mathcal{O} = \mathrm{p}(\mathcal{F}(\mathbf{c})|\mathbf{q}, \boldsymbol{\lambda}) \tag{9.37}$$

A quasi-Newton Method is a numerical method used to solve non-linear equation systems, which is used in particular to estimate local maxima (respectively minima) of an objective function f for which the analytical expression is unknown.

$$\widehat{\mathbf{x}} \quad | \quad f'(\widehat{\mathbf{x}}) = 0 \tag{9.38}$$

Quasi-Newton methods iteratively determine the stationary point of an objective function f based on local quadratic approximation and Newton's method. In particular, Newton's method assumes that the objective function f can be locally approximated as a quadratic in the region around the optimum, and use the first and second derivatives (gradient and Hessian) to determine the stationary point. In the used quasi-Newton method, gradient and Hessian of the function f are approximated by finite differences and the *Broyden-Fletcher-Goldfarb-Shanno* BFGS ([Broyden, 1970]) formula, respectively.

Let $\mathbf{x}_0$ be the initialization of $\mathbf{x}$, $\mathbf{x}_i$ the estimate of the solution $\widehat{\mathbf{x}}$ at iteration i, and $\mathbf{s}_i = \mathbf{x}_{i+1} - \mathbf{x}_i$ a direction at iteration i. Let g_i be the gradient of the function f at point $\mathbf{x}_i$, and B_i the approximation to the Hessian Matrix H_i of the function f at point $\mathbf{x}_i$ and along the direction $\mathbf{s}_i$.

The function f is locally approximated at point $\mathbf{x}_i$ by its second-order Taylor approximation:

$$f(\mathbf{x}) \quad = f(\mathbf{x}_i) + (\mathbf{x} - \mathbf{x_i}) f^{(1)}(\mathbf{x}_i) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_i) f^{(2)}(\mathbf{x}_i)(\mathbf{x} - \mathbf{x}_i) \tag{9.39}$$

Thus, the estimate $\mathbf{x}_{i+1}$ of the stationary point can be formulated according to the modified Newton method:

$$\mathbf{x}_{i+1} = \mathbf{x}_i - \alpha_i \mathbf{p}_i \tag{9.40}$$

where $\mathbf{p}_i$ is the Newton direction and α_i ($\alpha_i \in [0, 1]$) is the Newton step-length used as an additional conditioning parameter.

The Newton direction $\mathbf{p}_i$ is directly derived from the Newton method:

$$B_i \mathbf{p}_i = -g_i \tag{9.41}$$

under the quasi-Newton condition:

$$B_{i+1} \mathbf{p}_i = g_{i+1} - g_i \tag{9.42}$$

Thus, the estimation can efficiently be solved from the estimation of the matrix B_i at $\mathbf{x}_i$ along direction $\mathbf{s}_i$[Broyden, 1970]:

$$B_{i+1} = B_i + \frac{g_i g_i^\top}{g_i^\top \mathbf{s}_i} - \frac{(B_i \mathbf{s}_i)(B_i \mathbf{s}_i)^\top}{\mathbf{s}_i^\top B_i \mathbf{s}_i} \tag{9.43}$$

Then, the step-length is determined by a local line-search optimization of the function f at point $\mathbf{x}_i$, given the Newton direction $\mathbf{p}_i$:

$$\widehat{\alpha_i} = \underset{\alpha_i}{\operatorname{argmin}} \, f(\mathbf{x}_i - \alpha_i \mathbf{p}_i) \tag{9.44}$$

The quasi-Newton method is proved to locally converge to a stationary point $\widehat{\mathbf{x}}$ of the objective function f. Practically, B_0 is generally initialized with the identity matrix $B_0 = \mathbb{1}$, so that the first step is equivalent to a gradient descent, but further steps are more and more refined by the approximation to the Hessian.

In the present study, the augmented observation sequence is initialized to the mean observation sequence which corresponds to the optimal sequence in the absence of dynamic constraints, and the Hessian is initialized to the identity matrix. Assuming prosodic independence across successive major prosodic groups, the local optimization is achieved on each of the major prosodic groups independently.

9.3.4 Parameters Inference Using *Global Variance* (GV)

The *Global Variance* (GV) method was introduced to alleviate the poor dynamic of the synthesized acoustic parameters. The principle of the Global Variance method is to estimate the global variance of a sequence of acoustic parameters over a time sequence (e.g., an utterance or a phoneme), and to use the estimated global variance as an additional constraint during the inference of the sequence of acoustic parameters. Various methods have been proposed to account for the global variance during the training and the synthesis. In particular, methods have been proposed to introduce the global variance in the estimation and the inference of the parameters of the trajectory model [Toda and Tokuda, 2007, Toda and Young, 2009].

In the present study, the global variance is separately estimated during the training, and not used during the context-clustering. During the synthesis, the global variance is simply used as an additional constraint for the inference of the sequence of prosodic parameters. Thus, the sequence of static observation is determined so as to maximize the conditional probability of the observation sequence $\mathbf{o}$ under the constraints of the dynamic sequence $\mathbf{\Delta^{(k)}c}$ and the global variance.

9.4 Evaluations

9.4.1 Evaluation of the Rich Linguistic Context

In this section, the role of linguistic context in speech prosody modelling is assessed. Linguistic information was extracted from text using the linguistic processing chain described in chapter 7 that includes surface and deep syntactic parsing. Evaluation was achieved using the conventional HMM-based speech synthesis system [Zen et al., 2009] trained with respect to different sets of linguistic contexts. Two sets of linguistic contexts were compared: a *baseline* set of linguistic contexts which is composed of the morpho-syntactic context solely; and a *rich* set of linguistic contexts which includes all of the linguistic contexts, in particular the deep syntactic contexts. The evaluation consisted in a subjective comparison of the models trained with the different linguistic contexts.

9.4.1.1 Stimuli

Linguistic Contexts

Linguistic information were extracted from text using the linguistic processing chain described in chapter 7. For the purpose of this experiment, two models were compared: a *baseline model* including segmental, prosodic, and morpho-syntactic features; and a *rich linguistic model* including segmental, prosodic, and all of the syntactic features. The used linguistic units are phoneme, syllable, and the syntactic units. Linguistic features are converted into linguistic contexts at the phoneme level by computing locational and weight contexts, and representing 1-order left-to-right contexts and 1-order child-to-parent contexts in the case of the dependency contexts.

The baseline and the rich contexts were defined as:

$$\begin{array}{lrcl} \text{baseline:} & Q^{(phone)}_{\text{baseline}} & = & Q_{\text{segment}} \cup Q_{\text{proso}} \cup Q_{\text{morpho}} \\ \text{rich:} & Q^{(phone)}_{\text{rich}} & = & Q_{\text{segment}} \cup Q_{\text{proso}} \cup Q_{\text{morpho}} \cup Q_{\text{dep}} \cup Q_{\text{chunk}} \cup Q_{\text{adj}} \end{array}$$

Training Corpus

Models were trained on 1 hour (956 utterances) of the *laboratory* corpus.

Evaluation Corpus

The precise and exhaustive evaluation of the relative influence of each of the extracted syntactic features on the synthesized speech prosody is unreachable, due to the high complexity of the

syntactic structure as well as their dependencies with speech prosody. More reasonably, one can evaluate the change in speech prosody with a limited set of well-defined and high-level syntactic constructions that cover large syntactic units. For this reason, a set of sentences was specifically designed to evaluate the role of syntactic adjunctions, only. This choice was motivated by several reasons:

1 adjunctions have been proved to significantly relate to speech prosody in chapter 8, and may be associated with specific prosodic patterns (e.g., incises, relative clauses or more generally *oral parenthesis*).

2 some specific adjunctions concern long-term temporal domains, and global changes in speech prosody are more easily perceived and evaluated than local details.

3 a limited vocabulary suffices for a reasonable description of various adjunctions. This is desired to design a limited set of representative and controlled sentences for the evaluation.

The text corpus used for the evaluation was designed in the following manner: 10 baseline sentences were chosen with a direct and minimal syntactic structure (for instance: *Le chat a mangé la souris*, *The cat ate the mouse*). These sentences were then enriched with various types of adjunctions (subordinate participial clauses, subordinate relative clauses, coordinate clauses, incises and enumerations). For each type of adjunction, the sentences were enriched according to two control parameters: position (initial, medial, final) and *complexity* (presence or not of adjunctions within the current adjunction) of the introduced adjunction. This finally results into an evaluation corpus composed of 54 sentences. A description of the text corpus construction is provided in table 9.2.

Finally, the evaluation corpus is composed of a subset of 20 sentences randomly extracted from the designed text corpus. Sentences were processed by the linguistic processing chain, without manual correction. The sequence of prosodic events was determined using the context-dependent discrete HMM described in chapter 8.2 with the full linguistic context, and shared among the models to be compared.

9.4.1.2 Participants

50 French native speakers (including 17 expert and 33 naïve listeners) participated in the evaluation. Meta-information were gleaned from the participants: speech expertise (expert, naïve), language (native French speaker, non-native French speaker), age, and listening condition (headphones or not). Participants were encouraged to use headphones.

9.4.1.3 Procedure

The evaluation consisted in a subjective comparison of the 2 models. A comparison category rating (CCR[International Telecommunication Union, 1996]) test was used to compare the *prosodic naturalness* of the speech utterances synthesized by the *baseline*-contexts and the *rich*-contexts models[1]. The evaluation was conducted according to a *crowd-sourcing* technique using social networks[2].
Participants compared a total of 20 pairs of speech utterances. Pairs of synthesized speech utterances were randomly presented to the participants. Participants were asked to compare the *prosodic naturalness* of each pair of synthesized speech utterances. They were asked to attribute a preference score according to the *prosodic naturalness* of the speech utterances being compared on the comparison mean opinion score (CMOS) scale (table 9.1).

Prosodic naturalness was referred as:

[1] the experiment is available at the following link: `http://recherche.ircam.fr/equipes/analyse-synthese/lanchant/index.php/Main/TestSP`

[2] *Ircam Analysis and Synthesis Perceptual Experiments* on Facebook: `http://www.facebook.com/group.php?gid=150354679034&ref=ts`

- a "correct" prosody: the utterance is pronounced as it could be expected from a native speaker.
- a "lively" prosody. The opposite of a lively prosody is a monotone prosody.

Participants were additionally asked to ignore speech synthesis artefacts.

Score	**Difference**
(+/-) 3	much better
(+/-) 2	better
(+/-) 1	slightly better
0	about the same

Table 9.1: Comparative MOS scale

sentence	Je me suis couché de bonne heure. *I got to bed early.*
enrichments	
subordinate	*Comme la nuit tombait*, je me suis couché de bonne heure. Je me suis couché, *Maman ayant fermé la porte et souffle´ma bougie*, de bonne heure. Je me suis couché de bonne heure, *songeant longtemps encore au charme d'Albertine*.
coordinate	Je me suis couché de bonne heure, *car le sommeil m'accablait*. Je me suis couché de bonne heure, *et je regardais les miroitements de ma lanterne magique*.
incise	*Longtemps*, je me suis couché de bonne heure. Je me suis couché, *à mon grand désespoir sans le baiser de Maman*, de bonne heure. Je me suis couché de bonne heure, *hélas*.
enumeration	*Marcel, Swann, et Madame de Guermantes*, se sont couchés de bonne heure. Je me suis couché de bonne heure *dans la chambre de mon enfance, dans cette autre, ou bien dans cette autre encore*.

Table 9.2: Description of sentence enrichment with different types of adjunction.

stream	source/filter	duration	f_0
corpus			
training corpus	laboratory corpus (1h)		
evaluation corpus	C-SYNTAX (24 sentences)		
feature extraction			
feature	5-order aperiodicity 39-order MFCC	state-duration	f_0
window	50-ms blackmann		
frame rate	5ms		
feature transform			
transform	-	log	log
dynamic	1-orderΔ ,Δ^2	-	1-orderΔ ,Δ^2
model			
topology	5-state HMM normal distribution semi-tied covariance	5-state HMM normal distribution	5-state MSD-HMM normal distribution semi-tied covariance
context	**M1**: baseline linguistic context, $Q^{(phone)}_{\text{baseline}} = Q_{\text{morpho}} \cup Q_{\text{proso}}$ **M2**: rich linguistic context, $Q^{(phone)}_{\text{rich}} = Q_{\text{adj}} \cup Q_{\text{proso}}$		
clustering	DT ML-MDL		

Table 9.3: Evaluation of the *Rich Linguistic Context*: model setup

9.4.1.4 Results

Overall CMOS and preference score (PS) are presented in figure 9.5. CMOS and PS with respect to the adjunction type are presented in table 9.4 and figure 9.6.

The rich-context model is overally significantly preferred to the baseline-context model (CMOS=+0.31, PS=31/49%). However, there are a significant differences depending on the the type of adjunction being modelled. The baseline utterances presents no difference between the baseline and rich model (CMOS=+0.19, PS=31/43%). Scores obtained for the enriched utterances reveals the strongest preference for the rich-context model, this result being however not systematic for all types of adjunction. The rich-context model is significantly preferred in the case of participial (CMOS=+0.64, PS=31/62%), coordinate (CMOS=+0.48, PS=25/52%), incise (CMOS=+0.40, PS=28/51%), and enumeration (CMOS=+1.19, PS=12/83%) adjunctions, The baseline-context model is not significantly preferred in the case of relative (CMOS=-0.25, PS=32/46%) adjunctions.

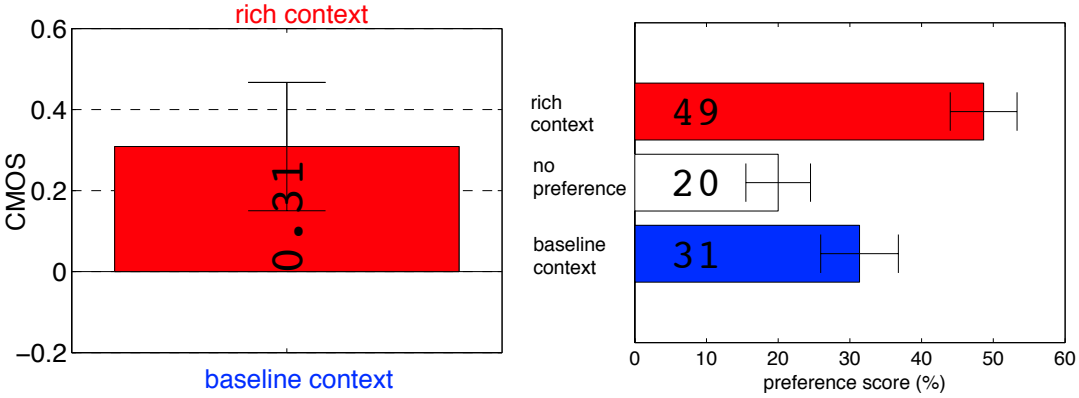


Figure 9.5: Overall CMOS and PS. Mean scores and 95% intervals.

adjunction type	CMOS		PS (%)					
	rich		baseline		rich		no preference	
no	+0.19	(± 0.16)	30.8	(± 9.2)	43.5	(± 8.6)	25.8	(± 10.6)
participial	+0.64	(± 0.17)	30.8	(± 7.8)	61.5	(± 12.0)	7.7	(± 7.6)
relative	-0.25	(± 0.15)	46.2	(± 9.0)	32.2	(± 9.7)	21.6	(± 5.9)
coordinate	+0.48	(± 0.10)	25.0	(± 4.6)	51.9	(± 6.0)	23.1	(± 7.4)
incise	+0.40	(± 0.12)	28.2	(± 8.8)	50.6	(± 8.9)	21.2	(± 0.0)
enumeration	+1.19	(± 0.10)	11.5	(± 5.2)	82.7	(± 5.6)	5.8	(± 5.1)

Table 9.4: Comparison of CMOS and PS with respect to type of adjunction. Mean scores and 95% intervals

9.4.1.5 Discussion

The baseline sentences were used as control sentences in the evaluation, in particular with a simple syntactic structure. Consequently, it was expected that no significant difference would be observed. The significant preference for the rich-context model is evidence that the rich-context model succeeds in modelling prosodic variations associated with specific syntactic constructions. However, the improvement depends on the type of adjunction being modelled. Such differences may result from two main causes: adjunctions are more or less associated with specific prosodic patterns; adjunctions with rare occurrence are poorly modelled.

In order to interpret the differences in preference, the synthesized prosodic variations were analysed. There were no difference with respect to the state-duration modelling. The model clearly failed in

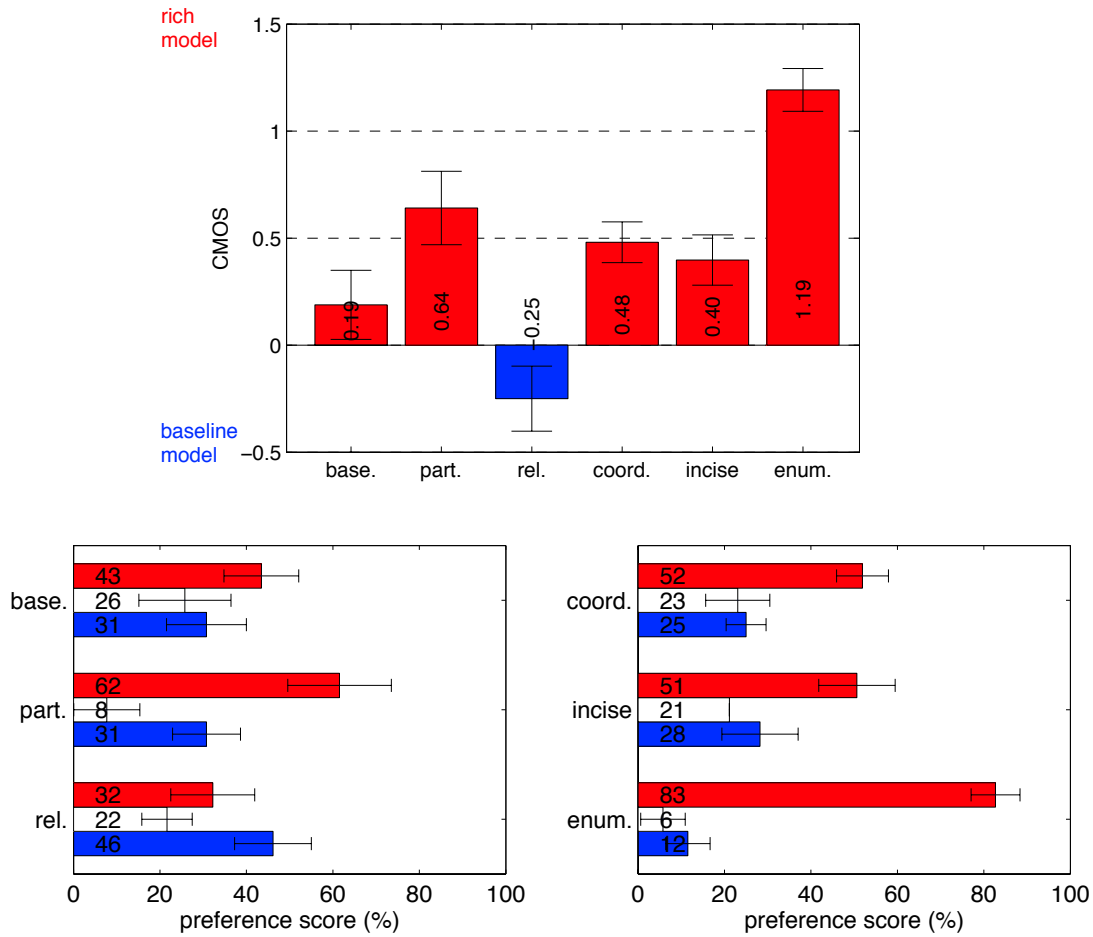


Figure 9.6: Overall CMOS and PS with respect to the type of adjunction type. Mean scores and 95% intervals

modelling the change in speech rate associated with specific syntactic constructions. A comparison of f_0 variations with respect to the linguistic-context is provided in figure 9.7 in the case of two the the evaluated utterances, with state-duration alignment.
The rich-context provides local prosodic changes that can be listed as follows:

prosodic prominence: modification of the prosodic contour associated with a prosodic prominence (e.g., conclusive or continuative contour); prominence dynamic;

prosodic phrasing: local change in prosodic phrasing, change in the sequence of prosodic contours. However, no global change, such as change in register, was observed.

Finally, the rich-context model mostly affects local prosodic variations, but failed into modelling global prosodic changes.

9.4.1.6 Conclusion

In this section, the role of linguistic context in speech prosody modelling was assessed. Linguistic information was extracted from text using the linguistic processing chain that includes surface and deep syntactic parsing. Evaluation was achieved using the HMM-based speech synthesis system with respect to different linguistic contexts. Two linguistic feature sets were compared: a *baseline* feature set which is composed of prosodic and morpho-syntactic features, solely; and a *rich* feature set which includes prosodic features and all of the extracted syntactic features, and in particular deep syntactic features. The evaluation consisted in a subjective comparison of the

models trained with the different linguistic contexts.

The rich syntactic description was proved to successfully refine the modelling of speech prosody variations, depending on the syntactic construction being modelled. The changes concern local prosodic variations only, and fails to model global variations of speech prosody. In particular, the rich-context model failed to model speech rate and register changes with respect to specific syntactic constructions. This may be due to the short-term modelling of speech prosody that is inherent in the conventional HMM-based speech synthesis system. In order to model local and global variations of speech prosody adequately, long-term variations have to be explicitly described and modelled. The formulation of a trajectory model based on the stylization and the modelling of f_0 variations simultaneously over various temporal domains is an attempt to model long-term variations of speech prosody, and will be evaluated in the following section.

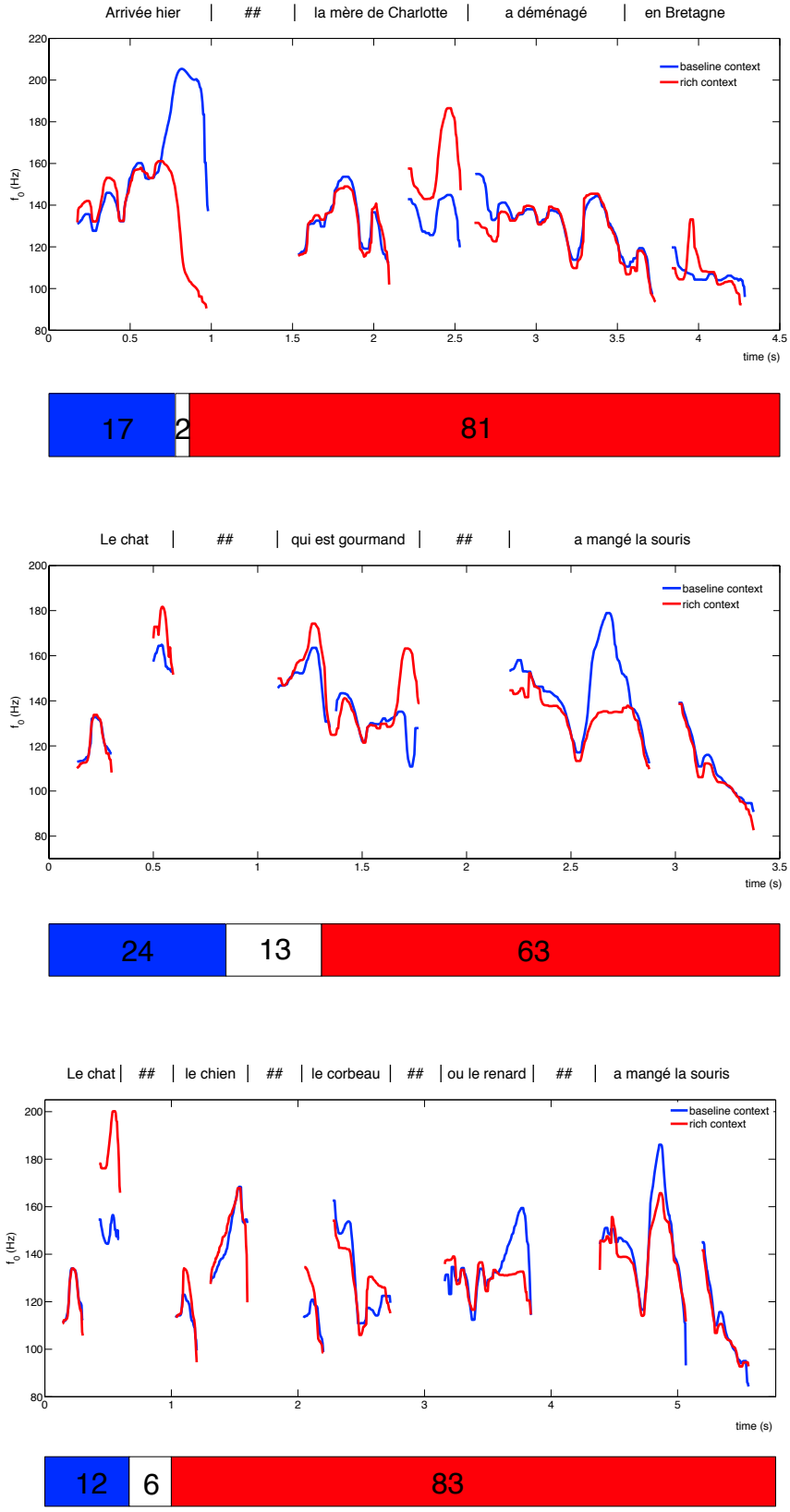


Figure 9.7: Comparison of synthesized f_0 variations depending on the linguistic context, with the associated preference scores.

9.4.2 Evaluation of the *Trajectory Model*

In this section, the trajectory model is evaluated using different long-term trajectories and compared to the conventional HMM-based model. The evaluation consisted in a subjective comparison of different speech prosody models in speech synthesis. Four speech prosody models were compared: trajectory models using various long-term temporal domains (1-order syllable context, minor prosodic group (AG), major prosodic group (PG)[3]), and the conventional HMM-based model. Evaluation was conducted using the HMM-based speech synthesis system [Zen et al., 2009].

9.4.2.1 Stimuli

Linguistic Contexts

Linguistic information were extracted from text using the linguistic processing chain described in chapter 7. For the purpose of this experiment, models were trained with the baseline feature set, which consists in segmental, prosodic, and morpho-syntactic features solely. The used linguistic units were (phoneme), syllable, and the syntactic units. Linguistic features were converted into linguistic contexts at the (phoneme), syllable level by computing locational and weight contexts, and representing 1-order left-to-right contexts and 1-order child-to-parent contexts in the case of the dependency contexts.

The baseline context was defined as:

$$\text{baseline:} \quad Q_{\text{baseline}}^{(phone,syllable)} \quad = \quad Q_{\text{segment}} \cup Q_{\text{proso}} \cup Q_{\text{morpho}}$$

Training Corpus

Speech synthesis model and speech prosody models were trained on 5 hours (1888 utterances) of the *multi-media* corpus.

Evaluation Corpus

The evaluation text corpus is composed of 8 sentences randomly extracted from the C-TALE text corpus (143 sentences). The C-TALE corpus is the fairy-tale "*Le Petit Poucet*" ("*Little Tom Thumb*") by French writer Charles Perrault [Perrault, 1697]. The sentences were processed by the linguistic processing chain, without manual correction. The sequence of prosodic events was determined using the context-dependent discrete HMM described in chapter 8.2 with the full linguistic context, and shared among the models to be compared.

Speech Prosody Models

Different models were compared: the conventional HMM-based model, and three trajectory models with different temporal domains. The conventional HMM-based model was trained at the phoneme level, and the trajectory models were trained at the syllable level.

1. HTS;
2. syllable + 1-order syllable-context unit;
3. syllable + minor prosodic group (AG) unit;
4. syllable + major prosodic group (PG) unit.

For each of the trajectory models, the inferred sequences of prosodic parameters (syllable duration and f_0 variations) were integrated into the HMM-based speech synthesis system [Zen et al., 2009] in

[3] AG stands for accentual group (minor prosodic group), and PG for prosodic group (major prosodic group)

the following manner: First, the inferred syllable duration sequence was used to modify the state-duration sequence as determined by the conventional HMM-based state-duration model: phone and sub-phone durations were homogeneously realigned according to the inferred syllable duration. Then, the inferred sequence of stylized f_0 parameters was converted into a sequence of f_0 variations with respect to the inferred syllable durations and the voice/unvoiced sequence as determined from the conventional HMM-based f_0 model. Finally, speech utterances were synthesized by the speech synthesizer. Each sentence was synthesized with the different models. This result into 8x4 = 32 synthesized utterances, and 8x6 = 48 pairs of speech utterances to be compared.

stream	duration	f_0
corpus		
training corpus	multi-media corpus (7h)	
evaluation corpus	fairy-tale text corpus (8 sentences)	
feature extraction		
feature	duration	f_0
window	syllable	50-ms hanning
frame rate	syllable	5ms
feature transform		
transform function	log	log + 5-order DCT
transform unit	syllable	**M1**: syllable; **M2.1**: Δ = 1-order context **M2.2**: Δ = AG **M2.3**: Δ = PG
model		
topology	single state HMM normal distribution diagonal covariance	
context	baseline linguistic context, $Q^{(phone,syllable)}$	
clustering	DT ML-MDL	

Table 9.5: Evaluation of the *Joint Trajectory Model*: model setup

9.4.2.2 Participants

20 native French speakers (including 13 expert and 7 naïve listeners) participated in the evaluation. Meta-information were gleaned from the participants: speech expertise (expert, naïve), language (native French speaker, non-native French speaker), age, and listening condition (headphones or not). Participants were encouraged to use headphones.

9.4.2.3 Procedure

The experiment consisted in a subjective comparison of the different models of speech prosody.

A comparison category rating test (CCR[International Telecommunication Union, 1996]) was used to compare the *naturalness* of the synthesized speech utterances[4]. The evaluation was conducted according to a *crowd-sourcing* technique using social networks[5].
Pairs of synthesized speech utterances were randomly presented to the participants. They were asked to attribute a preference score according to the *naturalness* of the speech utterances being compared on the comparison mean opinion score (CMOS) scale.

[4]the experiment is available at the following link: `http://recherche.ircam.fr/equipes/analyse-synthese/obin/pmwiki/pmwiki.php/Main/HTSProsoModel`

[5]*Ircam Analysis and Synthesis Perceptual Experiments* on Facebook: `http://www.facebook.com/group.php?gid=150354679034&ref=ts`

9.4.2.4 Results

Overall CMOS and preference score (PS) are presented in figure 9.8. Pair CMOS and preference rate are presented in table 9.6 and figure 9.9.

The 1-order trajectory model significantly outperforms all of the other prosodic models regardless to the preference measure. In particular, the 1-order trajectory model is overally significantly preferred to the other prosodic models (CMOS=0.53, PS=30%), and is individually significantly preferred to each of the other prosodic models (MOS=+0.54,0.51,054 and PS=52.1%,56.3%,55.1% compared with HTS, AG, and PG models respectively). The AG trajectory model is preferred to the HTS model but not significantly (overall: CMOS=-0.18, PS=22%; pair: CMOS=+0.15, PS=46%); and significantly preferred to the PG trajectory model. Finally, the HTS model is preferred to the PG trajectory model, but not significantly (overall: CMOS=-0.34, PS=18%; pair: CMOS=+0.10, PS=28.7%). In particular, trajectory models decrease in preference when increasing the temporal domain of the trajectory constraint (1-order:CMOS=0.53,PS=30%; AG: CMOS=-0.18, PS=22%; PG: CMOS=-0.38, PS=17%).

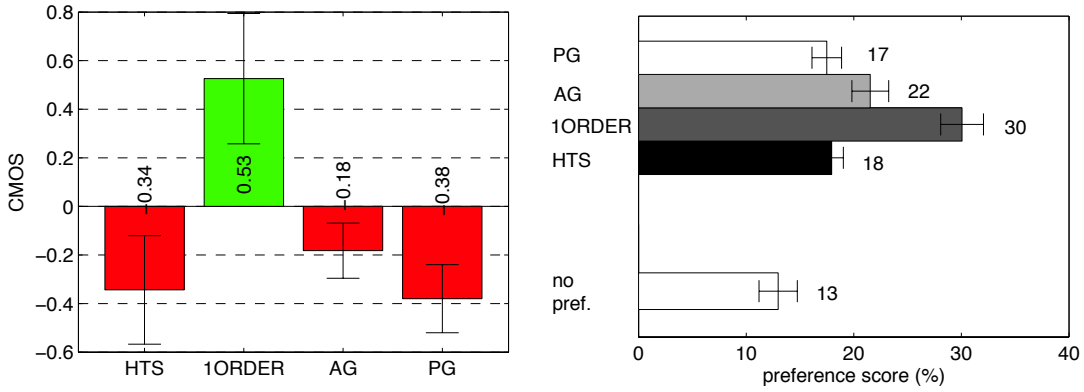


Figure 9.8: Overall CMOS and PS. Mean scores and 95% confidence intervals

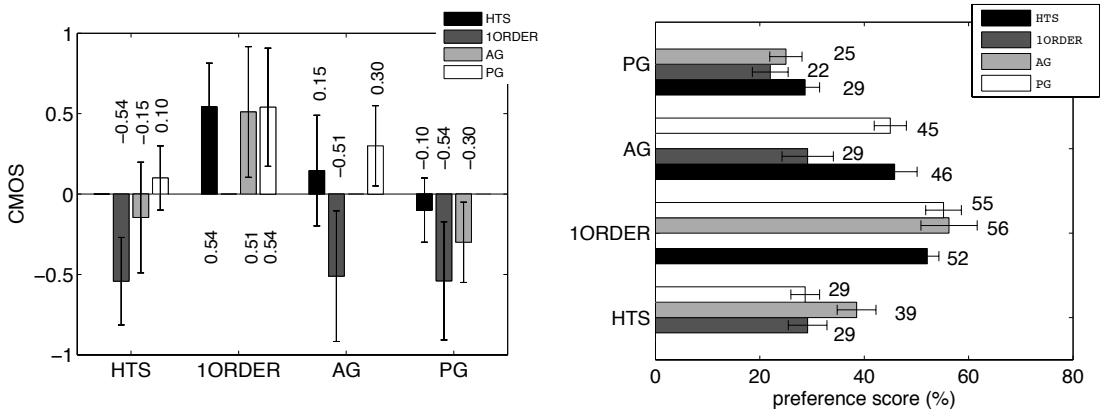


Figure 9.9: Pair Comparison of CMOS and PS. Mean scores and 95% intervals

A comparison of the preference scores depending on the expertise of the participant reveals a significant difference in the perception of speech prosody between naïve and expert listeners : naïve listeners have clearly marked preferences, but with more variability, while expert listeners have less marked preferences, but with less variability (table 9.7). Naïve listeners have a strong preference for the 1-order trajectory model (CMOS=+0.88, PS=37.4%), and a general preference for the trajectory models compared with the HTS model. Expert listeners have a significant preference for the 1-order trajectory model but in a less pronounced manner than naïve listeners (CMOS=+0.41,

CMOS	HTS	1-order	AG	PG
HTS	-	-0.54	-0.15	+0.10
1-order	**+0.54**	-	**+0.51**	**+0.54**
AG	+0.15	-0.51	-	+0.30
PG	-0.10	-0.54	-0.30	-

preference (%)	HTS	1-order	AG	PG
HTS	-	29.1	38.6	28.7
1-order	**52.1**	-	**56.3**	**55.2**
AG	45.9	29.2	-	45.1
PG	28.7	22.2	24.9	-

Table 9.6: CMOS and PS matrices depending on the model.

PS=28.1%), and have a significant preference for the HTS model compared to the longer-units trajectory models. In particular, the PG trajectory model is significantly rejected (CMOS=-0.52, PS=16.7%), and comparable with the indecision (PS=13.5%). In both cases, trajectory models decrease in preference when increasing the temporal domain of the trajectory constraint (1-order:CMOS=+0.88,+0.41,PS=37.4%, 28.1%; AG: CMOS=-0.10, -0.21, PS=20.7%,20.8%, ; PG: CMOS=-0.20, -0.52, PS=11.4%, 16.7% for the naïve and expert listeners respectively).

preference	naive			expert		
(%)	score		rank	score		rank
HTS	16.0	(± 2.7)	3	20.9	(± 3.8)	2
1-order	**37.4**	(± 7.2)	**1**	**28.1**	(± 6.7)	**1**
AG	20.7	(± 4.7)	2	20.8	(± 4.7)	3
PG	11.4	(± 2.8)	4	16.7	(± 3.5)	4
no preference	14.5	(± 14)	-	13.5	(± 7.4)	-

CMOS	naive			expert		
	score		rank	score		rank
HTS	-0.77	(± 0.44)	4	-0.20	(± 0.27)	2
1-order	**+0.88**	(± 0.43)	**1**	**+0.41**	(± 0.26)	**1**
AG	-0.10	(± 0.50)	2	-0.21	(± 0.28)	3
PG	-0.20	(± 0.44)	3	-0.52	(± 0.24)	4

Table 9.7: Comparison of CMOS and PS with respect to the expertise of the participant. Mean scores and 95% intervals.

9.4.2.5 Discussion

A comparison of duration modelling reveals no significant differences between state-based and syllable-based modelling, with the exception to slight improvements of local speech rate and fluency for the later. In order to interpret the differences in speech prosody, study cases of synthesized f_0 variations with respect to the speech prosody model are provided in figure 9.10 with prior state duration alignment. Speech prosody differences mostly concern $f0$ variations.

The 1-order trajectory model clearly succeeds to model the local variations and dynamic of speech prosody. Compared to the HTS model, the synthesized f_0 variations appear more flat than those synthesized by the HTS model when considering the micro-prosodic details, but more pronounced on prosodic prominences. Thus, naïve listeners may focus on global variations only, when expert listeners may pay a closer attention to finer prosodic details. The AG trajectory model appears to model middle-term prosodic variations such as initial f_0 reset and local f_0 declination, compared with the 1-order trajectory model and the HTS model. However, prosodic prominences are less

pronounced, and prosodic phrasing is more flat.

A comparison of the different trajectory models reveals that differences in speech prosody concern local (syllable f_0 variations, prominence form) and global f_0 variations. However, it is observed that the increase of the trajectory domain results into noisy local f_0 variations, and partially (AG) or totally (PG) inadequate global f_0 contours. In particular, the PG trajectory model failed in modelling global f_0 declination. The degradation is probably due to the increase in the dimensionality of the optimization problem when accounting for long-term trajectory constraints. In the absence of an explicit formulation of the gradient, the optimization method obviously failed to account for the long-term dependencies. Not surprisingly, this results both into local and global degradation in the synthesized f_0 variations.

9.4.2.6 Conclusion

In this section, a trajectory model based on the stylization and the joint modelling of f_0 variations over various temporal domains was proposed. In the proposed approach, f_0 variations are stylized with a Discrete Cosine Transform, and modelled simultaneously over various temporal domains that covers short-term and long-term variations. During the training, a context-dependent model is estimated according to the joint stylized f_0 contours over the syllable and a set of long-term temporal domains. During the synthesis, f_0 variations are inferred using the long-term variations as trajectory constraints. The evaluation consisted in a subjective comparison of different speech prosody models in speech synthesis. Four models were compared: syllable-based trajectory models trained with respect to different long-term temporal domains (1-order syllable context, minor prosodic group (AG), major prosodic group (PG)), and the conventional HTS model. Evaluation was conducted using the HMM-based speech synthesis system.

The 1-order trajectory model proved to be significantly preferred to the conventional model, and to the other trajectory models. Each of the trajectory models succeeds in modelling f_0 contours that are consistent with the considered temporal domains. However, the ability of the trajectory model to account for long-term variations decreases when the temporal domain increases, due to the increase in complexity of the optimization process. In further studies, the relationship between static and dynamic trajectories will be explicitly formulated [Latorre and Akamine, 2008, Qian et al., 2009], and different combinations of trajectory constraints will be evaluated. Finally, the formulation of the trajectory model will be extended to the modelling of the local speech rate variations.

9.5 Conclusion

In this chapter, a context-dependent model based on continuous HMMs was presented to model the acoustic variations of speech prosody. Linguistic and statistical modelling refinements were proposed so as to enrich the description of the linguistic context used to model the speech prosody variations, and to improve the modelling of short-term and long-term speech prosody variations.

Firstly, the role of linguistic context in speech prosody modelling was assessed. A linguistic chain that includes surface and deep syntactic parsing was presented in order to enrich the description of the linguistic contexts that are used to model speech prosody variations. The enrichment of the linguistic context was shown in a subjective evaluation to significantly vary the synthesized speech prosody. A significant change in speech prosody was obtained for the modelling of f_0 variations, but not for the state-durations. The modification of speech prosody mostly concerns the realization of prosodic contours, the dynamic of prosodic prominences, and prosodic phrasing. However, the observed modifications remain local and not systematic. In particular, no global change in speech prosody was observed either for f_0 or for state-duration variations depending on specific syntactic constructions (e.g., incises, relative clauses).

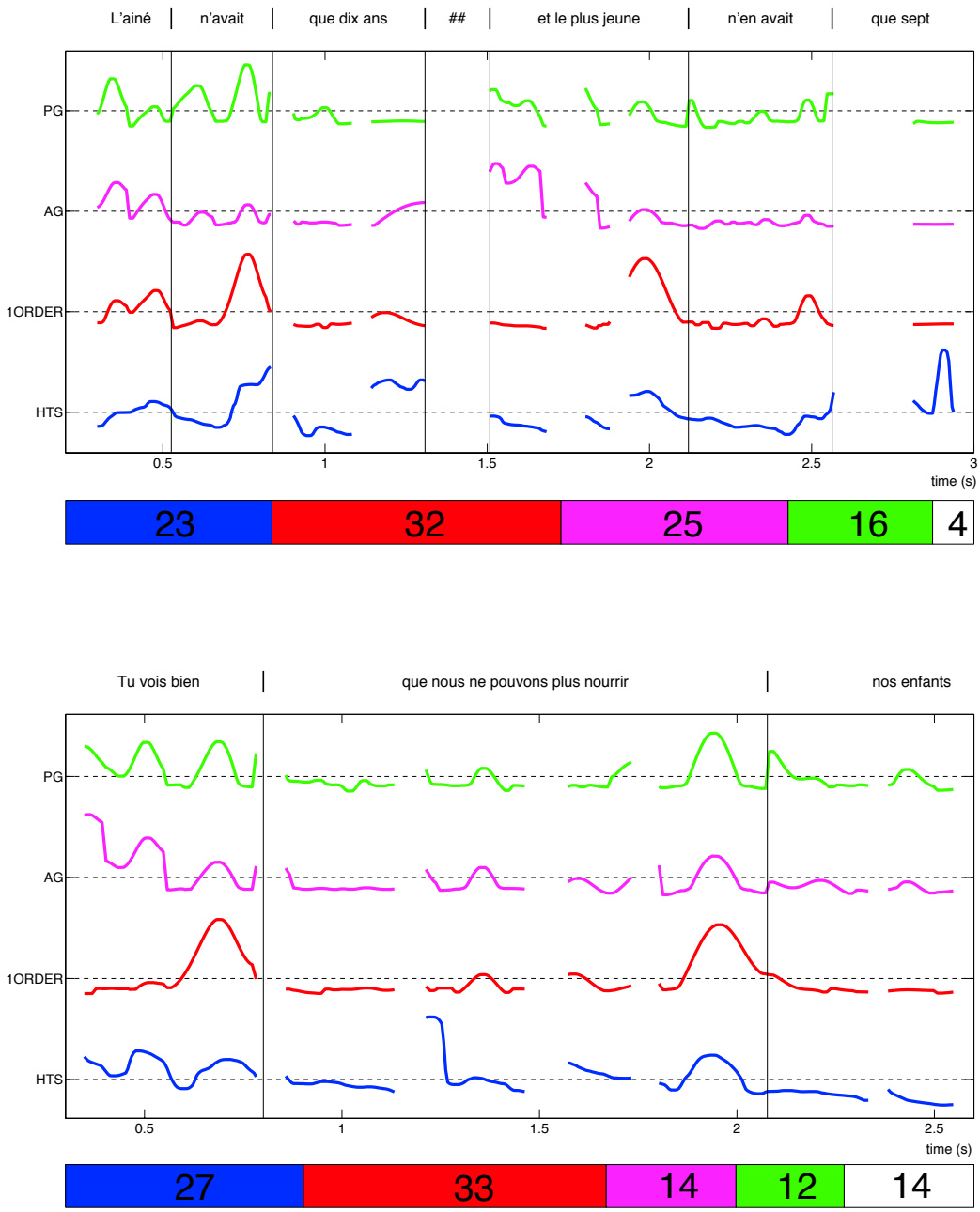


Figure 9.10: Comparison of synthesized f_0 variations depending on the trajectory model, with the associated preference scores.

Secondly, a trajectory model was proposed to model simultaneously the speech prosody variations over various temporal domains. In the proposed trajectory model, f_0 variations are represented simultaneously over different temporal domains, and stylized with a Discrete Cosine Transform, and duration variations are directly modelled at the syllable-level unit. During the training, the long-term variations are used to refine the clustering of the linguistic contexts. During the synthesis, f_0 variations are determined under the constraint of the long-term trajectories. The

proposed approach was compared to the conventional HTS model, and different trajectory models were compared with respect to the temporal domain considered. The 1-order syllable-context trajectory model was proved to significantly outperform the other speech prosody models in a subjective evaluation. Each of the trajectory models consistently models the f_0 contours associated with the temporal domains considered. However, the trajectory model failed to efficiently account for long-term trajectory constraints due to the increase in complexity of the optimization procedure.

Incidentally, the evaluation of speech prosody models is questioned. In particular, the comparison of different speech prosody is problematic, since various equally possible alternatives in speech prosody exist for a given utterance. Additionally, the speech prosody to compare frequently differ in many respects : a given speech prosody may be preferred to another one according to some local and/or global differences, while the other may be preferred according to some other differences. Consequently, a global preference does not account for such multi-decision preferences. Finally, the evaluation should clearly distinguish the *naturalness*, the *liveliness*, and the *variety* of the synthesized speech prosody.

correctness : the speech prosody adequate reproduces that which can be expected from a native speaker. The correctness directly relates to the intelligibility of speech, since speech prosody is used by a speaker and a listener to organize the acoustic content so as to clarify the meaning that the speaker intends to convey.

variety : the variety of speech prosody refers to the variation of speech prosody depending on the context. The variety of speech prosody relates both to the intra and inter speech prosody variations that occur over utterances. Intra-variations denote the variations that occur within a speech utterance, while inter-variations denote the variations that occur across speech utterances.

liveliness : The liveliness of a speech prosody includes the variety and the dynamic of speech prosody variations. Variety refers to the variety of prosodic contours that are observed within an utterance and across utterances depending on the context. Dynamic refers to the actual dynamic in the realization of a particular prosodic contour, the speech prosody variations over an utterance, and the speech variations across utterances.

The variety of the synthesized speech prosody is a major criterion to ensure the naturalness of speech synthesis that is required for high-quality applications. However, this variety cannot be evaluated with separated utterances solely. The design of an efficient methodology to assess and compare speech prosody would clearly benefit to the evaluation of speech synthesis systems.

Part III

Speaking with Style: Modelling Speaking Style

Abstract

Each speaker has his own *speaking style* that constitutes his vocal signature, and a part of his identity. Nevertheless, a speaker continuously adapts his speaking style according to specific situations of speech communication, and emotional states. Each situation determines a specific mode of production associated with it - a *genre* - which is defined by a set of conventions of form and content that are shared among all of its productions [Bühler, 1934, Benvéniste, 1966, Bakhtin, 1984, Koch and Oesterreicher, 2001]. In particular, a specific discourse genre (DG) relates to a specific *speaking style* [Fonagy, 1983, Léon, 1993, Lacheret et al., 2009, Simon et al., 2009, Degand and Simon, 2009]. Consequently, a speaker adapts his speaking style to each specific situation depending on the formal conventions that are associated with the situation, his a-priori knowledge about these conventions, and his ability to adapt his speaking style. Thus, each communication act instantiates a style which is composed of a style that is *particular* to the individual, and a *conventional* speaking style that is conditioned by a specific situation.

In speech synthesis, methods have been proposed to model the acoustic speech characteristics of a speaking style, with applications to emotional HMM-based speech synthesis and adaptation [Yamagishi et al., 2004, Yamagishi, 2006]. In the meanwhile, methods have been proposed to model and adapt the symbolic speech characteristics of a speaking style [Schmid and Atterer, 2004, Bell et al., 2006]. However, no study exists on the simultaneous modelling of symbolic and acoustic characteristics of speaking style, and speaking style acoustic modelling is generally limited to the modelling of emotion, with rare extensions to other sources of speaking styles variations [Krstulović et al., 2007]. The high-quality synthesis of speech and the adaptation of speaking style is a desired requirement in many multi-media applications (e.g., avatars, video games, interactive systems).

This part is dedicated to the study of *discourse genres* (DGs) and *speaking style* modelling. This part presents a study on the modelling of speaking style for speech synthesis, and addresses the issue of speaking style from the cognitive description of speaking styles to the modelling in speech synthesis. A preliminary experiment investigates whether listeners can distinguish speaking styles related to different situations of communication (chapter 10). The identification ability of speaking styles and the similarity that exists across different speaking styles is used to instantiate a reference for the evaluation of speaking style modelling in speech synthesis. In parallel, an average discrete/continuous context-dependent HMM is used to model the symbolic/acoustic characteristics of speaking style in speech synthesis. The ability of the model to model the speech characteristics of a speaking style is assessed (chapter 11). Finally, a speaker-independent modelling of speaking style based on shared context-dependent modelling and speaker normalization is presented to adapt the speaking style of a speaker in speech synthesis (chapter 12). The ability of listeners to distinguish speaking styles (natural speech and synthetic speech) is based on a perception experiment with delexicalized speech, and the identification obtained with synthetic speech is compared to that obtained with natural speech, and discussed.

Chapter 10

Expectations for Speaking Style: a Preliminary Study

Contents

In this chapter, the ability of listeners to associate a speaking style with a situation of speech communication - a discourse genre (DG) - is addressed, and the extent to which a speaking style is shared among speakers and listeners depending on their language background is investigated. The concept of discourse genre has been widely studied in rhetoric and literature and more recently extended to the oral domain ([Halliday, 1985, Biber, 1988], and [Broth et al., 2005] for a comprehensive study on French media speech.). Each situation and each given social context correspond to a specific mode of production - a *genre* - [Bühler, 1934, Benvéniste, 1966, Bakhtin, 1984] which is defined by a set of conventions of form and content (semantic, syntactic, phonological) that are shared among all of its productions.

The concept of genre originates from research in textual typology whose primary aim is to: 1) describe the diversity of discourses (e.g., literary, legal, political, religious); 2) understand the articulation of discourses in genres ([Rastier, 1989]); and 3) determine the formal markers of discourses genres, in particular the co-occurrence of specific linguistic cues that can be considered as being typical of a genre. In oral, studies focused on the definition and the description of *phonostyles* ([Fonagy, 1983, Léon, 1993, Simon et al., 2009]). In particular,

public discourse (such as political, religious, journalistic and sports discourses), considered as cultural stereotypes, are related to expressive strategies that act as markers of a phonostyle ([Lacheret-Dujour and Beaugendre, 1999, Degand and Simon, 2009]). However, studies on speaking style remain generally descriptive, and no study exists that addresses whether a speaking style is shared among speakers and listeners of a language, whether a speaking style is specific to a language or universal shared, and finally to which extent a similarity in the genre relates to a similarity in the style.

In this chapter, a preliminary experiment on the identification of speaking style is presented. In this study, four DGs are compared: church service (M), political speech (P), journalistic review (J), and sports commentary (S). The text and speech material are presented in section 10.1. A formal description of the DGs used is presented in section 10.2. Then, an experiment for the identification of speaking style is presented and discussed in section 10.3. In particular, the formal description of a situation of communication and the identification of a speaking style are compared and discussed. The identification experiment with natural speech will be used in further chapters as a reference for the evaluation of speaking style modelling in speech synthesis.

10.1 Design of a Speaking Style Database

10.1.1 Corpus Design

For the modelling of speaking style in speech synthesis, a 4-hour multiple-speakers speech database was designed from which the stimuli for the present experiment were selected. The corpus consists of four different DG's: catholic mass ceremony (M), political (P), journalistic (J), and sports commentary (S). In order to limit the intra-variability of the DGs, the different DGs were restricted to male speakers only and to specific discourse situations.

The following is a description of the four selected DG's:

1. **mass** : Christian church sermon (pilgrimage and Sunday high-mass sermons); single speaker monologue, no interaction.

2. **political** : new Year's speech; single speaker monologue; no interaction.

3. **journal** : radio review (press review; political, economical, technological chronicles); almost single speaker monologue with a few interactions with a lead journalist.

4. **sports commentary** : soccer; two speakers engaged in monologues with speech overlapping during intense soccer sequences and speech turn changes; almost no interactions.

The speech database consists of *natural speech* audio contents in compressed audio format (mp3 format with various encoding) that were collected from various multi-media applications on the internet, and with strongly variable audio quality (background noise: crowd, audience, recording noise, and reverberation). Recordings date from the 2000's with the exception of the political speech that homogeneously ranges from 1975 to 2007. The sample collection was especially designed to provide a well-balanced speaking-style speech database in terms of total duration of a speaking style and mean duration per speaker[1]. The characteristics of the speech database are summarized in table 10.1, and described in more details in table 10.2.

The speech database was processed in the same manner to that described in chapter 5.

10.1.2 Text Analysis

The text analysis includes manual orthographical transcription, automatic form and sentence segmentation, and automatic surface and deep syntactic parsing.

[1] This was reached with the exception of the sports commentary which has half duration than the other DG's

speaking style	media	# speaker	speaker gender	mean duration / sample	mean duration / speaker	total duration
mass	-	7	7M	12mn	11mn	**1h20**
political	TV	5	5M	12mn	14mn	**1h10**
journal	radio	5	5M	4mn	14mn	**1h10**
sport	radio	4	4M	20mn	9mn	**35mn**

Table 10.1: Description of the speaking style speech database.

text transcription The text was manually orthographically transcribed.

text segmentation The linguistic processing chain was used to segment the transcribed text into forms and sentences [Sagot and Boullier, 2005].

text analysis The linguistic processing chain was used to perform surface and deep syntactic parsing on the segmented sentences [Sagot, 2010, Villemonte de La Clergerie, 2005b].

The text analysis instantiates the following syntactic units: form, chunk, adjunction, and sentence.

10.1.3 Speech Analysis

The speech analysis includes acoustic feature extraction, speech segmentation and automatic prosodic transcription.

speech segmentation The speech material was phonetically aligned to the text transcription using the HMM-based phoneme segmentation IRCAMALIGN system [Lanchantin et al., 2008] based on the *hidden Markov model toolkit* (HTK, [Young et al., 2002]), then manually corrected.

prosody transcription automatic prosodic transcription was performed using the IRCAMPROM system [Obin et al., 2008c, Obin et al., 2009b]

The speech analysis instantiates the following speech units: phoneme, syllable, minor prosodic group, major prosodic group, and utterance.

The fundamental frequency f_0 and periodicity measure were estimated using the STRAIGHT algorithm [Kawahara et al., 1999a], and manually adapted so as to fit the characteristics of the speakers.

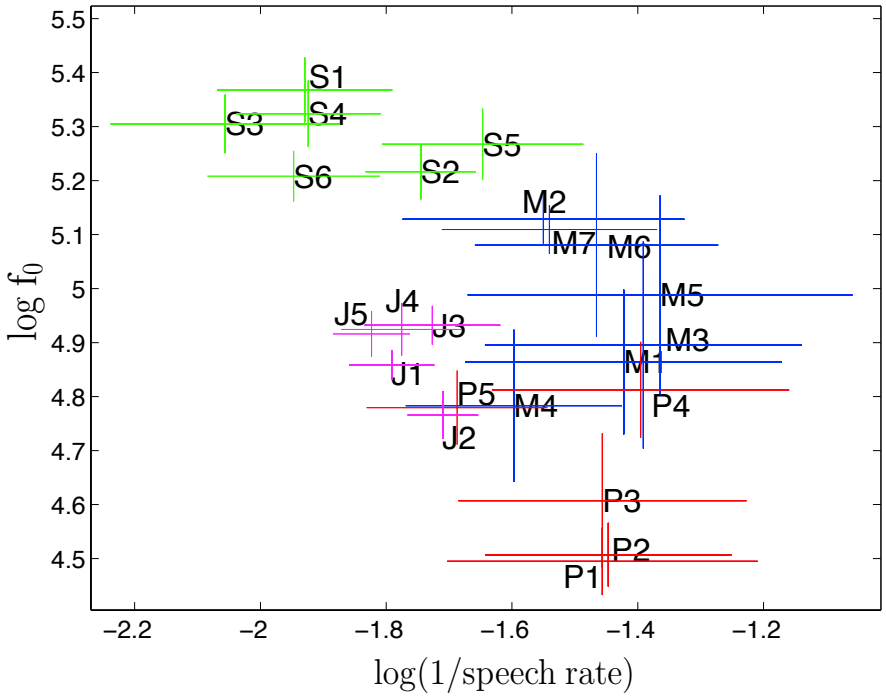


Figure 10.1: Prosodic description of the DGs. Mean and variance of f_0 and speech rate.

DG	M			P			J			S		
acoustic												
segmentation												
utterance	598			454			840			743		
syllable	#	# / utt.		#	# / utt.		#	# / utt.		#	# / utt.	
	17698	30		15430	**34**		23395	28		12059	**16**	
phone	#	# / utt.		#	# / utt.			# / utt.		#	# / utt.	
	37320	62		32710	**72**		50846	60		24633	**33**	
prosodic analysis	#	# / utt.	%	#	# / utt.	%	#	# / utt.	%	#	# / utt.	%
T	5132	8.6	**35**	3973	8.8	30	4751	5.6	22	2844	4.0	27
FM	2065	**3.5**	14 (**40**)	1686	**3.7**	13 (**42**)	1364	1.6	6 (28)	1322	1.8	13 (**46**)
Fm	1417	2.4	10 (27)	1001	2.2	7 (25)	1764	2.1	8 (**37**)	768	1.0	7 (27)
P	1650	**2.7**	11 (**33**)	1286	**2.8**	10 (27)	1623	1.9	7 (**35**)	754	1.0	7 (27)
lingusitic												
segmentation												
sentence	598			454			840			743		
forms	9755	16		8212	**18**		13006	15		6959	**9**	
parsing												
coverage %	70			71			73			**62**		
ambiguity	1.22			1.18			1.28			**0.78**		

Table 10.2: Descriptive analysis of the text and speech material depending on the DG.

10.2 Formal Description: Speech in Situation

This section investigates whether discourse genres (DGs) can be distinguished with respect to the situation in which a speech communication is produced. The situation of a DG is described according to the *conceptual scale* proposed in [Koch and Oesterreicher, 2001]. The conceptual scale is used to provide a formal description that aims at "*classifying the communicative behavior of interlocutors according to the constraint that are associated with the situation and the context in which the communication is produced*[2]" [Koch and Oesterreicher, 2001]. In particular, the conceptual scale accounts for the more or less formal nature of a communication, and in particular of a DG. In the experiment, four DGs are compared: church service (M), political speech (P), journalistic review (J), and sports commentary (S). Three expert linguists described the situational context of each of the DGs according to the conceptual scale. The formal description is briefly discussed and will be compared to that obtained from the speaking style identification experiment to assess to which extent a similarity in the situation relates to a similarity in the speaking style.

10.2.1 Experimental Design

10.2.1.1 Participants

Three expert linguists participated to this experiment. Participants are expert in DG analysis with various linguistic backgrounds: speech prosody, syntactic-prosodic interface, and speech synthesis.

10.2.1.2 Stimuli

Stimuli consisted of the list of the four DGs used: church service (M), political speech (P), journalistic review (J), and sports commentary (S). However, the situational context of a DG may vary significantly from one production to the other. For instance, a political speech may be associated with a large range of situations, from TV allocution, to journalistic interview, political debates, or political meetings. In the proposed speaking style speech database, each DG was restricted to a very specific situation (see section 10.1 for a precise description of the situation associated with each of the DGs) so as to limit their variability. In order to describe precisely the context in which the DGs occurred, participants were additionally given information about their specific context, and the possibility to access the speech database.

10.2.1.3 Procedure

The experiment consisted of the description of the DGs with respect to the *conceptual* scale. The conceptual scale consists of a set of ten cues that are used to describe different aspects of the situation in which a communication is produced. Each cue is associated with a scale from formal (*distance* language) to informal (*proximity* language). For instance, an *informal* communication would be associated with a spontaneous, interactive, and emotional dialogue in the presence of intimate interlocutors, while a *formal* communication would be associated with a prepared monologue discourse addressed to unknown interlocutors, with a spatial and temporal separation, no interaction, and no emotional content. Then, any communication instantiates an intermediate configuration that stands between a purely formal and purely informal communication, according to the more or less formal aspects of the situation in which it occurs. A description of the conceptual scale is presented in figure 10.2. The number of degrees used for the description depends on the desired precision and the number of communications to be described and distinguished: roughly, the larger the set of communications is, the finer the description must be. The precise number of degrees varies depending on the study (10 [Koch and Oesterreicher, 2001], and 5 degrees [Simon et al., 2009], respectively). In this study, four DGs were being compared, and a 3-degree scale was used for the description: immediate, distance, and an additional intermediate degree.

[2] such a description is independent of the nature of the medium, graphic or oral.

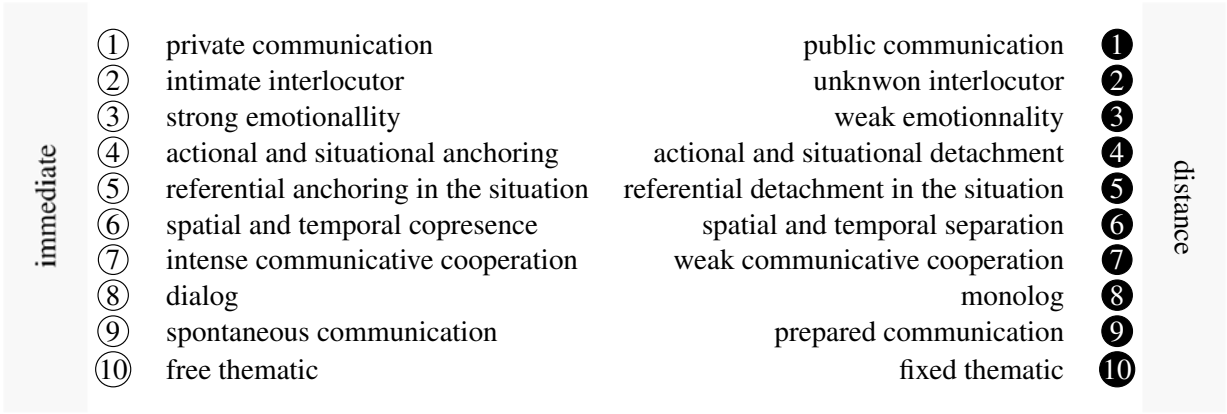


Figure 10.2: Description of the [Koch and Oesterreicher, 2001] conceptual scale.

Participants were asked to describe the four DGs according to the 3-degree conceptual scale. DGs were presented to the participants with respect to their nominal designation (church service, political speech, journalistic review, and sports commentary), additional information about their specific context, and the possibility to access the speech database. Participants accomplished the description independently from each other.

10.2.2 Results & Discussion

The rounded mean situational profile of each of the DGs is presented in figure 10.3. Journalistic review (J) appears as the more formal DG (100% of the cues were rated as strictly formal, 0% as strictly informal), and sports commentary (S) as the more informal (40% of the cues were rated as strictly informal, 20% as strictly formal). However, sports commentary appears fairly informal only, due to the medium separation, the public nature of the communication, the constrained thematic and the formal structure of the commentary (speakers are sequentially engaged in monologues with a relatively slight degree of interaction). Church service (M) and political speech (P) appear as mostly formal, but in an intermediate position compared to the journalistic review (J) (60% and 70% of the cues were rated as strictly formal, 10% and 0% for the church service and the political speech respectively). Church service and political speech distinguish with respect to a single cue only (church service is produced in the presence of the audience, while a media separation is observed for the political speech).

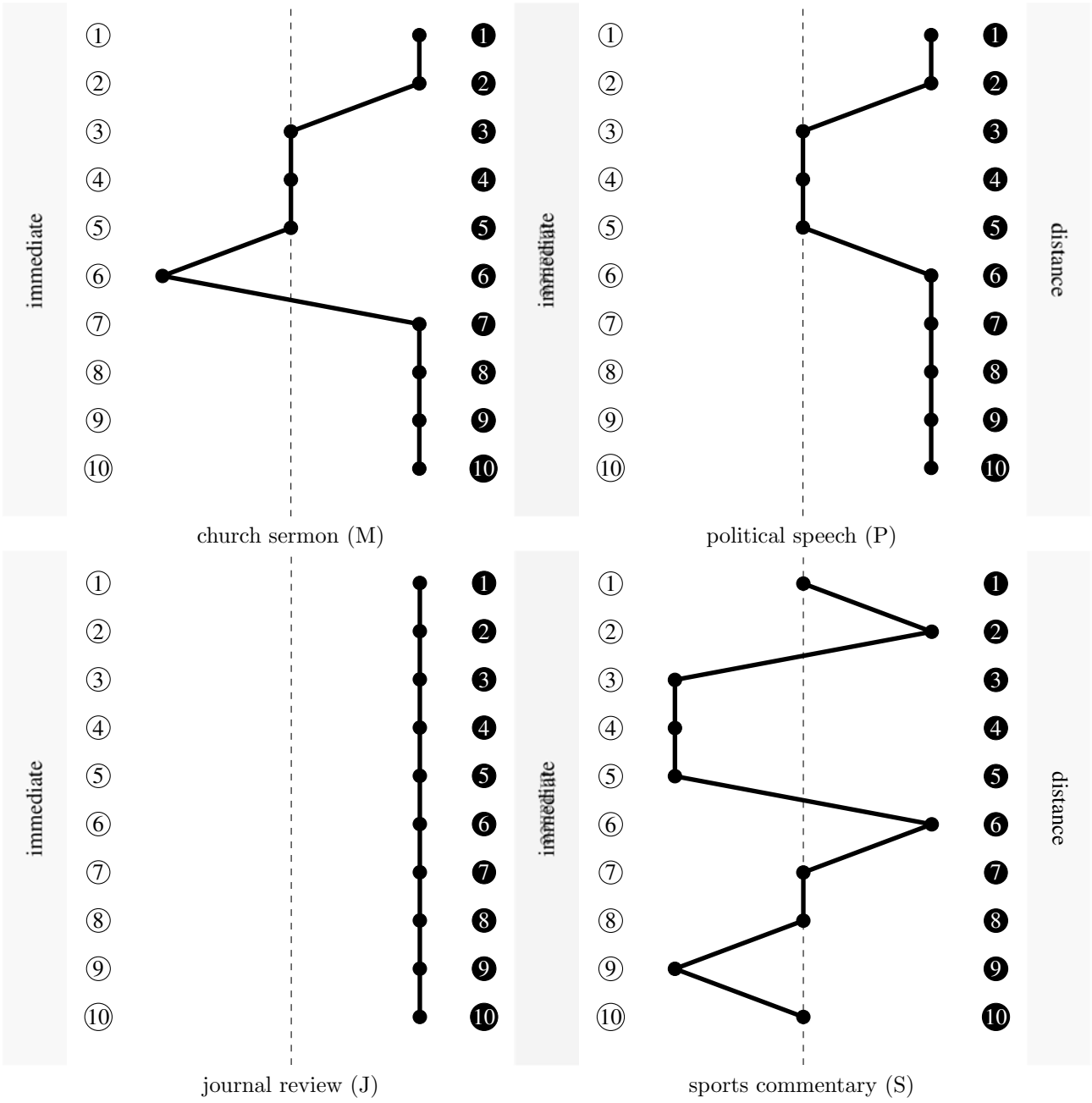


Figure 10.3: Profile of the DGs with respect to the conceptual scale. Description is provided according to the rounded mean annotation.

10.3 Expectations for Speaking Style

This section investigates whether listeners can distinguish speaking style that relates to different discourse genres (DGs). In a perception experiment with delexicalized speech, 70 listeners with varying experience in French (native speakers, non-native speakers, and non-speakers) were asked to identify the speaking style that relates to a DG. In the experiment, four DGs are compared: church service (M), political speech (P), journalistic review (J), and sports commentary (S).

10.3.1 Experimental Design

The experiment consisted of a multiple choice identification task based on the perception of speaking style. Speech utterances were collected from the speaking style speech database, filtered to remove lexical content[3] and presented as a multiple choice identification experiment to listeners with various language backgrounds in a crowd-sourcing framework.

10.3.1.1 Participants

72 subjects participated to this experiment. This includes: 37 native French speakers, 20 non-native French speakers, 15 non-French speakers; 46 expert participants, 26 naïve participants. *Expert* participants were actually coming from various domains (speech and audio technologies, linguistic, musicians). 7 participants were removed because they did not process the experiment entirely or because they did the experiment several times. In the case of multiple participation of a participant, his first participation was used for analysis only. Participants were aged from 20 to 65 years, with a strong proportion (65%) within the 20-35 year range.

10.3.1.2 Stimuli

40 speech utterances (10 per DG) were selected in the speaking style speech database to provide various and representative prosodic patterns of each speaking style[4].

Firstly, speech utterances were segmented into *prosodic periods* [Avanzi et al., 2008].

Secondly, the selection of speech utterances was derived from an attempt to classify speech utterances into discursive sequences. In particular, *archetypal speech utterances* were extracted depending on the DG[5]. Additionally, speech utterances were classified into *discursive sequences* depending on the DG. For instance, journalistic chronicles can be formally described as a sequence of topic sequences with punctual interaction with a lead speaker during topic changes (introduction/development/transition/conclusion). Speech utterances were thus classified into global introduction from a lead speaker, and initial, medium, terminal and transitional sequences for each topic. sports commentary sequences were classified depending on the context of the action and the situation (e.g., on-line comment of the current action, summary of the past actions, off-line comments), and emotional content (intensity of the action being commented). Other DG's speech utterances were classified in the similar manner.

Thirdly, speech utterances were classified into short (4 ± 0.5s.) and long ($10 \pm 1s$.) utterances that were homogeneously distributed for each DG.

[3]Lexical content of a speech utterance is an evident cue for DG's identification, a single word or lexical construction being potentially a non-ambiguous cue to distinguish DGs: "*Dieu*" ("*God*"), "*Mes chers compatriotes*" ("*My fellow countrymen*"), "*l'actualité*" ("*news*"), "*but*" ("*goal*").

[4]In the absence of a comprehensive and systematic framework for the description of speech prosody and DGs, the following proposal for the segmentation into prosodic units and the selection of relevant prosodic pattern remains ad-hoc.

[5]For instance, "*Au nom du père et du fils, et du Saint-Esprit, ainsi soit-il, Amen.*" ("*In the name of the Father, the Son, and the Holy Spirit, Amen.*"), "*Mes chers compatriotes, vive la République et vive la France!* " ("*My fellow countrymen, long live the republic! Long live France!*"), "*Oh le but de Babel! le but de Babel! le but de Babel!*" ("*What a goal by Babel! Goal by Babel! Goal by Babel!*") were considered as archetypal utterances that comes with a stereotypical speech prosody.

Finally, 2 speech utterances were selected for each speaker in order to remove any identification based on the speaker.

Then, speech utterances were processed as follows for audio normalization and delexicalization:

1. background noise and reverberation removal with a noise cancellation algorithm ([Bogaards and Roebel, 2005]);

2. delexicalization using a low band-pass filter. Pass-band was chosen so as to insure that the lowest frequency of the fundamental frequency and the highest frequency of its first harmonic was included ([Bogaards and Roebel, 2005]);

3. active speech mean level normalization at -20dBov [Kabal, 1999];

4. compression in mp3 format at 192Kb/s.

10.3.1.3 Procedure

The experiment consisted of a multiple choice identification task based on the perception of the speaking style[6]. The experiment was conducted according to a source-crowding technique using web social networks[7]. Participants were given a brief description of the different speaking styles. No speech example of the different speaking styles was presented to the participants prior to the experiment. This was adopted in order to focus the participant on his own mental representation of the different DGs and their expected speaking styles.

P	political	(TV new year's speech)
J	journalistic	(radio review)
S	sports commentary	(soccer)
M	mass	(Christian sermon)

Participants were asked to associate a speaking style to each of the speech utterances. For this purpose, participants were given three options:

total confidence : select only one DG when certain of the choice;

confusion : select two different DGs when a confusion between two likely DGs exists;

total indecision : select "indecision" when completely unsure. Participants were asked to use this possibility only as a very last resort.

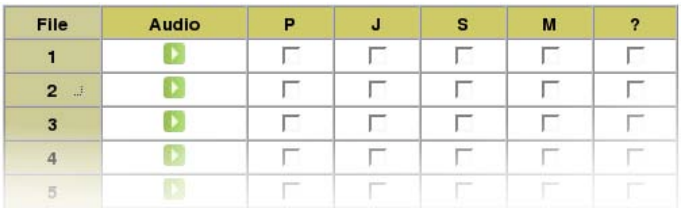


Figure 10.4: Illustration of the web interface used for the speaking style identification experiment.

Additional information were gleaned from the participants: speech expertise (expert, naïve), language (native French speaker, non-native French speaker, non-French speaker), age, and listening condition (headphones or not). Participants were encouraged to use headphones.

[6]the experiment is available at the following link: `http://recherche.ircam.fr/equipes/analyse-synthese/obin/pmwiki/pmwiki.php?n=Main.SSRecoProso`, and the original speech utterances on: `http://recherche.ircam.fr/equipes/analyse-synthese/obin/pmwiki/pmwiki.php/Main/SSORIG`

[7]*Ircam Analysis and Synthesis Perceptual Experiments* on Facebook: `http://www.facebook.com/group.php?gid=150354679034&ref=ts`

10.3.2 Results

Identification performance was estimated using a measure based on Cohen's Kappa statistic [Cohen, 1960]. Cohen's Kappa statistic measures the proportion of agreement between two raters with correction for random agreement. Our measure monitors the agreement between the ratings of the participants and the ground truth. The resulting Kappa value is considered as a measure of identification performance. The measure varies from -1 to 1: -1 is perfect disagreement; 0 is chance; 1 is perfect agreement. *Confusion* ratings (15% of the total ratings) were considered as equally possible ratings. *Total indecision* ratings (7% of the total ratings) were removed.

Overall score reveals fair identification performance ($\kappa_{\text{natural}} = 0.45 \pm 0.03$). Actually, the identification performance significantly depends on the DG. sports commentary is substantially identified ($\kappa^{(S)}_{\text{natural}} = 0.70 \pm 0.03$), journalistic review is fairly identified ($\kappa^{(J)}_{\text{natural}} = 0.54 \pm 0.05$), church service and political speech are slightly identified ($\kappa^{(M)}_{\text{natural}} = 0.38 \pm 0.04$ and $\kappa^{(P)}_{\text{natural}} = 0.34 \pm 0.05$, respectively).

Significant differences exist depending on the native language background of the listeners (figure 11.2). Native French speakers performed substantial identification ($\kappa_{\text{natural,native}} = 0.58 \pm 0.02$), non-native French speakers fair identification ($\kappa_{\text{natural,non-native}} = 0.44 \pm 0.02$), and non-French speakers only slight identification ($\kappa_{\text{natural,non-speaking}} = 0.26 \pm 0.05$). ANOVA analysis (*one-way analysis of variance*) was conducted to assess whether the identification performance depends on the language of the participants. Analysis reveals a significant effect of the language ($F(2, 59) = 15, p < 0.001$). Post-hoc analysis reveals significant difference between native French speakers and the others ($F(1, 52) = 13, p < 0.001$, $F(1, 43) = 24, p < 0.001$) but no effect between non-native French speakers and non-French speakers ($F(1, 23) = 3, p = 0.07$).

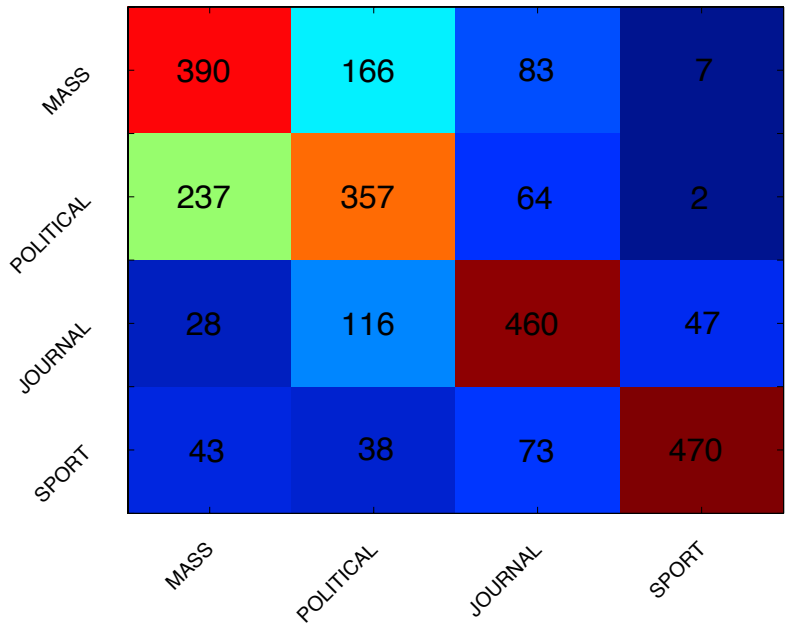


Figure 10.5: DGs confusion matrix. Rows represent synthesized speaking style. Columns represent identified speaking style.

Finally, the effect of the language background varies depending on the speaking style 10.3: sports commentary is widely identified regardless to the language background of the listener, while the identification of political speech drastically depends on the language.

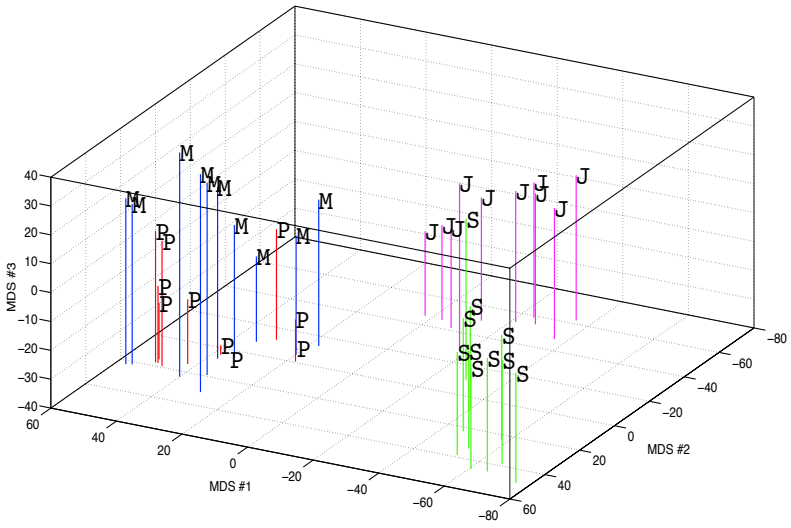


Figure 10.6: Representation of speech utterances according to their similarity after Multi Dimensional Scaling.

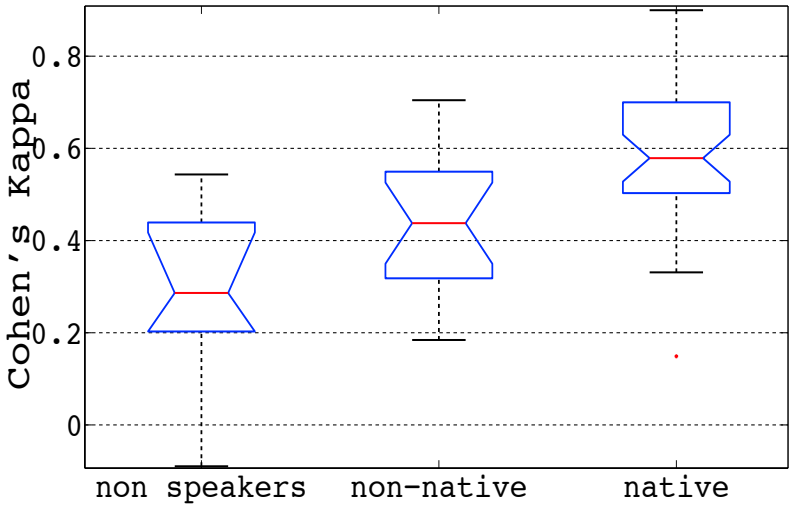


Figure 10.7: Identification performance according to the language background of the listener: non-French speaker, non-native French speaker, native French speaker (median, inter-quartiles, standard deviation).

Multi-Dimensional Scaling (MDS, [Borg and Groenen, 2005]) was used to represent and classify the speaking styles with regard to the observed perceptual confusion. For each speech utterance, the observed confusion matrix was used to estimate DGs confusion, and to define coordinates of speech utterance. A similarity distance between speech utterances was then estimated according to the L-1 metric (cumulative sum of absolute differences) (table 10.3).. Speech utterances similarity was then used to represent speech utterances into a 3-dimensional space according to multi-dimensional scaling 10.6 .

	native	non-native	non-speaking
M	0.48 (± 0.06)	0.33 (± 0.08)	0.30 (± 0.11)
P	0.45 (± 0.06)	0.27 (± 0.09)	0.02 (± 0.15)
J	0.63 (± 0.05)	0.54 (± 0.10)	0.33 (± 0.10)
S	0.78 (± 0.03)	0.65 (± 0.07)	0.60 (± 0.09)

Table 10.3: Identification of speaking styles depending on the language background of the listeners: native speaker, non native speaker, and non speaking. Mean Cohen's Kappa and 95% interval.

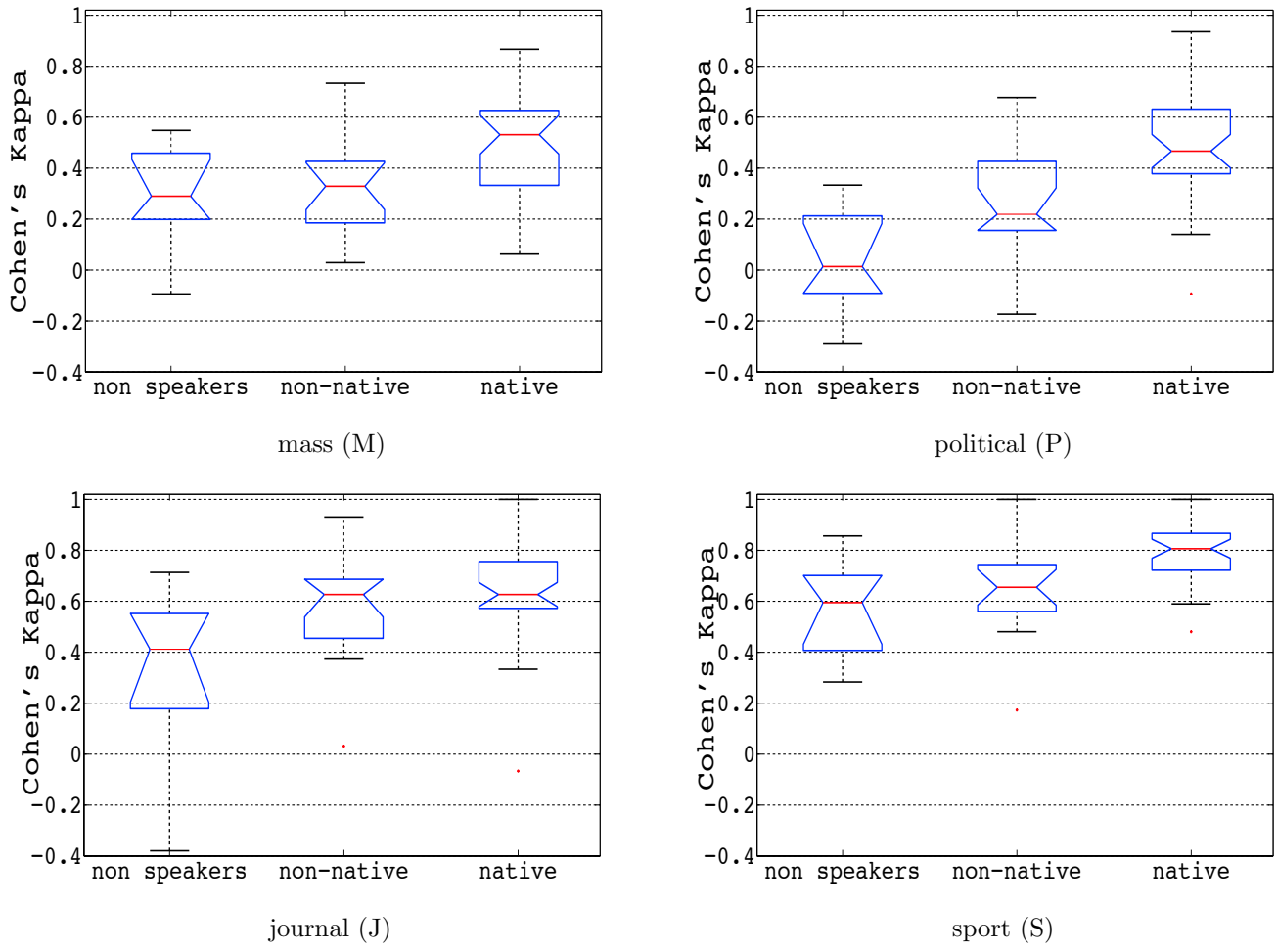


Figure 10.8: Comparison of identification performance with respect to the language background of the listener: non-French speaker, non-native French speaker, native French speaker (median, inter-quartiles, standard deviation).

10.3.3 Discussion

The experiment reveals fair identification performance based on the perception of the speaking style. This confirms evidence for the hypothesis that DGs relate to a specific speaking style that is shared among speakers and listeners.

The ability to identify a speaking style significantly depends on the language background. This shows that the abstract representation of a speaking style depends on the language, or more generally on the culture background[8]. More precisely, the experiment reveals that the language

[8]such hypothesis is supported by non-French speaking participants who report that they could not represent

factor significantly varies depending on the DG. This suggests that DG relate to conventions associated with a speaking style are more or less shared regardless to a specific language. In particular, the speaking style of sports commentary appears clearly common to all languages while the speaking style of church service and political speech dramatically depend on the language.

Non-native French speakers systematically stand between native French speakers and non-French speakers in the identification of a speaking style. However, this intermediate position varies depending on the speaking style, and non-native French speakers group either with the native French speakers or the non-French speakers (non-native French speakers group with non-French speakers for the church service and the sports commentary, with native French speakers for the journalistic review, and group either with the native French speakers neither with the non-French speakers in the case of the political speech). This indicates that non-native French speakers have a representation of a speaking style which results of the combination of their native and secondary languages.

The Multi-Dimensional-Scaling suggests that 3 dimensions suffice to distinguish the DGs. The first two dimensions significantly distinguish the journalistic review from the sport-commentary, and these from a group formed by the church service and the political speech. Finally, the third dimension fairly distinguishes the church service and the political speech. Naturally, the obtained dimensions remain to be interpreted and correlated to specific prosodic markers. Nevertheless, the dimensions may possibly correspond to a simple combination of global and local cues, such as speech rate, intensity, pausing, prosodic regularity, and prosodic contours[9].

A comparison of DG clusters as estimated from the formal description and the perception of a speaking style reveals a similar cluster structure. This confirms that a discourse context (situational, spatio-temporal, ... context) consistently relate to specific prosodic strategies. Moreover, the actual perceptual clusters emphasize and precise the distance across the different DGs (e.g., the journalistic review clearly distinguishes from the political speech and the church service on the prosodic dimension.) This result supports the hypothesis that prosodic strategies act as markers of a specific *speech act* ([Searle, 1969]) (for instance: neutrally describing an event with distanciation for the journalistic discourse vs. arguing and persuading for the political speech and mass discourse). In particular, the significant confusion that exists between the mass and the political speaking styles suggests that a similarity in the situation may relate to a similarity in the speaking style. sports commentary stands significantly apart from the other DGs. This confirms previous studies on the very specific nature of the sports commentary ([Deulofeu, 1998]), in particular in its iconical dimension: sportscaster does not only describe but vocally mimics the action being observed. This dimension is particularly emphasized in the case of radio sports commentary, where sportscaster must supply the absence of the image media.

10.3.4 Conclusion

An experiment on the identification of DGs based on the perception of speaking style was presented. The experiment consisted in the identification of different speaking styles from a set of 40 delexicalized natural speech utterances. Participants with various degrees of expertise (naïve/expert listener) and language backgrounds (non-French speaker, non-native French speaker, native French speaker) participated to the experiment. Four DGs were compared: church service (M), political speech (P), journalistic review (J), and sports commentary (S).
The experiment provided evidence that a specific communicative situation relate to a speaking style that is shared among speakers and listeners. Factorial analysis reveals that the identification significantly depends on the language of the listener, and is additionally clearly dependent on the DG. Interestingly, the clustering indicates that the characteristics of a situation consistently relates to the characteristics of a speaking style, then a similarity in the situation conducts to a similarity in the speaking style. Perceptual distances are even more salient compared to those observed for the situational classification, suggesting that the differences among DGs are orally

themselves "how sounds" a Christian sermon (religious dependency) nor political new year's speech (cultural dependency)

[9] as explicitly reported by the participants in their commentaries.

more pronounced and more stereotypical than those obtained from the description of the situation only.

The identification performance from natural speech will be used as a reference for the evaluation of speaking style speech synthesis in chapters 11 and 12.

Chapter 11

Average Discrete/Continuous Modelling of Speaking Style

Contents

11.1 Introduction

This chapter assesses the ability of HMM-based speech synthesis to model the speech characteristics of various speaking styles. In addition, the robustness of the HMM-based speech synthesis is evaluated in the conditions of real-world applications.

In speech synthesis, methods have been proposed to model the acoustic characteristics of a speaking style, with application to emotional HMM-based speech synthesis and adaptation [Yamagishi et al., 2004, Yamagishi, 2006]. In the meanwhile, methods have been proposed to model and adapt the symbolic characteristics of a speaking style [Schmid and Atterer, 2004, Bell et al., 2006]. However, no study exists on the simultaneous modelling of the symbolic and acoustic characteristics of speaking style, and speaking style acoustic modelling is generally restricted to the modelling of emotion, with rare extensions to other sources of speaking style variations [Krstulović et al., 2007]. The high-quality synthesis of speech and the adaptation of speaking style is a desired requirement in many multi-media applications (e.g., avatars, video games, interactive systems).

In this chapter, an average discrete/continuous HMM is presented to model the symbolic and the acoustic characteristics of a speaking style. The proposed model is used to model the average characteristics of a speaking style that is shared among various speakers depending on specific situations of speech communication. In particular, it is assumed that the speaking style is not restricted to the conventional characteristics of speech prosody (f_0 variations, durations) and includes timbre, voice quality, and phonatory strategies. A discrete/continuous HMM is used to model the average characteristics of four speaking styles associated with different situations of speech communication: church service (M), political speech (P), journalistic review (J), and sports commentary (S).

Firstly, the context-dependent discrete HMM described in chapter 8 is used to model the average symbolic characteristics of a speaking style. Secondly, the conventional HMM-based speech synthesis system is used to model the average acoustic characteristics of a speaking style [1]. During the training, the discrete/continuous context-dependent HMMs are estimated separately. During the synthesis, the symbolic/acoustic parameters are determined in cascade, from the symbolic representation to the acoustic variations. For each speaking style, an average speaking-style model is estimated based on the rich linguistic description presented in chapter 7. The evaluation consisted of an identification experiment of the speaking style based on delexicalized speech, and compared to a similar experiment conducted with natural speech.

The average discrete/continuous HMM model is briefly described in section 11.2 and used to model four speaking styles. The evaluation is presented and discussed in section 11.3. The identification of synthesized speaking style is compared to that obtained for natural speech in section 11.4.

11.2 Average Modelling of Speaking Style

A speaking style model $\boldsymbol{\lambda}^{(style)}$ is composed of discrete/continuous context-dependent HMMs that model the symbolic/acoustic speech characteristics of a speaking style.

$$\boldsymbol{\lambda}^{(style)} = \left(\boldsymbol{\lambda}^{(style)}_{\text{symbolic}}, \boldsymbol{\lambda}^{(style)}_{\text{acoustic}}\right) \tag{11.1}$$

11.2.1 Average Discrete Modelling

For each speaking style, an average symbolic model $\boldsymbol{\lambda}^{(style)}_{\text{symbolic}}$ is estimated from the pooled speakers associated with the speaking style.

The symbolic model consists of the context-dependent discrete HMM described in chapter 8 based on the RHAPSODIE prosodic grammar described in chapter 8 : *major prosodic boundary* (F_M, boundary of a major prosodic group), *minor prosodic boundary* (F_m, boundary of a minor prosodic group), and *prosodic prominence* (P, prosodic prominence). Speech prosody is automatically transcribed based on *Analor* [Avanzi et al., 2008] and *ircamProm* [Obin et al., 2008c], then converted into a sequential structure, and represented over syllable so as to account for all of the prosodic events simultaneously.

Let R be the number of speakers from which an average model $\boldsymbol{\lambda}^{(style)}_{\text{symbolic}}$ is to be estimated. Let $\mathbf{l} = \left(\mathbf{l}^{(1)}, \dots, \mathbf{l}^{(R)}\right)$ the total set of prosodic events observations, and $\mathbf{l}^{(r)} = \left[l^{(r)}(1), \dots, l^{(r)}(N_r)\right]$ is the sequence of prosodic events associated with speaker r, where $l^{(r)}(n)$ is the prosodic label associated with the n-th syllable.
Let $\mathbf{q} = \left(\mathbf{q}^{(1)}, \dots, \mathbf{q}^{(R)}\right)$ the total set of linguistic contexts observations, and $\mathbf{q}^{(r)} = \left[\mathbf{q}^{(r)}(1), \dots, \mathbf{q}^{(r)}(N_r)\right]$ is the linguistic context sequence associated with speaker r, where $\mathbf{q}^{(r)}(n) = [q_1^{(r)}(n), \dots, q_L^{(r)}(n)]^\top$ is the (Lx1) linguistic context vector which describes the linguistic

[1] At the time of this study, the context-dependent continuous HMM described in chapter 9 was not available

characteristics associated with the n-th syllable.

An average context-dependent discrete HMM $\boldsymbol{\lambda}^{(style)}_{\text{symbolic}}$ is estimated from the pooled speakers observations based on the speaker-dependent model described in chapter 8. Firstly, an average context-dependent tree $\text{T}^{(style)}_{\text{symbolic}}$ with terminal nodes $S^{(style)}_{\text{symbolic}} = (S^{(style)}_{\text{symbolic},1}, \dots, S^{(style)}_{\text{symbolic},M})$ is derived so as to maximize the information gain of the prosodic events $\mathbf{l}$ conditionally to the linguistic contexts $\mathbf{q}$. Then, a context-dependent HMM model $\boldsymbol{\lambda}^{(style)}_{\text{symbolic}}$ is estimated with respect to the context-dependent tree $\text{T}^{(style)}_{\text{symbolic}}$.

11.2.2 Average Continuous Modelling

For each speaking style, an average acoustic model $\boldsymbol{\lambda}^{(style)}_{\text{acoustic}}$ that includes source/filter variations, f_0 variations, and state-durations, is estimated from the pooled speakers associated with the speaking style.

Let R be the number of speakers from which an average model is to be estimated. Let $\mathbf{o} = \left(\mathbf{o}^{(1)}, \dots, \mathbf{o}^{(R)}\right)$ the total set of observations, and $\mathbf{o}^{(r)} = \left[\mathbf{o}^{(r)}(1), \dots, \mathbf{o}^{(r)}(T_r)\right]$ is the observation sequences associated with speaker r, where $\mathbf{o}^{(r)}(t) = \left[o^{(r)}_t(1), \dots, o^{(r)}_t(D)\right]^\top$ is the (Dx1) observation vector which describes the acoustical property at time t. Let $\mathbf{q} = \left(\mathbf{q}^{(1)}, \dots, \mathbf{q}^{(R)}\right)$ the total set of linguistic contexts observations, and $\mathbf{q}^{(r)} = \left[\mathbf{q}^{(r)}(1), \dots, \mathbf{q}^{(r)}(T_r)\right]$ is the linguistic context sequence associated with speaker r, where $\mathbf{q}^{(r)}(t) = [q^{(r)}_1(t), \dots, q^{(r)}_L(t)]^\top$ is the (L'x1) augmented linguistic context vector which describes the linguistic properties at time t.

An average context-dependent continuous HMM $\boldsymbol{\lambda}^{(style)}_{\text{symbolic}}$ is estimated from the pooled speakers observations based on the conventional HTS system ([Zen et al., 2009]), in a similar manner to the speaker-dependent model described in chapter 9. Firstly, a context-dependent HMM model is estimated for each of the linguistic contexts. Then, an average context-dependent tree $\text{T}^{(style)}_{\text{acoustic}}$ with terminal nodes $S^{(style)}_{\text{acoustic}} = (S^{(style)}_{\text{acoustic},1}, \dots, S^{(style)}_{\text{acoustic},M})$ is derived so as to minimize the description length of the context-dependent HMM model $\boldsymbol{\lambda}^{(style)}_{\text{acoustic}}$.

The acoustic module models at once source/filter variations, f_0 variations, and the temporal structure associated with a speaking style. Speakers f_0 were normalized with respect to the speaking style prior to modelling. Source, filter, and normalized f_0 observation vectors and their dynamic vectors are used to estimate context-dependent HMM models $\boldsymbol{\lambda}^{(style)}_{\text{acoustic}}$. Multi-Space probability Distributions (MSD) [Tokuda et al., 1999] are used to model variable dimensional parameter sequence such as the f_0 variations with respect to voiced regions. Each context-dependent HMM $\boldsymbol{\lambda}^{(style)}_{\text{acoustic}}$ has state-duration probability density functions (PDFs) to model its temporal structure [Zen et al., 2004].

11.2.3 Parameters Inference

During the synthesis, the text is first converted into a concatenated sequence of context-dependent HMM models $\boldsymbol{\lambda}^{(style)}_{\text{symbolic}}$ associated with the linguistic context sequence $\mathbf{q} = [\mathbf{q}_1, \dots, \mathbf{q}_N]$, where $\mathbf{q}_n = [q_1, \dots, q_L]^\top$ denotes the (Lx1) linguistic context vector associated with the n-th syllable.

Firstly, the sequence of prosodic events $\widehat{\mathbf{l}}$ is determined so as to maximize the probability of the sequence of prosodic events $\mathbf{l}$ conditionally to the linguistic context sequence $\mathbf{q}$ and the model $\boldsymbol{\lambda}^{(style)}_{\text{symbolic}}$.

$$\widehat{\mathbf{l}} = \underset{\mathbf{l}}{\text{argmax}}\ \text{p}(\mathbf{l}|\mathbf{q}, \boldsymbol{\lambda}^{(style)}_{\text{symbolic}}) \tag{11.2}$$

Then, the linguistic context sequence $\mathbf{q}$ augmented with the sequence of prosodic events $\widehat{\mathbf{l}}$ is

converted into a concatenated sequence of context-dependent models $\boldsymbol{\lambda}^{(style)}_{\text{acoustic}}$.

The acoustic sequence $\widehat{\mathbf{o}}$ is determined so as to maximize the probability of the acoustic sequence $\widehat{\mathbf{o}}$ conditionally to the model $\boldsymbol{\lambda}^{(style)}_{\text{acoustic}}$ and the sequence length T.

$$\widehat{\mathbf{o}} = \underset{\mathbf{o}}{\text{argmax}} \max_{\mathbf{q}} \text{p}(\mathbf{o}|\mathbf{q}, \boldsymbol{\lambda}^{(style)}_{\text{acoustic}}, T)\ \text{p}(\mathbf{q}|\boldsymbol{\lambda}^{(style)}_{\text{acoustic}}, T) \tag{11.3}$$

First, the state sequence $\widehat{\mathbf{q}}$ is determined so as to maximize the probability of the state sequence conditionally to the model $\boldsymbol{\lambda}^{(style)}_{\text{acoustic}}$ and the sequence length T.

$$\widehat{\mathbf{q}} = \underset{\mathbf{q}}{\text{argmax}}\ \text{p}(\mathbf{q}|\boldsymbol{\lambda}^{(style)}_{\text{acoustic}}, T) \tag{11.4}$$

Then, the observation sequence $\widehat{\mathbf{c}}$ is determined so as to maximize the probability of the observation sequence conditionally to the state sequence $\widehat{\mathbf{q}}$, the model $\boldsymbol{\lambda}^{(style)}_{\text{acoustic}}$ under dynamic constraint $\mathbf{o} = \mathbf{W}\mathbf{c}$.

$$\widehat{\mathbf{c}} = \underset{\mathbf{c}}{\text{argmax}}\ \text{p}(\mathbf{W}\mathbf{c}|\widehat{\mathbf{q}}, \boldsymbol{\lambda}^{(style)}_{\text{acoustic}}) \tag{11.5}$$

Assuming that each state probability density function is assumed to be a single normal distribution, the above equation can be reformulated as a set of linear equations which can be solved efficiently:

$$\mathbf{R}_{\widehat{\mathbf{q}}}\widehat{\mathbf{c}} = \mathbf{r}_{\widehat{\mathbf{q}}} \tag{11.6}$$

where:

$$\mathbf{R}_{\widehat{\mathbf{q}}} = \mathbf{W}^\top \boldsymbol{\Sigma}^{-1}_{\widehat{\mathbf{q}}} \mathbf{W}. \tag{11.7}$$

$$\mathbf{r}_{\widehat{\mathbf{q}}} = \mathbf{W}^\top \boldsymbol{\Sigma}^{-1}_{\widehat{\mathbf{q}}} \boldsymbol{\mu}_{\widehat{\mathbf{q}}}. \tag{11.8}$$

and $\boldsymbol{\Sigma}_{\widehat{\mathbf{q}}}$ and $\boldsymbol{\mu}_{\widehat{\mathbf{q}}}$ are respectively the covariance matrix and the mean vector for the sate sequence $\widehat{\mathbf{q}}$.

11.3 Evaluation

The experiment consisted of a multiple choice identification task based on the perception of speaking style, in parallel to that conducted for natural speech. This was achieved in order to compare the identification ability of natural speech and synthesized speech. For the purpose of the comparison, both experiments were based on delexicalized speech in order to alleviate the problem due to the presence of an explicit linguistic content.

11.3.1 Experimental Design

The experiment consisted of a multiple choice identification task based on the perception of speaking style. The procedure was identical to that described for natural speech in chapter 10. In particular, the same sentences were used to synthesize speech utterances with respect to the corresponding speaking style model. Then, the synthesized speech utterances were filtered to remove lexical content and presented as a multiple choice identification experiment to listeners with various language background in a crowd-sourcing framework.

11.3.2 Stimuli

The stimuli selection and processing were identical to those used for the identification of natural speech in chapter 10. In particular, the same 40 speech utterances (10 per DG) were selected, and processed in the same manner for normalization and filtering. Additionally, the speech utterances to be synthesized were removed from the speech database prior to the modelling.

11.3.2.1 Linguistic Contexts

Linguistic information were extracted from text using the linguistic processing chain described in chapter 7. The symbolic model was trained with the full rich linguistic feature set, which consists in segmental, prosodic, morpho-syntactic, dependency, constituency, and adjunction features. The acoustic model was trained with the full rich linguistic and the prosodic structure feature sets. The used linguistic units were (phoneme), syllable, and the syntactic units. Linguistic features were converted into linguistic contexts over the (phoneme) syllable by computing locational and weight contexts, and representing 1-order left-to-right contexts and 1-order child-to-parent contexts in the case of the dependency contexts.

Finally, the linguistic contexts used are defined as:

$$\begin{array}{lrcl} \text{symbolic:} & Q_{\text{symbolic}}^{(syllable)} & = & Q_{\text{segment}} \cup Q_{\text{morpho}} \cup Q_{\text{dep}} \cup Q_{\text{chunk}} \cup Q_{\text{adj}} \\ \text{acoustic:} & Q_{\text{acoustic}}^{(phone)} & = & Q_{\text{segment}} \cup Q_{\text{morpho}} \cup Q_{\text{dep}} \cup Q_{\text{chunk}} \cup Q_{\text{adj}} \cup Q_{\text{proso}} \end{array}$$

11.3.2.2 Training Corpus

Average speaking-style models were estimated on the speaking style speech database with respect to the considered speaking style minus the speech utterances to be synthesized.

11.3.2.3 Evaluation Corpus

The evaluation corpus was composed of the 40 speech sentences (10 per DG) that were used for the identification experiment based on natural speech in chapter 10.3.

11.3.2.4 Speaking style models

A speaking style model was estimated for each of the DGs:

$$\begin{array}{lllll} \boldsymbol{\lambda}^{(M)}: & 598 - 10 \text{ utterances}, & 1\text{h}20, & 7 \text{ speakers} \\ \boldsymbol{\lambda}^{(P)}: & 454 - 10 \text{ utterances}, & 1\text{h}10, & 5 \text{ speakers} \\ \boldsymbol{\lambda}^{(J)}: & 840 - 10 \text{ utterances}, & 1\text{h}10, & 5 \text{ speakers} \\ \boldsymbol{\lambda}^{(S)}: & 743 - 10 \text{ utterances}, & 35\text{mn}, & 4 \text{ speakers} \end{array}$$

For the symbolic modelling, a speaking-style model $\boldsymbol{\lambda}_{\text{symbolic}}^{(style)}$ is estimated based on average modelling. For the acoustic modelling, a speaking-style model $\boldsymbol{\lambda}_{\text{acoustic}}^{(style)}$ is estimated based on average modelling.

During the synthesis, the text is first converted into a concatenated sequence of context-dependent HMM models. Firstly, the sequence of prosodic events is determined so as to maximize the probability of the sequence of prosodic events conditionally to the linguistic context sequence and the model $\boldsymbol{\lambda}_{\text{symbolic}}^{(style)}$. Then, the sequence of acoustic variations is determined so as to maximize the probability of the acoustic sequence conditionally to the linguistic context sequence, the sequence of prosodic events, and the model $\boldsymbol{\lambda}_{\text{acoustic}}^{(style)}$. This finally results into 10x4 = 40 synthesized speech utterances to be identified.

dimension	**structure**	**acoustic**		
stream	prosodic structure	source/filter	duration	f_0
corpus				
training corpus	C-STYLE (4h, 1h per DG)			
evaluation corpus	C-STYLE (40 utterances, 10 per DG			
feature extraction				
feature	hierarchical prosodic structure F_M, F_m, P	5-order aperiodicity 39-order MFCC	state-duration	f_0
window	syllable	50-ms blackmann		
frame rate	syllable	5ms		
feature transform				
transform function	linearization	-	log	log
transform unit	syllable	1-orderΔ ,Δ^2	-	1-orderΔ ,Δ^2
model				
topology	discrete HMM ergodic	5-state HMM normal distribution semi-tied covariance	5-state HMM normal distribution	5-state MSD-HMM normal distribution semi-tied covariance
context	$Q^{(syllable)}_{\text{symbolic}}$	$Q^{(phone)}_{\text{acoustic}}$		
clustering	average context-clustering DT CART	average context-clustering DT ML-MDL		

Table 11.1: Evaluation of the *Average Voice Model*: model setup

11.3.3 Participants

50 subjects participated in this evaluation. This includes: 25 native French speakers, 15 non-native French speakers, 10 non-French speakers; 34 expert and 16 naïve participants. *Expert* participants were actually coming from various domains (speech and audio technologies, linguistic, musicians). In comparison, 72 subjects participated in the natural speech experiment and 23 did both.

11.3.4 Procedure

The evaluation consisted of an identification experiment of the speaking style based on delexicalized speech[2] The experiment was conducted according to a source-crowding technique using web social networks[3]. The procedure was identical to that described for natural speech in chapter 10.

11.4 Results & Discussion

Identification performance was measured using the measure based on Cohen's Kappa statistic that was presented in chapter 10. *Confusion* ratings were considered as equally possible ratings. *Total indecision* ratings were relatively rare (3% of the total ratings) and removed. Table 12.2 presents the recognition confusion matrix.

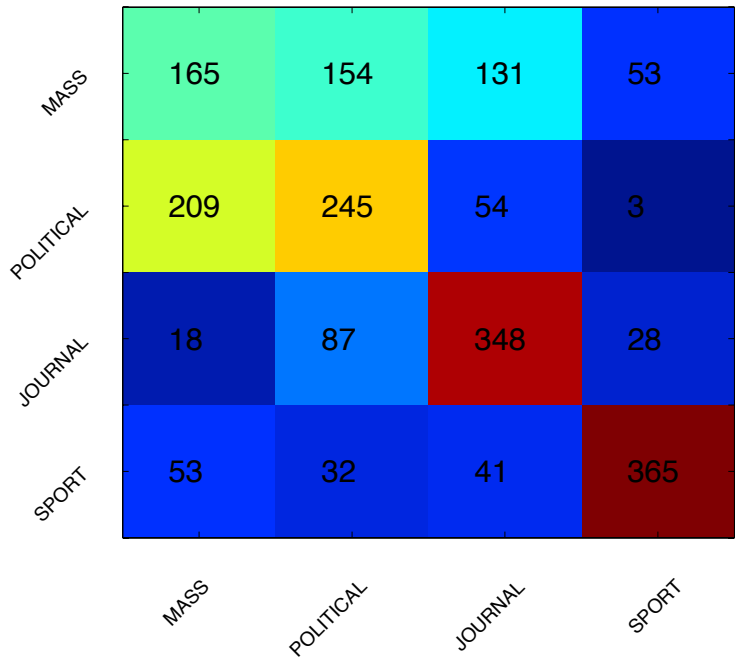


Figure 11.1: Average speaking style confusion matrix. Rows represent synthesized speaking style. Columns represent identified speaking style.

Overall score reveals a fair identification performance ($\kappa_{\text{average}} = 0.38 \pm 0.04$) which is comparable to that observed for natural speech ($\kappa_{\text{natural}} = 0.45 \pm 0.03$). The identification performance significantly depends on the speaking style (figure 11.2): sports commentary is substantially identified

[2]the experiment is available at the following link: `http://recherche.ircam.fr/equipes/analyse-synthese/obin/pmwiki/pmwiki.php/Main/HTSSSProsoEvaluation`.

[3]*Ircam Analysis and Synthesis Perceptual Experiments* on Facebook: `http://www.facebook.com/group.php?gid=150354679034&ref=ts`

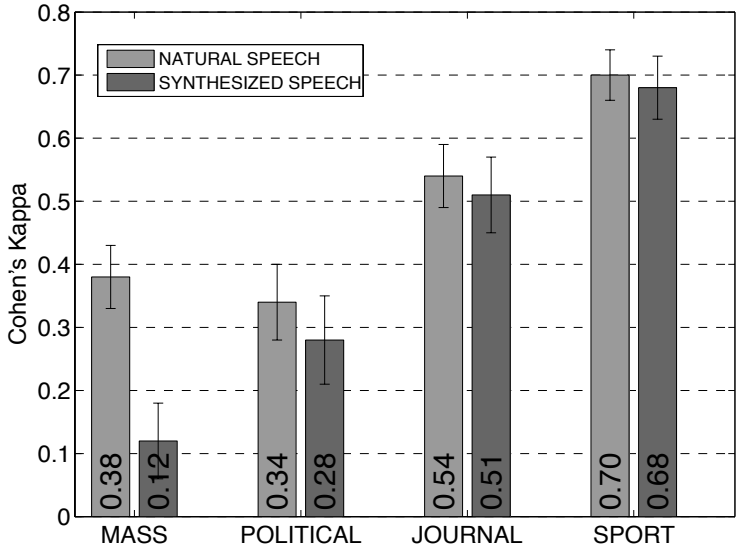


Figure 11.2: Comparison of identification scores for natural and synthesized speech. (mean distance and 95% confidence interval).

($\kappa^{(S)}_{\text{average}} = 0.68 \pm 0.05$), journal fairly identified ($\kappa^{(J)}_{\text{average}} = 0.50 \pm 0.06$), political speech moderately identified ($\kappa^{(P)}_{\text{average}} = 0.28 \pm 0.07$), and mass only slightly identified ($\kappa^{(M)}_{\text{average}} = 0.12 \pm 0.06$).

In comparison with the identification of natural speech, the identification is comparable in the case of the sports commentary and the journal speaking styles ($\kappa_{\text{natural}} = 0.70 \pm 0.03$ and $\kappa_{\text{natural}} = 0.54 \pm 0.05$, respectively). However, there is a drop in identification for the political and the mass speaking styles, especially significant for the latter ($\kappa_{\text{natural}} = 0.34 \pm 0.05$ and $\kappa_{\text{natural}} = 0.38 \pm 0.04$, respectively). This indicates that the model somehow failed to capture the relevant characteristics of the corresponding speaking style. Nevertheless, a large confusion exists between the political and the mass speech that is inherent to a similarity in the speech communication and in the speaking style. Additionally, the conventional HMM-based speech synthesis system failed into modelling adequately the breathiness and the creakiness that is specific to the political speaking style, especially within unvoiced segments.

ANOVA analysis was conducted to assess whether the identification performance depends on the language background of the listeners. Analysis reveals a significant effect of the language ($F(2, 59) = 15, p < 0.001$) (F(48,2)=5.9, p-value=0.005), and confirms results obtained for natural speech. This confirms evidence that a speaking style varies depending on the language and/or cultural background.

Finally, an informal evaluation of the quality of the synthesized speech suggests that the speaking style modelling is robust to the large variety of audio quality.

distance	mass	political	journal	sport
mass	-	0.07 ($\pm$ 0.1)	0.55 ($\pm$ 0.07)	0.67 ($\pm$ 0.08)
political	0.07 ($\pm$ 0.1)	-	0.60 ($\pm$ 0.08)	0.89 ($\pm$ 0.06)
journal	0.55 ($\pm$ 0.07)	0.60 ($\pm$ 0.08)	-	0.84 ($\pm$ 0.06)
sport	0.67 ($\pm$ 0.08)	0.89 ($\pm$ 0.06)	0.84 ($\pm$ 0.06)	-

Table 11.2: DGs distance in the perceptual space (mean distance and 95% confidence interval).

11.5 Conclusion

In this chapter, the ability and the robustness of HMM-based speech synthesis to model the speech characteristics of varsious speaking styles were assessed[4]. A discrete/continuous HMM was presented to model the symbolic and acoustic speech characteristics of a speaking style, and used to model the average characteristics of a speaking style that is shared among various speakers, depending on specific situations of speech communication. For each speaking style, an average speaking style model is estimated based on discrete/continuous HMM and rich linguistic contexts. During the training, discrete/continuous context-dependent HMMs are estimated separately. During the synthesis, the symbolic/acoustic parameters are determined in cascade, from the symbolic representation to the acoustic variations. The evaluation consisted of an identification experiment of four speaking styles based on delexicalized speech, and compared to a similar experiment conducted for natural speech. The evaluation showed that the discrete/continuous HMM consistently models the speech characteristics of a speaking style, and is robust to the differences in audio quality. This provides evidence that the discrete/continuous HMM speech synthesis system successfully models the speech characteristics of a speaking style in the conditions of real-world applications.

[4]Examples of synthesized speech are available on: `http://recherche.ircam.fr/equipes/analyse-synthese/obin`

sentence	L' ainé n'avait que dix ans, et le plus jeune n'en avait que sept. *The eldest was only ten years old, and the youngest was seven.*
M	L' ainé n'avait // que dix ans // et le plus jeune // n' en avait que sept
P	L' ainé n'avait // que dix ans // et le plus jeune // n' en avait que sept
J	L' ainé n'avait / que dix ans // et le plus jeune // **N'EN** avait que sept
S	L' ainé n' avait / que dix ans // et // le plus jeune n' en avait // que sept
sentence	Il se leva de bon matin, et alla au bord d'un ruisseau, où il emplit ses poches de petits cailloux blancs, et ensuite revint à la maison. *In the morning he rose very early and went to the edge of a brook.* *There he filled his pockets with little white pebbles and came quickly home again.*
M	Il se leva // de **BON** matin // et alla au bord // d' un ruisseau // où il emplit ses poches // de petits cailloux blancs // et ensuite revint à la maison
P	Il se leva de bon matin // et alla au bord / d' un ruisseau // où il emplit ses poches de petits cailloux blancs // et ensuite revint à la maison
J	Il se leva de bon matin / et alla au bord d' un ruisseau // où il emplit ses poches / de petits / cailloux / blancs // et ensuite / revint à la maison
S	Il se leva de bon matin / et alla // au bord d' un ruisseau où il // emplit ses poches de petits cailloux blancs // et ensuite // revint à la maison
sentence	Il leur dit donc, ne craignez point, mes frères : mon Père et ma Mère nous ont laissés ici, mais je vous ramènerai bien au logis, suivez-moi seulement. *"Don't be afraid, brothers", he said presently; "our parents* *have left us here, but I will take you home again. Just follow me."*
M	Il leur dit donc // ne **CRAIGNEZ** point // mes frères // mon Père et ma Mère / nous ont laissés ici // **MZIS** je vous ramènerai bien au logis // suivez -moi seulement
P	Il leur dit donc / ne **CRAIGNEZ** point // mes frères // mon Père // et ma Mère // nous ont laissés ici // mais je vous ramènerai bien au logis // suivez-moi // seulement
J	Il leur dit donc ne **CRAIGNEZ** point / mes frères // mon Père et ma Mère nous ont laissés ici // mais je vous ramènerai bien au logis **SUIVEZ-MOI** seulement
S	Il leur dit donc / ne craignez point // mes frères // mon **PERE** // **et** ma **MERE** nous ont // laissés ici // mais je vous ramènerai bien au logis // suivez -moi seulement
sentence	Qu'à la vérité, il n'avait pas fait conscience de lui prendre ses bottes de sept lieues, parce qu'il ne s'en servait que pour courir après les petits enfants. *[...] Little Tom Thumb [...] only took the seven-league boots, about which he had no compunction,* *since they were only used by the ogre for catching little children.*
M	Qu' à la vérité // il n' avait pas fait conscience / de lui prendre // ses bottes de sept lieues // **PAR**ce qu' il ne s' en **SER**vait // que pour courir / après les petits enfants
P	Qu' à la vérité / il n' avait pas fait conscience // de lui prendre ses bottes de sept lieues // **PAR**ce **QU'IL** ne s' en servait // que pour courir / après les petits enfants
J	Qu' à la vérité / il n' avait pas fait conscience / de lui **PRENDRE** ses bottes de sept lieues // **PARCE** qu' il ne s' en servait / que pour courir / après les **PETITS** enfants
S	Qu' à la vérité // il n' avait pas / fait conscience de lui **PRENDRE** // ses bottes // de sept lieues // parce qu' il ne s' en servait // **QUE** pour courir / après les petits enfants

Table 11.3: Study cases of the average symbolic modelling of speech prosody. // denotes a major prosodic boundary, / a minor prosodic boundary, and bold font a prosodic prominence.

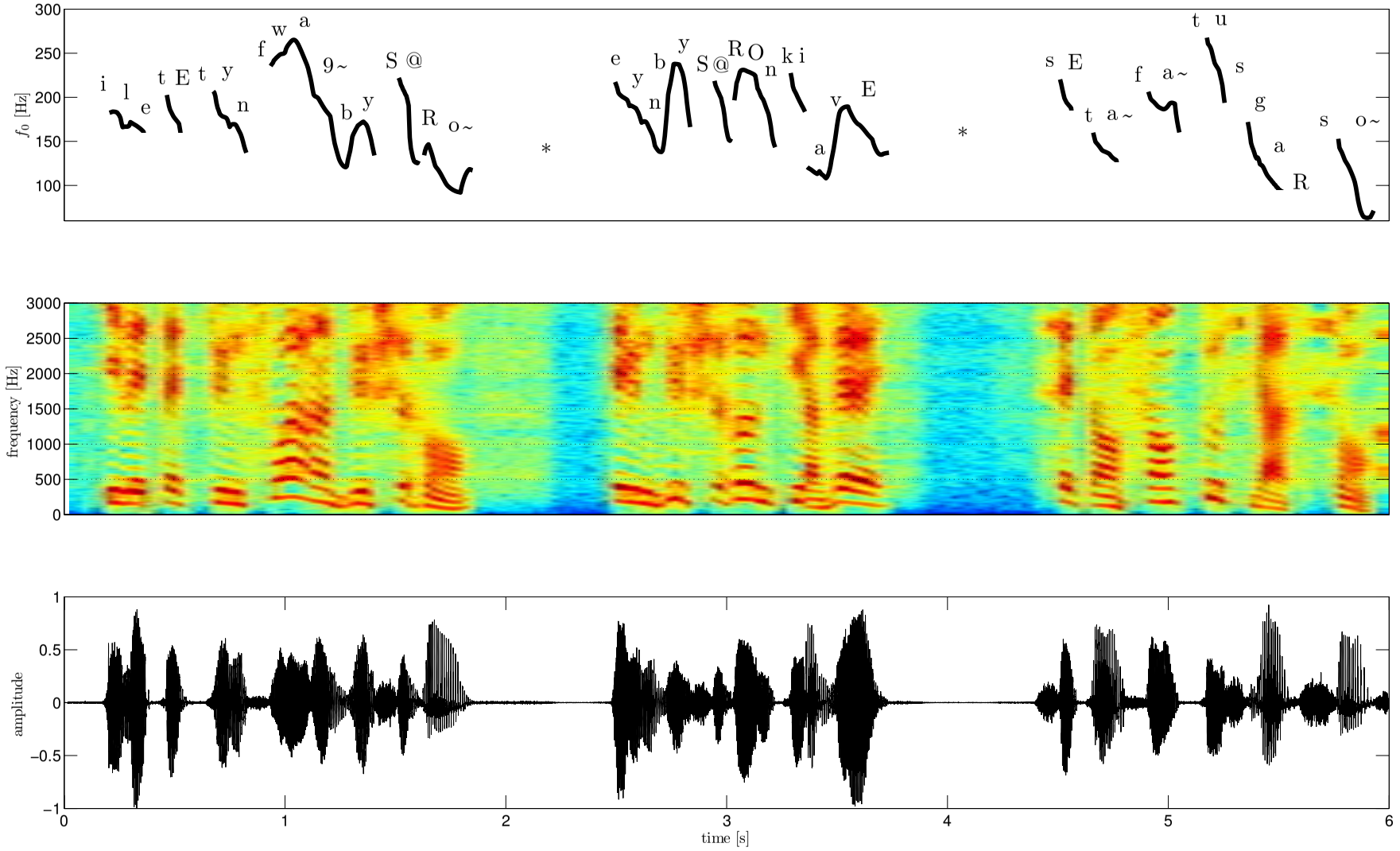


Figure 11.3: Synthesized prosodic contours of the *mass* speaking style for the utterance "*Il était une fois un bûcheron et une bûcheronne qui avaient sept enfants tous garçons.*" ("*Once upon a time there lived a woodcutter and his wife, who had seven children, all boys.*"). On top, melodic contour and phonemic labels. On middle, spectral variations. On bottom, speech waveform.

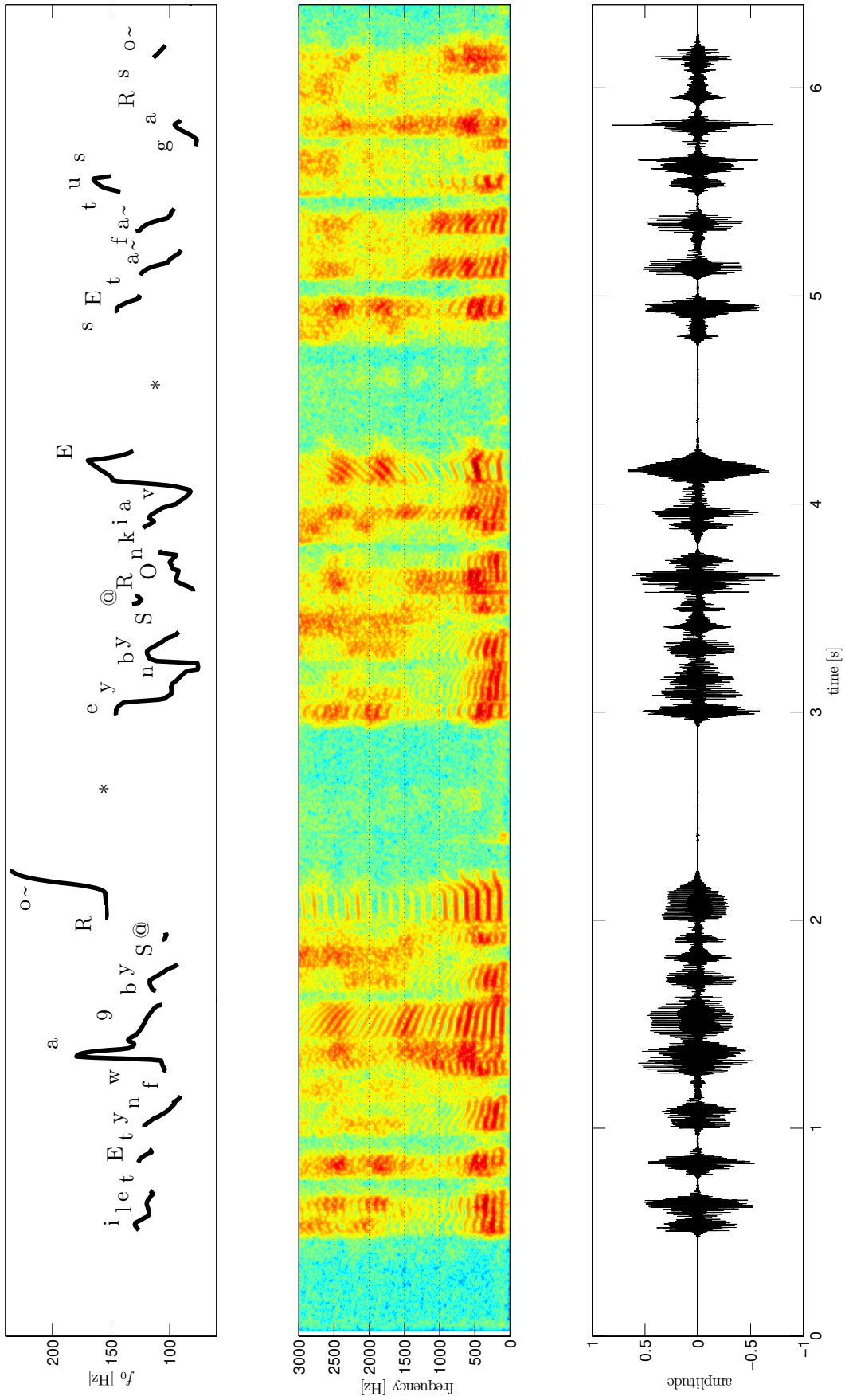


Figure 11.4: Synthesized prosodic contours of the *political* speaking style for the utterance "*Il était une fois un bûcheron et une bûcheronne qui avaient sept enfants tous garçons.*" ("*Once upon a time there lived a woodcutter and his wife, who had seven children, all boys.*"). On top, melodic contour and phonemic labels. On middle, spectral variations. On bottom, speech waveform.

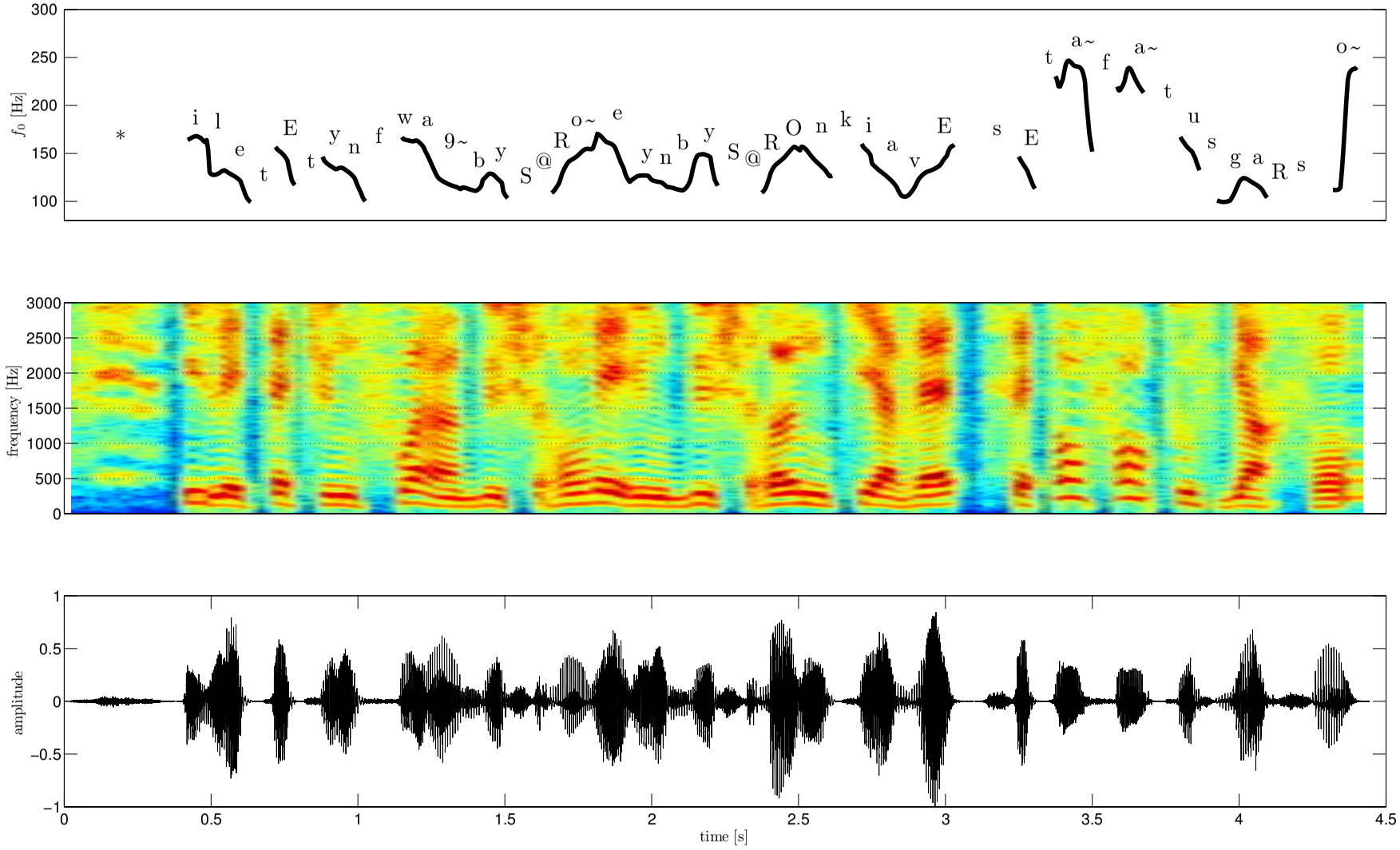


Figure 11.5: Synthesized prosodic contours of the *journalistic* speaking style for the utterance "*Il était une fois un bûcheron et une bûcheronne qui avaient sept enfants tous garçons.*" ("*Once upon a time there lived a woodcutter and his wife, who had seven children, all boys.*"). On top, melodic contour and phonemic labels. On middle, spectral variations. On bottom, speech waveform.

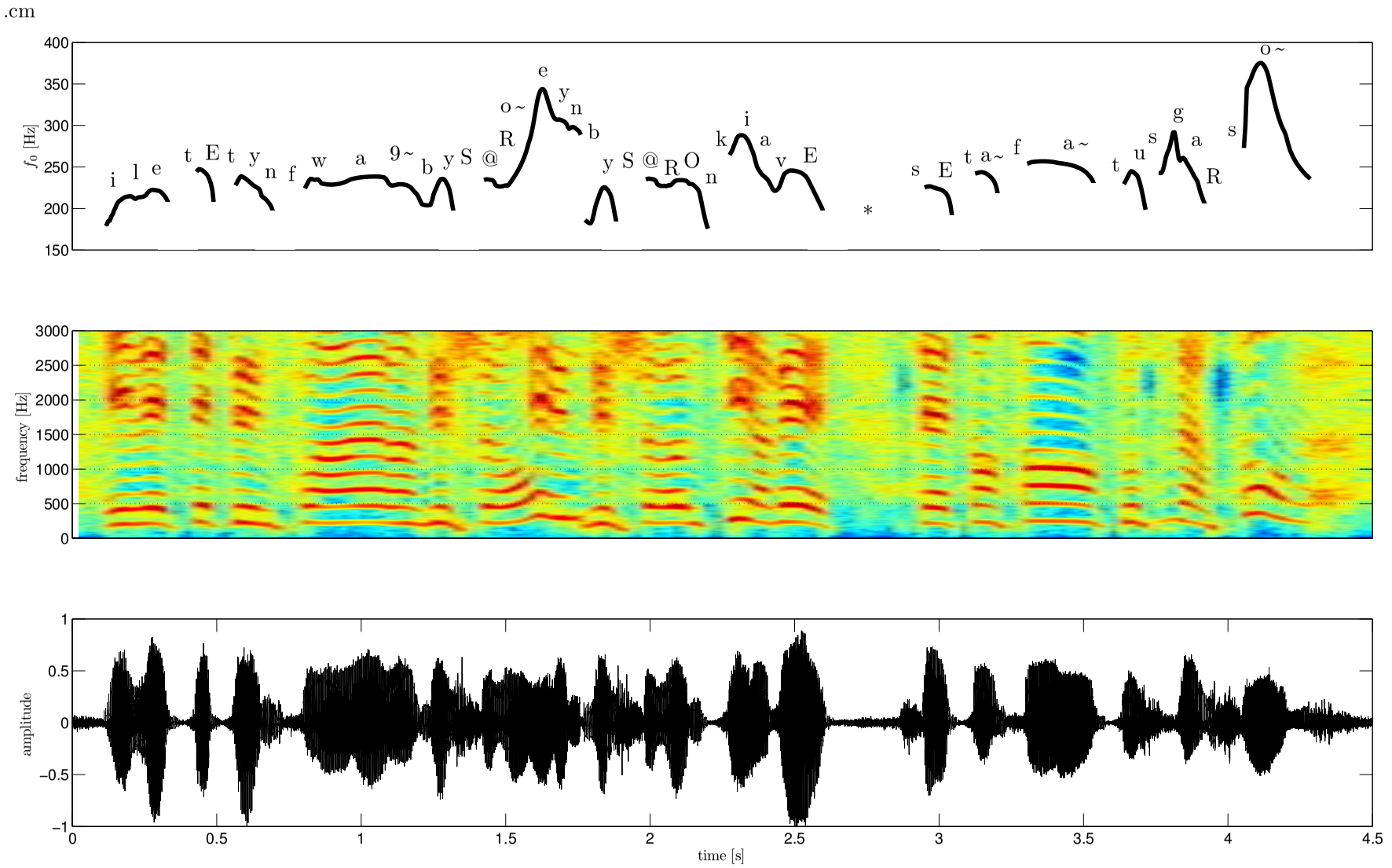


Figure 11.6: Synthesized prosodic contours of the *sport-commentary* speaking style for the utterance "*Il était une fois un bûcheron et une bûcheronne qui avaient sept enfants tous garçons.*" ("*Once upon a time there lived a woodcutter and his wife, who had seven children, all boys.*"). On top, melodic contour and phonemic labels. On middle, spectral variations. On bottom, speech waveform.

Chapter 12

Shared Modelling of Speaking Style

Contents

12.1 Introduction

A discourse genre (DG) is associated with a speaking style that is shared among listeners and speakers of a language. In particular, the speaking style that is shared relates to a set of conventions associated with the specific situation of speech communication, i.e. a set of prosodic variations that does not depend on a specific speaker. However, this speaking style varies across speakers depending on individual and temporal variations. This includes the individual speaking style of a speaker (physiological, idiolectal, geographical, and sociological variations), his ability to adapt his speaking style, and the variations of a speaking style over time. Consequently, modelling a speaking style is to estimate the speech prosody characteristics that are shared among speakers regardless of those particular of a speaker. Such a model is referred as a speaker-independent speaking-style model.

However, there is no known method to distinguish precisely the variations that are specific to a speaker (speaker-dependent) from those that are shared among speakers (speaker-independent). Nevertheless, methods have been proposed to estimate speaker-independent models based on *shared context-dependent models* [Yamagishi, 2006, Yamagishi, 2007] and *speaker normalization* [Anastasakos et al., 1997, Yamagishi, 2006] in the context of HMMs. In particular, speaker-independent modelling of speaking style has been introduced for emotional speech synthesis and adaptation [Yamagishi, 2007]. The shared modelling techniques present substantial advantages over conventional average modelling. Firstly, a shared context-dependent model is derived so as to equally account for the individual contribution of each speaker. Secondly, the parameters of the speaker-independent model are adequately estimated based on the normalization of the variations across speakers.

In this chapter, a speaker-independent speaking-style model is proposed to model the speaking style associated with various situations of speech communication. The proposed approach is the extension of the stylization/trajectory model presented in chapter 9 to the shared context-dependent HMM modelling and speaker normalization.The syllable is used as the minimal temporal domain for the description of speech prosody, and f_0 variations are stylized and modelled simultaneously over various temporal domains which cover short-term and long-term variations. During the training, a speaker-independent speaking-style symbolic model is estimated based on conventional average modelling previously described, and a speaker-independent speaking-style acoustic model is estimated based on shared context-clustering and speaker adaptive training. During the synthesis, the text is first converted into a concatenated sequence of context-dependent models with respect to the speaking style. Firstly, the sequence of prosodic events is determined conditionally to the symbolic model. Then, the acoustic sequence is determined conditionally to the sequence of prosodic events and the acoustic model. Thus, the inferred speaking-style is used to adapt the speaking style of a speaker to be synthesized. The proposed method is evaluated according to the identification of the speaking style that corresponds to the adapted speech synthesis of "*neutral*" text sentences.

12.2 Speaker-Independent Modelling of Speaking-Style

12.2.1 Shared Decision-Tree-Based Context-Clustering

The conventional average context-clustering suffers from inconsistency in the case of speaker-independent modelling [Yamagishi, 2006]. In particular, the average context-clustering does not ensure that the individual characteristics are equally considered to cluster the observed contexts. Thus, a illness-conditioned context-dependent model may be estimated in which the characteristics of a set of speakers are emphasized over others depending on the context. This may result into discontinuities in the inferred acoustic variations depending on the sequence of linguistic contexts.

A solution consists of deriving a speaker-independent context-dependent model from speaker-dependent models [Yamagishi, 2007]. Context-dependent models are estimated for each speaker to account for the individual contribution of each speaker, then merged to derive a context-dependent model that is shared among speakers. The speaker-independent context-clustering method is referred as the *shared context-clustering*.

In the shared-decision-tree context-clustering, a speaker-independent context-tree that is shared by all of the speakers is derived based on the Maximum-Likelihood Minimum-Description-Length. The derivation of the shared-decision-tree is a reformulation of the conventional decision-tree-based context-clustering (section 9.2.3) in the case of the speaker-independent modelling.

Let R be the number of speakers from which a speaker-independent model is to be estimated. Let T be a binary tree with root node S_0 and leaf nodes $\mathbf{S} = (S_1, \ldots, S_M)$. Let $\boldsymbol{\lambda}_{\boldsymbol{S}} = (\boldsymbol{\lambda}_{S_1}, \ldots, \boldsymbol{\lambda}_{S_M})$ the speaker-independent model associated to the set of leaf nodes S, and $\boldsymbol{\lambda}_S^{(r)} = (\boldsymbol{\lambda}_{S_1}^{(r)}, \ldots, \boldsymbol{\lambda}_{S_M}^{(r)})$ the speaker-dependent model associated with speaker r and the set of leaf nodes S. Thus,

$\boldsymbol{\lambda}_{S_m} = (\boldsymbol{\lambda}_{S_m}^{(1)}, \ldots, \boldsymbol{\lambda}_{S_m}^{(R)})$ for each node S_m, $m, \in [1, M]$.

Thus, the change of log-likelihood$\Delta\ {}_{L}^{q}(S')$ and description length$\Delta\ {}_{MDL}^{q}(S')$ by splitting leaf node S_m through context q into nodes $S_{m,q+}$ and $S_{m,q-}$ are simply reformulated in the case of speaker-independent modelling, and the optimal question $\widehat{q}_{MDL}$ is selected so as to minimize the speaker-independent change in description length.

Let $\mathrm{L}(S_m)$ denote the log-likelihood of the speaker-independent model $\boldsymbol{\lambda}_{S_m}$ given the observation sequences $\mathbf{o}_m$ associated with the contexts corresponding to the node S_m, and $\mathrm{L}(S_m^{(r)})$ the log-likelihood of speaker-dependent model $\boldsymbol{\lambda}_{S_m}^{(r)}$ given the observation sequences $\mathbf{o}_m^{(r)}$ associated with the context corresponding to the node S_m.

The log-likelihood of the speaker-independent model $\boldsymbol{\lambda}_S$ is given by the sum of the log-likelihood of the speaker-dependent models $\boldsymbol{\lambda}_S^{(r)}$:

$$L(S) = \sum_{r=1}^{R} L^{(r)}(S) \tag{12.1}$$

$$= \sum_{r=1}^{R} \sum_{m=1}^{M} L^{(r)}(S_m) \tag{12.2}$$

The description length of the speaker-independent model $\boldsymbol{\lambda_S}$ is given by the sum of the description length of the speaker-dependent models $\boldsymbol{\lambda}_S^{(r)}$::

$$DL(S) = \sum_{r=1}^{R} DL^{(r)}(S) \tag{12.3}$$

$$= \sum_{r=1}^{R} \Big(-L^{(r)}(S) + DM \log \Gamma^{\,(r)}(S) + \log I^{(r)} \Big) \tag{12.4}$$

where$\Gamma^{\ (r)} = \sum_{m=1}^{M} \Gamma_m^{(r)}$ is the speaker-dependent total state occupancy probability, $\Gamma_m^{(r)} = \sum_{t=1}^{T} \gamma_t^{(r)}(m)$ is the speaker-dependent total state occupancy probability at node S_m, $\gamma_t^{(r)}(m)$ is the speaker-dependent state occupancy probability at node S_m, D is the dimensionality of the observation feature vector, and I is the number of possible models.

The increase in the speaker-independent model log-likelihood $L(S')$ by splitting leaf node S_m through question q into nodes $S_{m,q+}$ and $S_{m,q-}$ is given by the sum of the increase in the speaker-dependent models log-likelihood $L^{(r)}(S')$ by splitting leaf node S_m through question q into nodes $S_{m,q+}$ and $S_{m,q-}$:

$$\begin{aligned}
\Delta_L^q(S') &= \sum_{r=1}^{R} \Delta_L^{q\,(r)}(S') \\
&= \sum_{r=1}^{R} \Big(L^{(r)}(S_{m,q+}) + L^{(r)}(S_{m,q-}) - L^{(r)}(S_m) \Big) \\
&= -\frac{1}{2} \sum_{r=1}^{R} \Big(\Gamma^{(r)}(S_{m,q+}) \log |\Sigma^{(r)}(S_{m,q+})| + \Gamma^{(r)}(S_{m,q-}) \log |\Sigma^{(r)}(S_{m,q-})| - \Gamma^{(r)}(S_{m,q}) \log |\Sigma^{(r)}(S_{m,q})| \Big)
\end{aligned}$$

where$\Gamma^{\ (r)}(.)$ and$\Sigma^{\ (r)}(.)$ denote the total state occupancy probability and the covariance matrix in speaker-dependent tree node, respectively.

The change in speaker-independent model description length $DL(S')$ by splitting leaf node S_m through question q into nodes $S_{m,q+}$ and $S_{m,q-}$ is given by the sum change in speaker-dependent

model description length $DL^{(r)}(S')$ by splitting leaf node S_m through question q into nodes $S_{m,q+}$ and $S_{m,q-}$:

$$\begin{aligned}\Delta_{DL}^{q}(S') &= \sum_{r=1}^{R} \Delta_{DL}^{q}{}^{(r)}(S') \\ &= \sum_{r=1}^{R} \left(-\Delta_{L}^{q\,(r)}(S') + DM(\log\Gamma^{\,(r)}(S_0))\right) \end{aligned} \tag{12.6}$$

where$\Gamma^{\,(r)}(.)$ denotes the total state occupancy probability in speaker-dependent tree node, and D the dimensionality of the observation feature.

The question $\widehat{q}_{MDL}$ which minimizes the increase of the speaker-independent model description length at node S_m is given by:

$$\widehat{q}_{MDL} = \underset{\mathbf{q}}{\text{argmax}} \; -\Delta_{DL}^{q}(S) \tag{12.7}$$

The shared context-dependent tree is then derived as follows:

1. **tree initialization**

   (a) speaker-independent tree initialization

   $$\begin{aligned} T^{(0)} &= T_0 \\ S^{(0)} &= S_0 \\ \boldsymbol{\lambda}_S^{(0)} &= \boldsymbol{\lambda}_{S_0} \end{aligned}$$

   (b) speaker-dependent trees initialization

   $$\begin{aligned} T^{(r)(0)} &= T_0^{(r)} \\ S^{(r)(0)} &= S_0 \\ \boldsymbol{\lambda}_S^{(0)} &= \boldsymbol{\lambda}_{S_0}^{(r)} \end{aligned}$$

2. **tree recursion**

   for each leaf node S_m of the speaker-independent context-tree $T^{(i)}$

   tree selection

   (a) speaker-dependent description length calculation:

   $$\Delta_{DL}^{q}{}^{(r)}(S), q \in [1, Q], r \in [1, R]$$

   (b) speaker-independent description length calculation:

   $$\Delta_{DL}^{q}(S) = \sum_{r=1}^{R} \Delta_{DL}^{q}{}^{(r)}(S), q \in [1, Q]$$

   (c) optimal speaker-independent splitting context:

   $$\widehat{q}_{MDL} = \underset{\mathbf{q}}{\text{argmax}} \; -\Delta_{DL}^{q}(S)$$

   tree derivation

   if$\Delta\ \widehat{q}_{DL}(S) < 0$, split node S_m, speaker-independent model parameters $\boldsymbol{\lambda}_{S_m}$, and speaker-dependent models parameters $\boldsymbol{\lambda}_{S_m}^{(r)}$:

   $$\begin{aligned} S'_m &\leftarrow (S_{m,\hat{q}-}, S_{m,\hat{q}+}) \\ \boldsymbol{\lambda}_{S'_m} &\leftarrow (\boldsymbol{\lambda}_{S_{m,\hat{q}-}}, \boldsymbol{\lambda}_{S_{m,\hat{q}+}}) \\ \boldsymbol{\lambda}_{S'_m}^{(r)} &\leftarrow (\boldsymbol{\lambda}_{S_{m,\hat{q}-}}^{(r)}, \boldsymbol{\lambda}_{S_{m,\hat{q}+}}^{(r)}) \end{aligned}$$

tree update

(a) speaker-independent tree update

$$\begin{aligned} T^{(i+1)} &= T' \\ S^{(i+1)} &= S' \\ \boldsymbol{\lambda}_S^{(i+1)} &= \boldsymbol{\lambda} S' \end{aligned}$$

(b) speaker-dependent trees update

$$\begin{aligned} T^{(r)(i+1)} &= T'^{(r)} \\ S^{(r)(i+1)} &= S' \\ \boldsymbol{\lambda}_S^{(r)(i+1)} &= \boldsymbol{\lambda}_{S'}^{(r)} \end{aligned}$$

3. **tree termination**

$$\begin{aligned} \widehat{T} &= T^{(i)} \\ \widehat{S} &= S^{(i)} \\ \widehat{\boldsymbol{\lambda}}_S &= \boldsymbol{\lambda}_S^{(i)} \end{aligned}$$

12.2.2 Speaker-Independent Modelling of Speaking-Style Based on *Speaker-Adaptive Training* (SAT)

Speaker normalization methods provide an approximation in which the speech characteristics of a given speaker are transformed so as to optimally fit a speaker-independent model. Statistical methods for speaker normalization have been proposed based on *Speaker Adaptive Training* (SAT) [Anastasakos et al., 1997] and *Maximum Likelihood Linear Regression* (MLLR) [Leggetter and Woodland, 1995]. The principle of speaker adaptation is to transform the characteristics of a speaker-independent model so as to optimally fit the characteristics of a specific speaker. The speaker-independent model is generally estimated from a large amount of multi-speaker observations which provide a robust a priori-knowledge about speech characteristics, while the speaker to be adapted is associated with a limited amount of observations. Speaker normalization based on speaker adaptive training is the inverse problem: the speaker-independent model is estimated so as to optimally fit both the speaker-independent model and the speaker-adapted models. The speaker-independent model remains hidden from observation and is to be estimated, while the speaker-dependent models are fully observed. In the context of maximum-likelihood linear-regression, a set of linear transformations is defined, and the speaker-independent model is estimated so as to maximize the likelihood of the speaker-independent model and the speaker-dependent transformations given the speaker-dependent observations. Such an approach has been proposed in the context of speaker-dependent and emotional speaking-style synthesis and adaptation [Yamagishi, 2007]. In this section, the speaking style of various situation of speech communication is modelled based on the stylization/trajectory model presented in chapter 9 extended to speaker-independent speaking-style modelling.

In the following, a speaker-independent speaking-style model based on speaker-normalization and maximum-likelihood-linear-regression is presented in the case of HMM in which each state is modelled with a single continuous multivariate normal distribution and diagonal covariance matrix. The generalization to a mixture of continuous normal distributions is straightforward.

12.2.2.1 *Maximum-Likelihood Linear Regression* (MLLR)

Let R be the number of speakers from which a speaker-independent model is to be estimated. Let $\mathbf{o} = (\mathbf{o}^{(1)}, \ldots, \mathbf{o}^{(R)})$ the total set of observations, and $\mathbf{o}^{(r)} = [\mathbf{o}^{(r)}(1), \ldots, \mathbf{o}^{(r)}(T_r)]$ is the observation sequences associated with speaker r, where $\mathbf{o}^{(r)}(t) = [o_t^{(r)}(1), \ldots, o_t^{(r)}(D)]^\top$ is the (Dx1) observation vector which describes the acoustical property at time t. Each speaker is modelled by a context-dependent HMM $\boldsymbol{\lambda}^{(r)} = (\boldsymbol{\Pi}^{(r)}, \mathbf{A}^{(r)}, \mathbf{B}^{(r)})$ where $b^{(r)}$ is the state output

probability distribution with (Dx1) mean vector $\boldsymbol{\mu}^{(r)}$ and (DxD) covariance matrix $\boldsymbol{\Sigma}^{(r)}$.

In the maximum-likelihood-linear-regression approach, the difference between each speaker-dependent model and the speaker-independent model is expressed as a linear regression function of the mean vectors of state output probability distributions:

$$\boldsymbol{\mu}^{(r)} = \mathbf{P}^{(r)}\boldsymbol{\mu} + \mathbf{Q}^{(r)} \tag{12.8}$$

where $\boldsymbol{\mu}$ is the speaker-independent mean vector, and $\boldsymbol{\mu}^{(r)}$ are the speaker-dependent mean vectors associated with speaker r.

This transformation can be factorized in a matrix form:

$$\boldsymbol{\mu}^{(r)} = \mathbf{W}^{(r)}\boldsymbol{\xi}^{(r)} \tag{12.9}$$

where $\mathbf{W}^{(r)}$ denotes the ((D+1)x(D+1)) transformation matrix associated with speaker r, and $\boldsymbol{\xi}^{(r)}$ the (D+1) augmented mean vector associated with speaker r.

$$\boldsymbol{\xi}^{(r)} = [\omega, \mu_1, \ldots, \mu_D]^\top \tag{12.10}$$

where ω represents the translation term in the regression.

12.2.2.2 Estimation of Model Parameters

The speaker-independent HMM model $\widehat{\boldsymbol{\lambda}}$ and the set of speaker transformations are jointly estimated to that which maximize the probability of the adapted observation sequence $\mathbf{o}$ given the model $\boldsymbol{\lambda}$ and the set of transformations $\mathbf{W}$.

$$\begin{aligned}(\widehat{\boldsymbol{\lambda}}, \widehat{\mathbf{W}}) &= \underset{\boldsymbol{\lambda},\mathbf{W}}{\text{argmax}}\ \mathrm{p}(\mathbf{o}|\boldsymbol{\lambda}, \mathbf{W}) \\ &= \underset{\boldsymbol{\lambda},\mathbf{W}}{\text{argmax}} \prod_{r=1}^{R} \mathrm{p}(\mathbf{o}^{(r)}|\boldsymbol{\lambda}, \mathbf{W}^{(r)})\end{aligned} \tag{12.11}$$

The estimation of speaker-independent model parameters $\boldsymbol{\lambda}$ and speaker-dependent transformation matrices $\mathbf{W}$ is achieved based on a cascade of Baum-Welch re-estimation procedures.

The SAT auxiliary function Q_{SAT} is defined as:

$$Q_{SAT}(\boldsymbol{\lambda}, \mathbf{W}; \boldsymbol{\lambda}', \mathbf{W}') = \sum_{\mathbf{q}} \mathrm{p}(\mathbf{o}, \mathbf{q}|\boldsymbol{\lambda}, \mathbf{W}) \log(\mathrm{p}(\mathbf{o}, \mathbf{q}|\boldsymbol{\lambda}', \mathbf{W}')) \tag{12.12}$$

The SAT model parameters which maximize the auxiliary function increase the value of the objective function. The SAT model parameters are re-estimated until convergence to a local maximum of the auxiliary function Q_{SAT}.

Speaker-independent state transition probabilities $\mathbf{A}$ are estimated based on the standard expectation-maximization formula. Speaker-dependent transformations $\mathbf{W}$, speaker-independent mean vector $\boldsymbol{\mu}$, and speaker-independent covariance matrix $\boldsymbol{\Sigma}$ are iteratively estimated. For each set of the parameters, the current parameters are estimated while the others are held constant. Then, the speaker-dependent transformation matrices are re-estimated given the current speaker-independent mean vector and covariance matrix, the speaker-independent mean vector is re-estimated given the re-estimated speaker-independent transformation matrices and the current speaker-independent covariance matrix, and the speaker-independent covariance matrix is re-estimated given the re-estimated speaker-dependent transformation matrices and the re-estimated speaker-dependent mean vector.

Firstly, the speaker-dependent transformation matrix is determined to that which maximize the probability of the speaker-dependent observation sequence $\mathbf{o}^{(r)}$ given the current model $\boldsymbol{\lambda}$ and the

speaker-dependent transformation matrix $\mathbf{W}^{(r)}$.

$$\widehat{\mathbf{W}}^{(r)} = \underset{\mathbf{W}^{(r)}}{\text{argmax}}\, \mathrm{p}(\mathbf{o}^{(r)}|\boldsymbol{\lambda}, \mathbf{W}^{(r)}) \tag{12.13}$$

The estimation of the optimal regression matrices $\mathbf{W}^{(r)}$ is achieved using Baum-Welch re-estimation procedure.
The MLLR auxiliary function Q_{MLLR} is defined as:

$$Q_{MLLR}(\mathbf{W}^{(r)}, \mathbf{W}'^{(r)}) = \sum_{\mathbf{q}} \mathrm{p}(\mathbf{o}, \mathbf{q}|\boldsymbol{\lambda}, \mathbf{W}^{(r)}) \log(\mathrm{p}(\mathbf{o}, \mathbf{q}|\boldsymbol{\lambda}, \mathbf{W}'^{(r)})) \tag{12.14}$$

The MLLR model parameters which maximize the auxiliary function increase the value of the objective function. The MLLR model parameters are re-estimated until convergence to a local maximum of the auxiliary function Q_{MLLR}.

The state output probability of the speaker-dependent observation vector $\mathbf{o}^{(r)}$ given the speaker-independent model $\boldsymbol{\lambda}$ and the speaker-dependent transformation matrix $\mathbf{W}^{(r)}$ is given by:

$$b^{(r)}(\mathbf{o}^{(r)}|\boldsymbol{\lambda}, \mathbf{W}^{(r)}) = \mathcal{N}(\mathbf{o}^{(r)}|\mathbf{W}^{(r)}\boldsymbol{\xi}^{(r)}, \boldsymbol{\Sigma}) \tag{12.15}$$

Thus, the MLLR auxiliary function can be rewritten as:

$$Q_{MLLR}(\mathbf{W}^{(r)}, \mathbf{W}'^{(r)}) = \sum_{\mathbf{q}} \sum_{t=1}^{T_r} \mathrm{p}(\mathbf{o}^{(r)}, \mathbf{q}|\boldsymbol{\lambda}, \mathbf{W}^{(r)}) \log(b_r(\mathbf{o}_t^{(r)}|\boldsymbol{\lambda}, \mathbf{W}^{(r)})) + C \tag{12.16}$$

The optimal solution is achieved by equating the partial derivative of $Q_{MLLR}(\mathbf{W}^{(r)}, \mathbf{W}'^{(r)})$ with respect to $\mathbf{W}^{(r)}$ to zero:

$$\frac{\partial Q_{MLLR}(\mathbf{W}^{(r)}, \mathbf{W}'^{(r)})}{\partial \mathbf{W}^{(r)}} = 0 \tag{12.17}$$

Hence, the re-estimation of the regression matrices $\widehat{\mathbf{W}}^{(r)}$ is given by:

$$\sum_{t=1}^{T_r} \gamma^{(r)}(t) \boldsymbol{\Sigma}^{(r)-1} \mathbf{o}_t^{(r)} \boldsymbol{\xi}_m^{(r)\top} = \sum_{t=1}^{T_r} \gamma^{(r)}(t) \boldsymbol{\Sigma}^{(r)-1} \mathbf{W}^{(r)} \boldsymbol{\xi}_m^{(r)} \boldsymbol{\xi}_m^{(r)\top} \tag{12.18}$$

Assuming that the transformation $\mathbf{W}^{(r)}$ is shared among the M distributions of the context-dependent model,

$$\sum_{t=1}^{T_r} \sum_{m=1}^{M} \gamma_m^{(r)}(t) \boldsymbol{\Sigma}_m^{(r)-1} \mathbf{o}_t^{(r)} \boldsymbol{\xi}_m^{(r)\top} = \sum_{t=1}^{T_r} \sum_{m=1}^{M} \gamma_m^{(r)}(t) \boldsymbol{\Sigma}_m^{(r)-1} \mathbf{W}^{(r)} \boldsymbol{\xi}_m^{(r)} \boldsymbol{\xi}_m^{(r)\top} \tag{12.19}$$

Then,

$$\mathbf{Z}^{(r)} = \sum_{m=1}^{M} V_m^{(r)} \mathbf{W}^{(r)} D_m^{(r)} \tag{12.20}$$

where:

$$\mathbf{Z}^{(r)} = \sum_{t=1}^{T_r} \sum_{m=1}^{M} \gamma_m^{(r)}(t) \boldsymbol{\Sigma}_m^{(r)-1} \mathbf{o}_t^{(r)} \boldsymbol{\xi}_m^{(r)\top} \tag{12.21}$$

$$V_m^{(r)} = \sum_{t=1}^{T_r} \gamma_m^{(r)}(t) \boldsymbol{\Sigma}_m^{(r)-1} \tag{12.22}$$

$$D_m^{(r)} = \boldsymbol{\xi}_m^{(r)} \boldsymbol{\xi}_m^{(r)\top} \tag{12.23}$$

Then, the re-estimation of the regression matrices $\widehat{\mathbf{W}}^{(r)}$ is given by:

$$\widehat{\mathbf{w}}_i^{(r)} = G_i^{-1}\mathbf{z}_i \tag{12.24}$$

where

$$G_i^{(r)} = \sum_{m=1}^{M} v_{i,i}^{(r)} D_m^{(r)} \tag{12.25}$$

and $\mathbf{w}_i$ and $\mathbf{z}_i$ denotes the i-th column of the $\mathbf{W}^{(r)}$ and $\mathbf{Z}^{(r)}$ matrices.

Secondly, the re-estimation of the speaker-independent mean vector $\boldsymbol{\mu}$ conditional to the set of speaker-dependent transformations $\widehat{\mathbf{W}}^{(r)}$ is given by:

$$\begin{aligned}\widehat{\boldsymbol{\mu}}_m = \quad & \left[\sum_{r=1}^{R}\sum_{t=1}^{T_r}\gamma_m^{(r)}(t)\widehat{\mathbf{A}}^{(r)\top}\boldsymbol{\Sigma}_m^{(r)-1}\widehat{\mathbf{A}}^{(r)}\right]^{-1} \times \\ & \left[\sum_{r=1}^{R}\sum_{t=1}^{T_r}\gamma_m^{(r)}(t)\widehat{\mathbf{A}}^{(r)\top}\boldsymbol{\Sigma}_m^{(r)-1}\left(\mathbf{o}_t^{(r)} - \mathbf{B}^{(r)}\right)\right]\end{aligned} \tag{12.26}$$

Thirdly, the re-estimation of the covariance matrix conditionnal to the speaker-dependent linear transformation is given by:

$$\widehat{\boldsymbol{\Sigma}}_m = \frac{\displaystyle\sum_{r=1}^{R}\sum_{t=1}^{T_r}\gamma_m^{(r)}(t)\left(\mathbf{o}_t^{(r)} - \widehat{\boldsymbol{\mu}}_m^{(r)}\right)\left(\mathbf{o}_t^{(r)} - \widehat{\boldsymbol{\mu}}_m^{(r)}\right)^{\top}}{\displaystyle\sum_{r=1}^{R}\sum_{t=1}^{T_r}\gamma_m^{(r)}(t)} \tag{12.27}$$

The speaker-independent model $\boldsymbol{\lambda}$ and the set of speaker-dependent transformations $\mathbf{W}$ are recursively estimated as follows:

1. **speaker-independent model initialization**

$$\begin{aligned}\boldsymbol{\mu} &= \boldsymbol{\mu_0} \\ \boldsymbol{\Sigma} &= \boldsymbol{\Sigma_0}\end{aligned}$$

2. **tree recursion**

estimation of speaker-dependent transformations
for each speaker r and observations $\mathbf{o}^{(r)}$

(a) update transformation matrix $\mathbf{W}^{(r)}$:

$$\widehat{\mathbf{w}}_i^{(r)} = G_i^{-1}\mathbf{z}_i$$

estimation of speaker-independent model

(a) update mean vector $\boldsymbol{\mu}$:

$$\begin{aligned}\hat{\boldsymbol{\mu}}_m = \quad & \left[\sum_{r=1}^{R}\sum_{t=1}^{T_r}\gamma_m^{(r)}(t)\widehat{\mathbf{A}}^{(r)\top}\boldsymbol{\Sigma}_{\mathbf{m}}^{-\mathbf{1}}\widehat{\mathbf{A}}^{(r)}\right]^{-1} \times \\ & \left[\sum_{r=1}^{R}\sum_{t=1}^{T_r}\gamma_m^{(r)}(t)\widehat{\mathbf{A}}^{(r)\top}\boldsymbol{\Sigma}_{\mathbf{m}}^{-\mathbf{1}}\left(\mathbf{o}_t^{(r)} - \mathbf{B}^{(r)}\right)\right]\end{aligned} \tag{12.28}$$

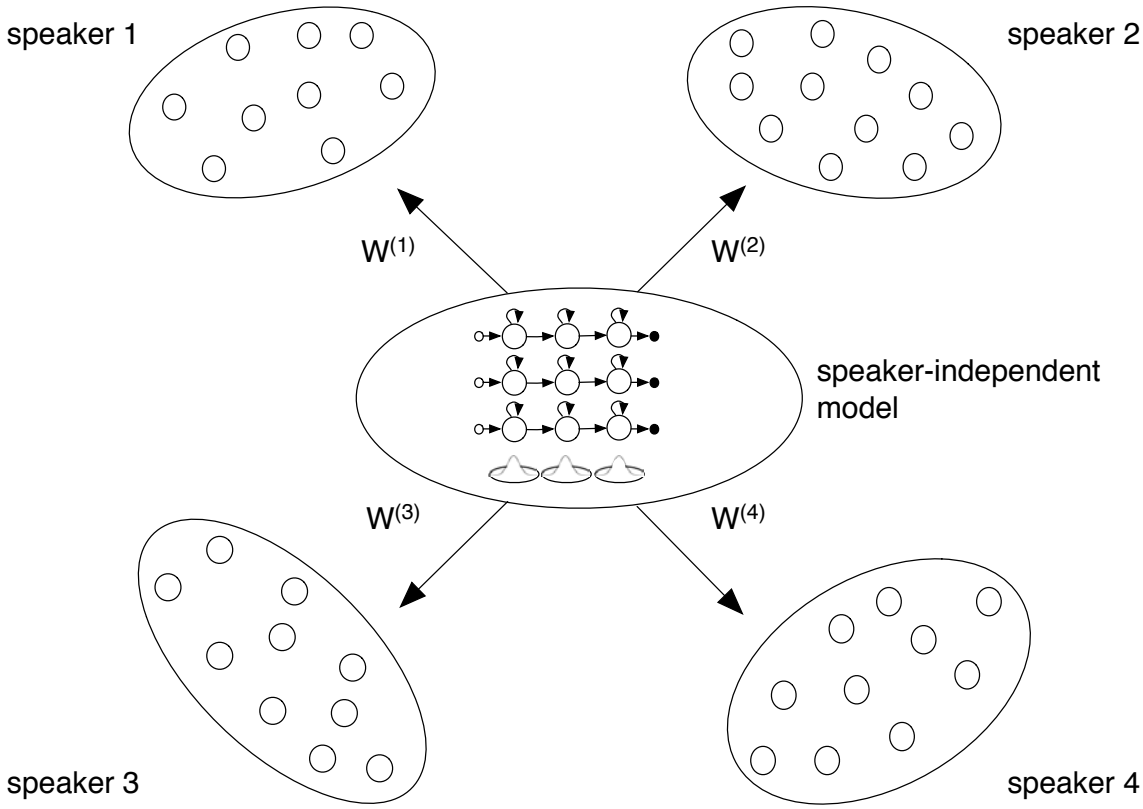


Figure 12.1: Illustration of the speaker-independent model estimation.

(b) update covariance matrix vector $\boldsymbol{\Sigma}$:

$$\widehat{\boldsymbol{\Sigma}}_m = \frac{\sum_{r=1}^{R}\sum_{t=1}^{T_r}\gamma_m^{(r)}(t)\left(\mathbf{o}_t^{(r)} - \hat{\boldsymbol{\mu}}_m^{(r)}\right)\left(\mathbf{o}_t^{(r)} - \hat{\boldsymbol{\mu}}_m^{(r)}\right)^\top}{\sum_{r=1}^{R}\sum_{t=1}^{T_r}\gamma_m^{(r)}(t)}$$

3. **tree termination**

$$\begin{aligned}\widehat{\mathbf{W}} &= \mathrm{W}\\ \widehat{\boldsymbol{\mu}} &= \boldsymbol{\mu}\\ \widehat{\boldsymbol{\Sigma}} &= \boldsymbol{\Sigma}\end{aligned}$$

12.3 Evaluation

In the following, the speaker-independent model is estimated from the speaker-dependent distributions of each of the terminal nodes of the shared decision tree. Additionally, the transformation matrix is assumed to be a full transformation matrix including translations ($\omega = 1$).

12.3.1 Experimental Design

The evaluation consisted of a speaking style identification experiment of the adapted speaking style in speech synthesis[1]. The evaluation material consisted of a set of synthesized speech utterances of a single speaker whose speaking style was adapted to that of a target speaking style. The text to be synthesized was chosen as being "*neutral*", i.e. whose genre and content differ with those of any of the considered DGs. From a set of randomly selected sentences, speech utterances were synthesized using a conventional speech synthesizer. Then, the speech prosody parameters corresponding to a speaking style were synthesized, and used to adapt the speaking style of the synthesized speaker. Finally, the adapted speech utterances were presented in a multiple choice identification experiment to listeners with various language backgrounds using crowd-sourcing.

12.3.2 Stimuli

12.3.2.1 Linguistic Contexts

Linguistic information were extracted from text using the linguistic processing chain described in chapter 7. The symbolic model was trained with the full rich linguistic feature set, which consists in segmental, prosodic, morpho-syntactic, dependency, constituency, and adjunction features. The acoustic model was trained with the full rich linguistic and the prosodic structure feature sets. The used linguistic units were syllable, and the syntactic units. Linguistic features are converted into linguistic contexts over syllable by computing locational and weight contexts, and representing 1-order left-to-right contexts and 1-order child-to-parent contexts in the case of the dependency contexts.

Finally, the linguistic contexts used are defined as:

$$\begin{aligned}\text{symbolic:}\quad & Q_{\text{symbolic}}^{(syllable)} &=\ & Q_{\text{segment}} \cup Q_{\text{morpho}} \cup Q_{\text{dep}} \cup Q_{\text{chunk}} \cup Q_{\text{adj}}\\ \text{acoustic:}\quad & Q_{\text{acoustic}}^{(syllable)} &=\ & Q_{\text{segment}} \cup Q_{\text{morpho}} \cup Q_{\text{dep}} \cup Q_{\text{chunk}} \cup Q_{\text{adj}} \cup Q_{\text{proso}}\end{aligned}$$

[1] Contrary to the previous identification experiments that were based on connotative text content and delexicalized speech, the present experiment is based on "*neutral*" text content, without delexicalization.

12.3.2.2 Training Corpus

Speaker-dependent speech synthesis source/filter models were estimated on 5 hours (1888 utterances) of the *multi-media* speech database using the conventional HMM-based speech synthesis system. Speaker-independent speaking-style models were estimated on the speaking style speech database with respect to the considered speaking style. Additionally, a universal speaking style model was estimated on the complete speaking style speech database.

12.3.2.3 Evaluation Corpus

The evaluation corpus was chosen as being "*neutral*", i.e. whose genre and content differ with any of the considered DGs. The evaluation corpus is composed of sentences extracted from the C-TALE corpus (143 sentences): the fairy-tale "*Le Petit Poucet*" ("*Little Tom Thumb*") by French writter Charles Perrault [Perrault, 1697]. The sentences were processed by the linguistic processing chain without manual correction.

12.3.2.4 Speaking Style models

A speaking style model was estimated for each of the DGs, and a universal model was additionally estimated from the pooled speakers among all DGs[2].

$\boldsymbol{\lambda}^{(M)}$:	598 utterances,	1h20,	7 speakers
$\boldsymbol{\lambda}^{(P)}$:	454 utterances,	1h10,	5 speakers
$\boldsymbol{\lambda}^{(J)}$:	840 utterances,	1h10,	5 speakers
$\boldsymbol{\lambda}^{(S)}$:	743 utterances,	35mn,	4 speakers
$\boldsymbol{\lambda}^{(N)}$:	2635 utterances,	4h,	21 speakers

For the symbolic modelling, a speaker-independent model $\boldsymbol{\lambda}^{(style)}_{\text{symbolic}}$ is estimated based on the conventional average model, solely (chapter 11). For the acoustic modelling, a speaker-independent model $\boldsymbol{\lambda}^{(style)}_{\text{acoustic}}$ is estimated based on shared-decision-tree context-clustering and speaker adaptive training (chapter 12).

During the synthesis, the text is first converted into a concatenated sequence of context-dependent HMMs. Firstly, the sequence of prosodic events is determined so as to maximize the probability of the sequence of prosodic events conditionally to the linguistic context sequence and the symbolic model $\boldsymbol{\lambda}^{(style)}_{\text{symbolic}}$. Then, the sequence of acoustic variations is determined so as to maximize the conditional probability of the acoustic sequence given the sequence of linguistic contexts, the sequence of prosodic events, and the acoustic model $\boldsymbol{\lambda}^{(style)}_{\text{acoustic}}$.

In parallel, the speech utterance is synthesized with respect to the speaker-dependent model, conditionally to the sequence of linguistic contexts and the sequence of prosodic events.

Then, the inferred speech prosody is used to adapt the speaking style of the speaker in a same manner as that described in chapter 9 with additional f_0 normalization with respect to that of the speaker. More precisely, the inferred f_0 variations are normalized in the log domain with respect to the speaker f_0 mean:

$$\log f_0^{speaker,style} = \log \bar{f}_0{}^{speaker} + \Delta \log f_0^{style} \tag{12.29}$$

Each sentence was synthesized and adapted according to the speaking styles considered and the universal speaking style. Then, 6 utterances were randomly selected for each DG. This finally results into 6x(4+1) = 30 adapted speech utterances to be identified.

[2]the universal model is referred as "N", since the universal model is assumed to be neutral.

stream	source/filter	duration	f_0
corpus			
training corpus	non-professionnal speaker (1h)		
evaluation corpus	-		
feature extraction			
feature	5-order aperiodicity 39-order MFCC	state-duration	f_0
window	50-ms blackmann		
frame rate	5ms		
feature transform			
transform	-	log	log
dynamic	1-orderΔ ,Δ ^{2}	-	1-orderΔ ,Δ ^{2}
model			
topology	5-state HMM normal distribution semi-tied covariance	5-state HMM normal distribution	5-state MSD-HMM normal distribution semi-tied covariance
context	$Q^{(phone)}_{acoustic}$		
clustering	DT ML-MDL		

Table 12.1: Evaluation of the *Speaker-Independent Speaking-Style Model*: speech synthesizer speaker-dependent model setup

dimension	**structure**	**acoustic**	
stream	prosodic structure	duration	f_0
corpus			
training corpus evaluation corpus	C-STYLE (4h, 1h per DG) C-TALE (30 utterances, 6 per DG + neutral)		
feature extraction			
feature	hierarchical prosodic structure F_M, F_m, P	duration	f_0
window	syllable	syllable	50-ms hanning
frame rate	syllable	syllable	5ms
feature transform			
transform function	linearization	log	log + 5-order DCT
transform unit	syllable	syllable	syllable Δ = 1-order context
model			
topology	discrete HMM ergodic	single state HMM single normal distribution diagonal covariance	single state HMM single normal distribution diagonal covariance
context	$Q^{(syllable)}_{\text{symbolic}}$	$Q^{(syllable)}_{\text{acoustic}}$	$Q^{(syllable)}_{\text{acoustic}}$
clustering	average context-clustering DT CART	shared context-clustering DT ML-MDL MLLR adaptation	shared context-clustering DT ML-MDL MLLR adaptation

Table 12.2: Evaluation of the *Speaker-Independent Speaking-Style Model*: speaking style model setup

12.3.3 Participants

47 subjects participated in this evaluation. This includes: 23 native French speakers, 15 non-native French speakers, 9 non-French speakers; 33 expert and 13 naïve participants. *Expert* participants were coming from various domains: 10 from speech synthesis, 9 from speech and audio technologies, 6 musicians, 5 from linguistics, and 3 non-specified. 2 participants were removed for the analysis because they did not process the experiment entirely.

12.3.4 Procedure

The experiment consisted of a multiple choice identification task based on the perception of speaking style[3]. The experiment was conducted according to a source-crowding technique using web social networks[4].

Firstly, participants were given a brief description of the 4 DGs augmented with a neutral one that corresponds to a neutral speaking style.

P	political	(TV new year's speech)
J	journalistic	(radio review)
S	sports commentary	(soccer)
M	mass	(Christian sermon)
N	neutral	(-)

In a preliminary experiment, participants were asked to identify the DG associated with *real* speech utterances that were extracted from the speaking style speech database (2 speech utterances per DG, each with a different speaker). This preliminary experiment aims at presenting the speaking style that can be expected for each DG, and controlling the identification ability of the participants.

In the main experiment, participants were asked to identify the speaking style associated with *synthesized* and *adapted* speech utterances (6 speech utterances per DG, and 6 neutral speech utterances). Firstly, 4 synthesized speech utterances of the speaker used for the speech synthesis were presented to the participants to familiarize with the speaking style of the speaker. Then, the adapted speech utterances were randomly presented, and participants were asked to associate each with a speaking style. For each speech utterance, participants were given two options:

total confidence : select only one DG when certain of the choice;

confusion : select two different DGs when a confusion between two likely DGs exists;

The experiment was conducted in a similar manner as those presented in chapters 10 and 11.

12.4 Results

Identification performance was estimated using the measure based on Cohen's Kappa statistic that was presented in chapter 10. *Confusion* ratings were considered as equally possible ratings. Table 12.2 presents the speaking style confusion matrix.

[3] the experiment is available at the following link: `http://recherche.ircam.fr/equipes/analyse-synthese/obin/pmwiki/pmwiki.php/Main/SpeakingStyleAdaptation`

[4] *Ircam Analysis and Synthesis Perceptual Experiments* on Facebook: `http://www.facebook.com/group.php?gid=150354679034&ref=ts`

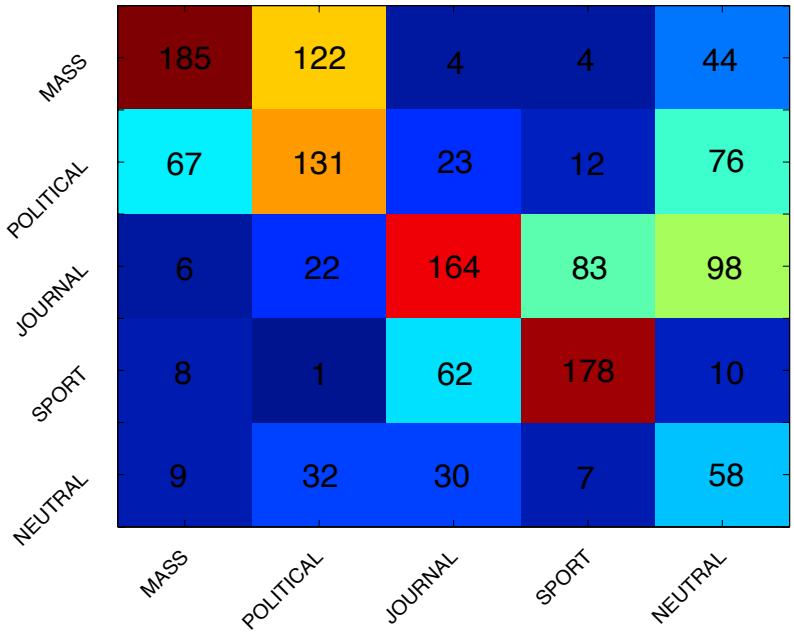


Figure 12.2: Adapted speaking style confusion matrix. Rows represent synthesized speaking style. Columns represent identified speaking style.

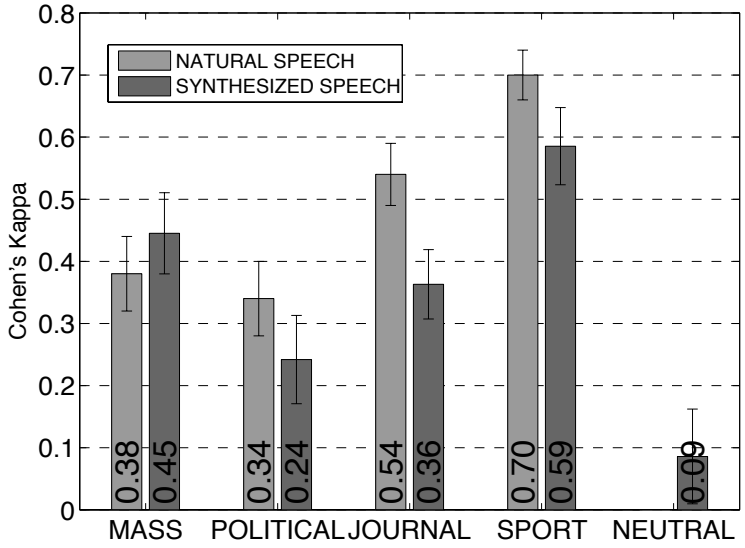


Figure 12.3: Overall identification performance depending on the speaking style (mean identification and 95% confidence interval).

Overall score reveals a fair identification performance ($\kappa_{\text{adaptation}} = 0.41 \pm 0.04$). In comparison with the experiment based on delexicalized speech, this is is comparable to the identification performance observed with natural speech ($\kappa_{\text{natural}} = 0.45 \pm 0.03$) and slightly outperforms that obtained with the average model ($\kappa_{\text{average}} = 0.38 \pm 0.04$). The identification performance significantly depends on the speaking style (figure 12.3): sports commentary shows substantial identification ($\kappa^{(S)}_{\text{adaptation}} = 0.59 \pm 0.07$), mass fair identification ($\kappa^{(M)}_{\text{adaptation}} = 0.45 \pm 0.07$), journal and political speech moderate identification ($\kappa^{(J)}_{\text{adaptation}} = 0.36 \pm 0.05$, $\kappa^{(P)}_{\text{adaptation}} = 0.24 \pm 0.07$). Finally, the neutral speaking style has almost random identification ($\kappa^{(N)}_{\text{adaptation}} = 0.09 \pm 0.07$).

A comparison with the previous experiments reveals substantial differences. Identification of the mass speaking style is significantly higher than that obtained with the average model ($\kappa^{(M)}_{\text{average}} = 0.12 \pm 0.06$) and even outperforms that obtained with natural speech ($\kappa^{(M)}_{\text{natural}} = 0.38 \pm 0.07$). Identification of the political speaking style is lower but not significantly than that obtained with the average model ($\kappa^{(P)}_{\text{average}} = 0.28 \pm 0.07$) and with natural speech ($\kappa^{(P)}_{\text{natural}} = 0.38 \pm 0.06$). Identification of journalistic speaking style significantly drops in performance compared to those obtained with the average model ($\kappa^{(J)}_{\text{average}} = 0.50 \pm 0.06$) and natural speech ($\kappa^{(J)}_{\text{natural}} = 0.54 \pm 0.07$). In particular, the journalistic speaking style is significantly more confused with the sport-commentary and the added neutral speaking styles. Identification of sport-commentary speaking is lower but not significantly to those obtained with the average model ($\kappa^{(S)}_{\text{average}} = 0.68 \pm 0.05$) and with natural speech ($\kappa^{(S)}_{\text{natural}} = 0.70 \pm 0.03$).

Post-hoc analysis conducted from the information provided by the participants (*multi-class one-way analysis of variance* (ANOVA)) confirms evidence for the previously reported *language* and *expertise* factors (figure 12.4).

Analysis reveals a significant effect of the language ($\text{F}(2, 45) = 3.45, \text{p} = 0.04$). In particular, native French speakers performs significantly better than the other participants ($\kappa_{\text{adaptation,native}} = 0.48 \pm 0.05$, $\kappa^{(M)}_{\text{adaptation,native}} = 0.54 \pm 0.07$, $\kappa^{(P)}_{\text{adaptation,native}} = 0.36 \pm 0.07$, $\kappa^{(J)}_{\text{adaptation,native}} = 0.40 \pm 0.09$, $\kappa^{(S)}_{\text{adaptation,native}} = 0.65 \pm 0.08$). Additionally, the language effect varies depending on the speaking style: there is a significant effect for the political and the sport-commentary speaking styles ($\text{F}(2, 45) = 8.12, \text{p} = 0.001$,$\text{F}(2, 45) = 4.05, \text{p} = 0.02$), but none for the mass and the journalistic speaking styles ($\text{F}(2, 45) = 1.96, \text{p} = 0.15$, $\text{F}(2, 45) = 1.64, \text{p} = 0.2$).

Analysis reveals a significant effect of the expertise ($\text{F}(1, 45) = 3.34, \text{p} = 0.007$). In particular, there is a clear significant differences between expert in speech synthesis and other expert participants ($\text{F}(1, 33) = 13.04, \text{p} = 0.001$). The expertise effect varies depending on the speaking style: there is a significant effect for the mass and the political speaking styles ($\text{F}(1, 45) = 7.76, \text{p} = 0.005$, $\text{F}(1, 45) = 3.1, \text{p} = 0.01$), but none for the journalistic and sport-commentary speaking styles ($\text{F}(1, 45) = 0.2, \text{p} = 0.65$, $\text{F}(1, 45) = 0.3, \text{p} = 0.57$).

A comparison with the previous experiments reveals a substantial change in the confusion of speaking styles. Political and mass speaking styles remains strongly confused but are better distinguished than for the average model ($\text{d} = 0.07 \pm 0.1$). The journalistic speaking style is clearly distinguished from the political and the mass speaking styles compared to the average model ($\text{d} = 0.55 \pm 0.07$, and $\text{d} = 0.60 \pm 0.08$), but is strongly confused with the sport-commentary speaking style ($\text{d} = 0.84 \pm 0.06$).

12.5 Discussion

The identification performance obtained in this experiment can not be precisely compared to those obtained with the previous identification experiments, due to their differences in the text used for the synthesis, the availability of linguistic content for the identification, the acoustic parameters used for the modelling, the normalization of the pitch range, and the add of a neutral

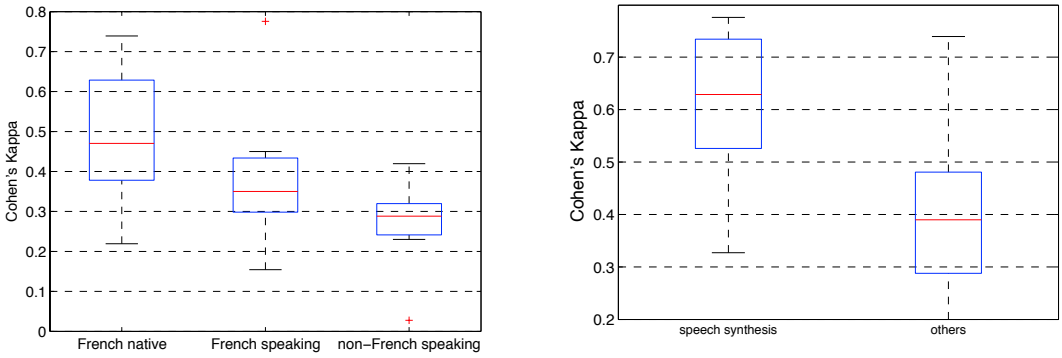


Figure 12.4: Identification performance distributions depending on the listener's language (median, inter-quartiles, standard deviation).

distance	mass	political	journal	sport	neutral
mass	-	0.24 (± 0.02)	0.95 (± 0.09)	0.93 (± 0.09)	0.67 (± 0.06)
political	0.24 (± 0.02)	-	0.75 (± 0.07)	0.91 (± 0.08)	0.24 (± 0.02)
journal	0.95 (± 0.09)	0.75 (± 0.07)	-	0.39 (± 0.04)	0.25 (± 0.02)
sport	0.93 (± 0.09)	0.91 (± 0.08)	0.39 (± 0.04)	-	0.88 (± 0.08)
neutral	0.67 (± 0.06)	0.24 (± 0.02)	0.25 (± 0.02)	0.88 (± 0.08)	-

Table 12.3: DGs distance in the perceptual space (mean distance and 95% confidence interval).

speaking style. The difference in the text used for the synthesis is not absolutely prohibitive for a comparison, but does not permit an exact comparison of the identification performance. The availability of a neutral linguistic content should not change the identification, but the neutral text content actually biased the identification process. The reduction of the acoustic parameters for the modelling of speaking style and the normalization of the pitch range clearly remove significant sources of information that can be used for the identification of a speaking style. Consequently, the identification performance of the adapted speaking style is expected to drop compared to that of the average speaking style. Nevertheless, several instructive conclusions can be pointed out from a comparative analysis.

The identification performance is comparable to that obtained with natural speech, and outperforms that obtained with the average model. However, the absence of a rich description that includes intensity, vocal quality, pitch range, and phonetic strategies may explain the increase in confusion across some of the speaking styles. For instance, the journalistic and the sport-commentary speaking styles could be simply distinguished based on intensity, vocal quality and pitch range: the sport-commentary has generally a high intensity with eventual non-linearities, a pressed voice, and a high pitch range compared to the journalistic speaking style. In a similar manner, the political speaking style has generally a low range and a breathy voice, while the mass speaking style has a high range and a relaxed voice. Thus, the conventional prosodic parameters do not suffice to convey the information associated with a speaking style. Finally, a speaking style appears to be characterized by a rich set of acoustic parameters from global characteristics, local variations of speech prosody (contours), to phonatory strategies (articulation and co-articulation).

The speaking style characteristics that have been modelled can be listed as follows. The mass speaking style is characterized by frequent and long pauses, slow speech rate, high pitch range, terminal high-pitch accents combined with long duration, and specific intonational structure associated with a high-pitch accent followed by a terminal medium-pitch accent with a long duration. The political speaking style is similar to the mass speaking style, with the exception of intermediate and terminal low-pitch accents. The journalistic speaking style is characterized

by a regular prosodic structure, focal prosodic prominences on the first syllable of internal prosodic groups or isolate forms eventually preceded by a short pause at some specific lexical or syntactic locations, fast speech rate, medium pitch range, and high intermediate and terminal pitch accents. Sport-commentary is characterized by an irregular prosodic structure, irregular rhythm, irregular intonational variations, focal prominence on or after specific lexical content and syntactic constructions, fast speech rate, high pitch range, and very high intermediate and terminal pitch accents.

Interestingly, participants report three main global prosodic cues to identify a speaking style: global speech rate, pausing (frequency and duration of pauses), and regularity (in intonation, rhythm, and prosodic structuring). In particular, the more regular was the speech prosody, the more neutral was considered the speaking style. Local prosodic variations (prominence location, prosodic grouping, and prosodic contours) were globally not expressed by participants as a determinant cue used for the identification. Experts in speech synthesis and expert native French participants were able to use local prosodic details to distinguish more accurately the speaking styles. This unfortunately may hide the accuracy of the speaker-independent model to model fine prosodic characteristics. A primary distinction is clearly made with respect to the global speech rate and the pausing. Then, the journalistic and the neutral speaking styles were distinguished from the sport-commentary mainly by the prosodic regularity. Finally, the remaining speaking styles were strongly confused, especially by non-native French and non-French participants who appears to be not able to use local prosodic details. In particular, the high confusion that is observed between the journalistic and the neutral speaking styles is due to the fact that the journalistic speaking style is considered as the more neutral, and thus close to a neutral style.

Finally, the choice of a neutral text for the evaluation is questioned: many French speakers participants have reported a clear difficulty to abstract the text content for the identification. In particular, the inadequacy of the text content and the speaking style was judged as a perturbation for the identification of a speaking style. In the meanwhile, the adequacy of a text content and a speaking style would facilitate the identification process. Thus, the partial lexical adequacy of the fairy tale content with that of a mass discourse may have favoured the identification of the mass speaking style.

12.6 Conclusion

In this chapter, a speaker-independent speaking-style model was proposed to model a speaking style that is shared among a set of speakers[5]. The proposed approach is based on shared context-dependent HMM modelling and speaker normalization that are combined with stylization and trajectory modelling of the acoustic variations over various temporal domains. During the synthesis, the sequence of prosodic events is determined conditionally to the average symbolic model. Then, the sequence of acoustic variations is determined conditionally to the sequence of prosodic events and the acoustic model. Thus, the inferred speaking-style is used to adapt the speaking style of a speaker. The proposed approach was evaluated in a speaking-style identification experiment with synthesized utterances from neutral text sentences. Five speaking-styles were including a neutral speaking style that was defined as the universal model from the pooled speaking styles.

The speaker-independent model succeeds in modelling the speaking-style regardless of a specific speaker. In particular, the model alleviates the discontinuities that are due to a illness-balanced context-dependent model and the predominance of a particular speaker. However, the speaker-independent modelling present some limitations, especially when a small amount of observations is available for each speaker used to estimate the speaker-independent model, or when the number of speakers is large compared to the amount of observations for each speaker. In particular, speaker-dependent models are poorly estimated due to the reduced number of observations

[5]examples of adapted speaking styles are available on: `http://recherche.ircam.fr/equipes/analyse-synthese/obin`

available, and the difference in the linguistic contexts that are observed for each speaker may cause an inadequate derivation of the shared context-dependent model.

The identification performance is comparable to that obtained with natural speech, and outperforms that obtained with the average speaking-style model, even with a reduced set of prosodic parameters and a normalization of the pitch range. However, some of the speaking styles were significantly more confused due to the reduced set of prosodic parameters and the pitch normalization. Thus, an extended set of acoustic dimensions is required to properly characterize a speaking-style, from global characteristics to phonatory strategies. Global prosodic variations were reported as the primary cues to distinguish the speaking styles: global speech rate, pausing (frequency and duration of pauses), and regularity (in intonation, rhythm, and prosodic structuring). Local prosodic variations (prominence location, prosodic grouping, and prosodic contours) were globally not reported as a determining cues for identification. Experts in speech synthesis and expert native French participants were able to use local prosodic details to distinguish the speaking styles more accurately. This unfortunately may mask the accuracy of the speaker-independent model to model fine prosodic variations.

Finally, the choice of a neutral text for evaluation is questioned: the inadequacy of a text content a speaking style tends to impair the identification process, while the adequacy of a text content and a speaking style tends to facilitate the identification process. A solution to manage the text content can be formulated in the selection of text that corresponds to a DG, to present the different adapted speaking style utterances, and to formulate the instructions for identification so as to select the speaking style that is the most appropriate to the text content. This identification scheme will be further evaluated and can be generalized to the identification of a speaking style, associated more accurately with a specific situations of speech communication or emotions.

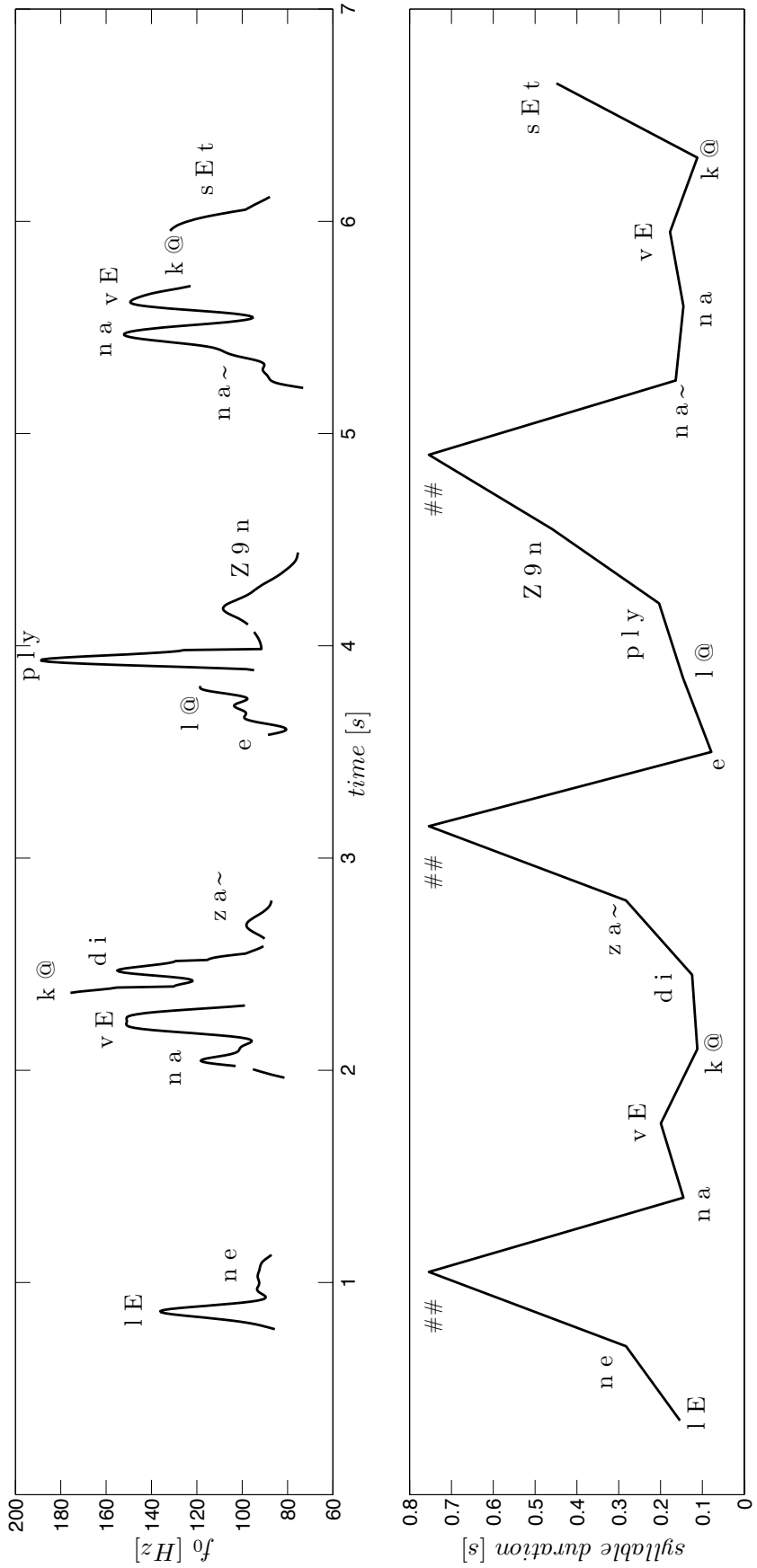


Figure 12.5: Synthesized prosodic contours of the mass speaking style for the utterance "*L'ainé n'avait que dix ans, et le plus jeune n'en avait que sept.*" ("*The eldest was only ten years old, and the youngest was seven.*"). On top, melodic contour. On bottom, rhythmic contour.

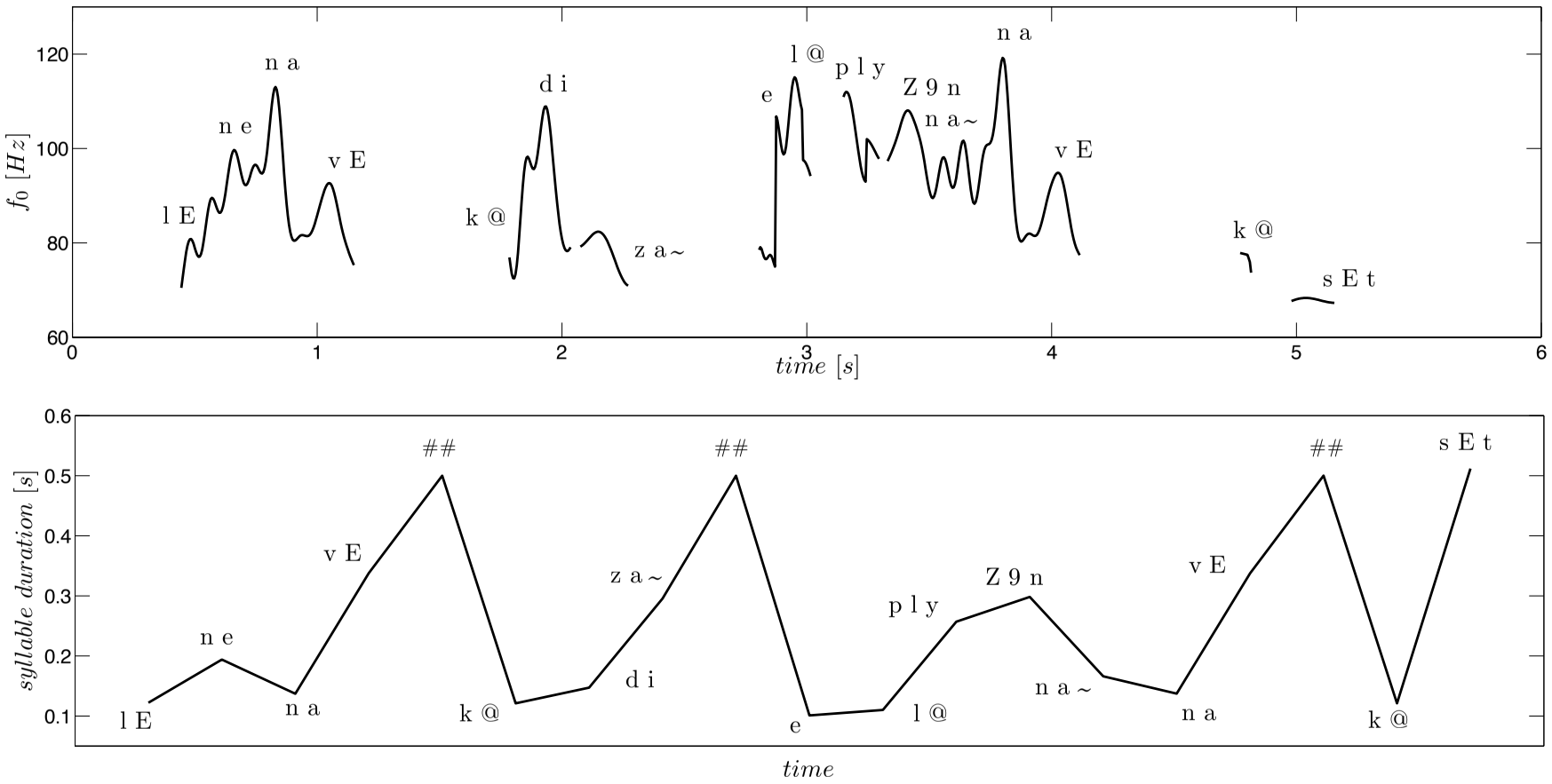


Figure 12.6: Synthesized prosodic contours of the political speaking style for the utterance "*L'ainé n'avait que dix ans, et le plus jeune n'en avait que sept.*" ("*The eldest was only ten years old, and the youngest was seven.*"). On top, melodic contour. On bottom, rhythmic contour.

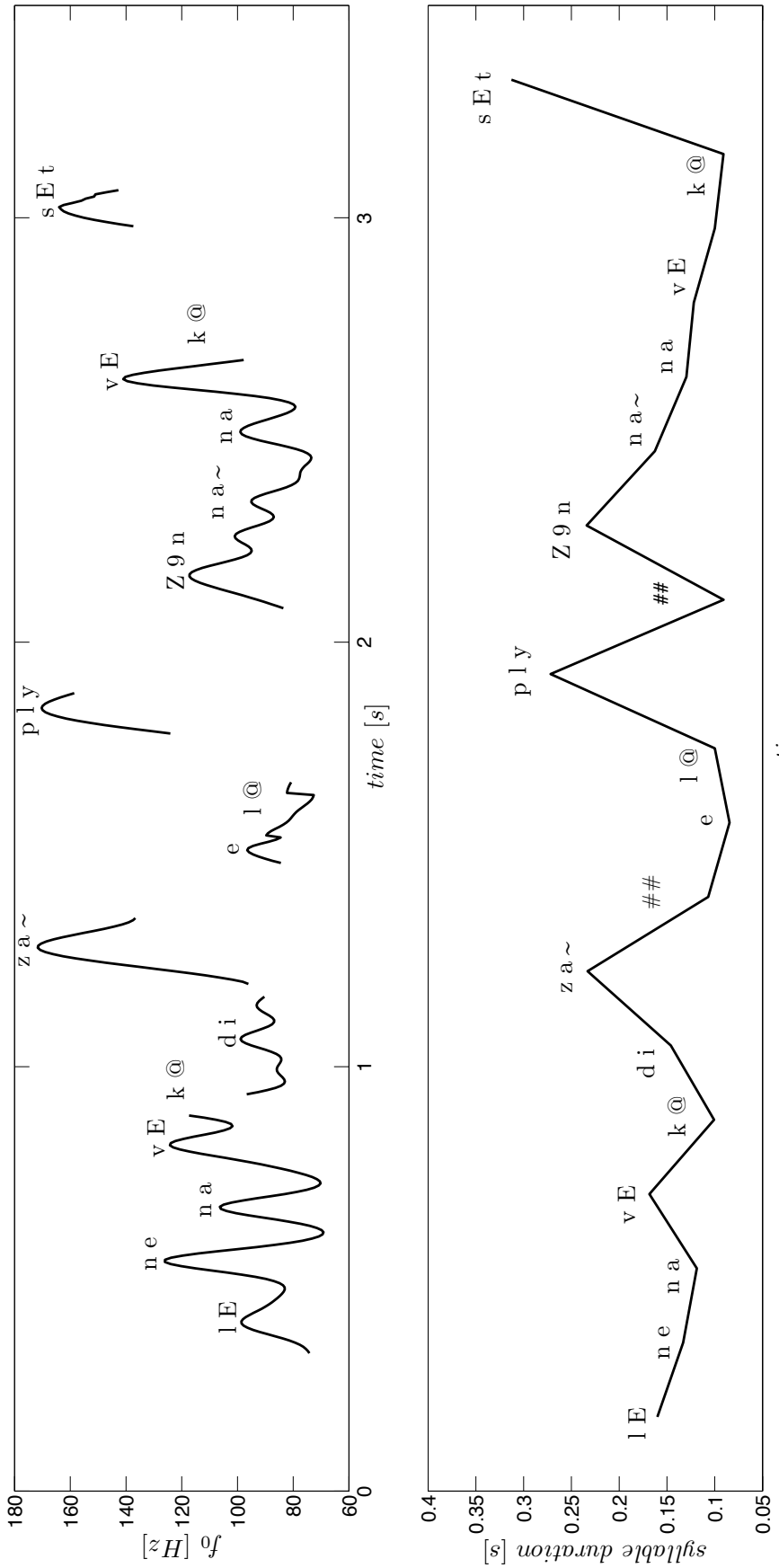


Figure 12.7: Synthesized prosodic contours of the journalistic speaking style for the utterance "*L'aîné n'avait que dix ans, et le plus jeune n'en avait que sept.*" ("*The eldest was only ten years old, and the youngest was seven.*"). On top, melodic contour. On bottom, rhythmic contour.

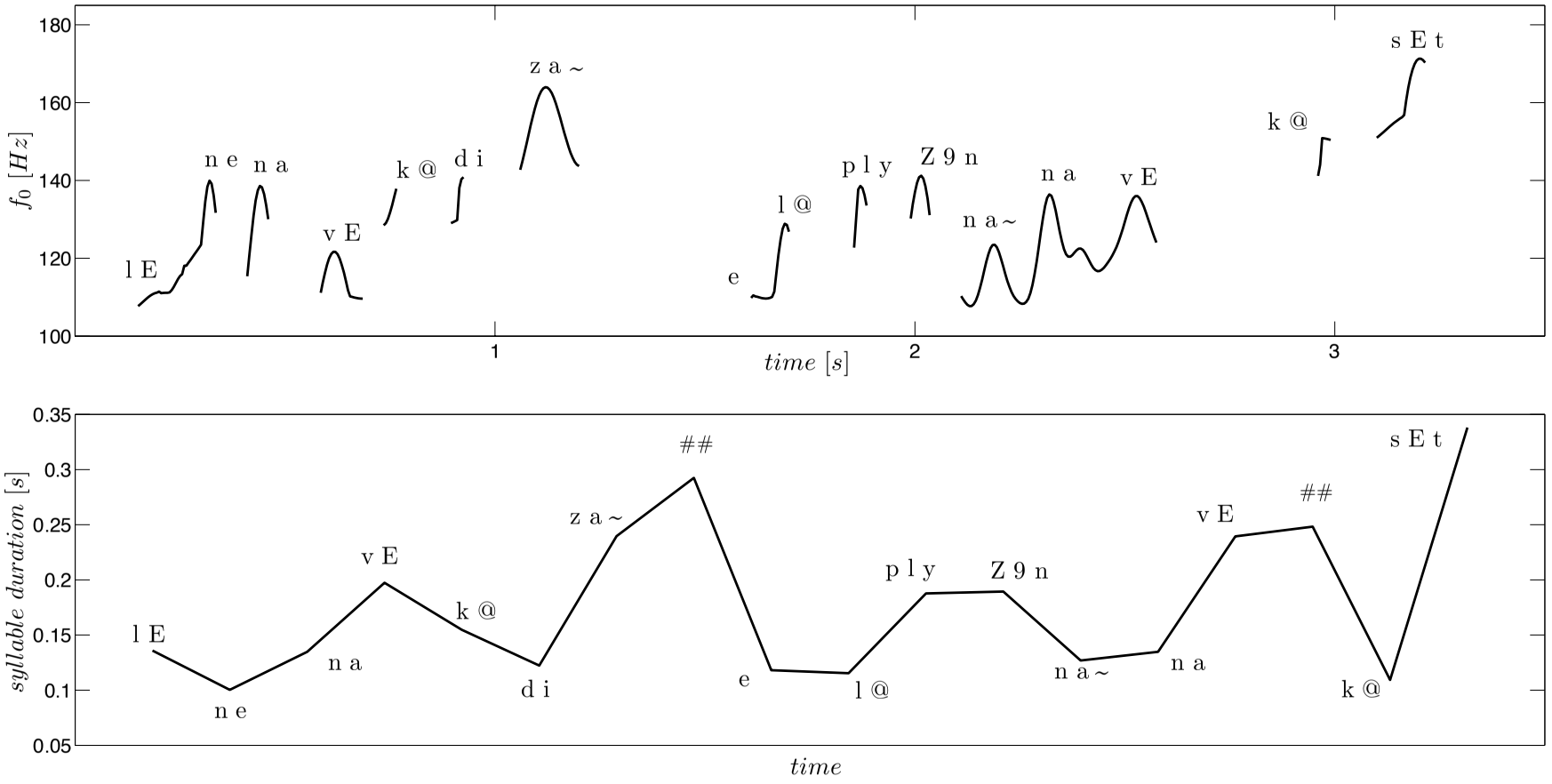


Figure 12.8: Synthesized prosodic contours of the sport-commentary speaking style for the utterance "*L'ainé n'avait que dix ans, et le plus jeune n'en avait que sept.*" ("*The eldest was only ten years old, and the youngest was seven.*"). On top, melodic contour. On bottom, rhythmic contour.

Chapter 13

General Conclusions & Further Directions

13.1 General Conclusions

In this thesis, the *Me$\mathcal{L}$os* system was presented for the analysis and synthesis of speech prosody and speaking style that can be used to control, vary, and adapt the speaking style of a speaker in speech synthesis. The major contribution of the present work was the special attention to combine theoretical linguistic and statistical modelling to provide a complete speech prosody system that can be used in speech synthesis systems. The main contributions consist of: 1) the design of a complete speech prosody system based on the discrete/continuous statistical modelling of the symbolic/acoustic characteristics of speech prosody, 2) the enrichment of the linguistic description that is used in context-dependent modelling, 3) the trajectory modelling of speech prosody based on the stylization of prosodic contours over various temporal domains, 4) the symbolic/acoustic modelling of speaking style, and to a lesser extent 5) the modelling of prosodic alternatives to vary speech prosody of a speaker in speech synthesis. The proposed methods were evaluated based on objective and subjective evaluations in speech synthesis.

The main conclusions of the present study are summarized below:

Integration of Rich Linguistic Context

An automatic linguistic processing chain was used to enrich the linguistic description of a text used for the modelling of speech prosody in context. The linguistic processing chain includes text pre-processing, surface parsing, and deep parsing based on *Tree Adjoining Grammar* (TAG) which represents both the dependency graph and the constituency structure derived from a sentence. The syntactic features extracted were classified into different sets depending on their nature: morpho-syntactic, dependency, constituency, and adjunction features that were additionally introduced.

The enrichment of the text description proved to improve dramatically the symbolic modelling of speech prosody compared to the conventional morpho-syntactic description. The gradual enrichment of the syntactic description was shown to correspond with an increase of the performance in parallel with a decrease in the complexity of the model. In particular, the constituency and the adjunction features proved to be highly relevant syntactic cues in the symbolic modelling of speech prosody. Morpho-syntactic and dependency features were only slightly relevant. The improvement is particularly significant for the modelling of major prosodic boundaries, and to a lesser extent for minor prosodic boundaries. However, the rich syntactic description failed to accurately model residual prosodic prominences that more likely relate to higher-level linguistic constraints. Finally, the combination of linguistic and metric constraints based on segmental HMMs and Dempster-Shafer fusion was shown to qualitatively improve the modelling of prosodic

breaks.

The role of linguistic description in the acoustic modelling of speech prosody is more contrasted. Enrichment of the text description was shown to significantly refine and vary the synthesized speech prosody compared to the conventional morpho-syntactic description. The improvements obtained concern local modification (local prosodic contour and dynamic), and to a lesser extent prosodic phrasing. However, the improvement obtained in speech prosody remains slight compared to the complexity of the syntactic description: the modifications obtained remain local and not systematic. In particular, the enrichment of the text description was expected to provide more clearly contrasted global variations of speech prosody (f_0 and speech rate) potentially related to specific syntactic constructions (e.g., relative clauses, incises, enumerations). However, no global change or contrast was observed either for the f_0 variations or for the state-duration.

Trajectory Modelling of Short and Long Term Variations Based on the Stylization of Speech Prosody

A trajectory model based on the stylization and the simultaneous modelling of f_0 variations over various temporal domains was presented. The syllable was used as the minimal temporal domain for the description of speech prosody, and f_0 variations are stylized and modelled simultaneously over various temporal domains that covers short-term and long-term variations. During the training, a context-dependent model is estimated according to the joint stylized f_0 contours over the syllable and a set of long-term temporal domains, and the clustering of context-dependent models is driven by long-term trajectories.. During the synthesis, f_0 variations are determined using the long-term variations as trajectory constraints. The trajectory model was used to model different temporal domains (1-order syllable context, minor prosodic group, major prosodic group), and compared to the conventional HTS model.

The 1-order syllable-context trajectory model proved to be significantly preferred to the conventional HTS model and the other trajectory models. In particular, the 1-order syllable trajectory constraint results in smooth f_0 variations and emphasized prosodic prominences, but with less micro-prosodic details, compared to the conventional HTS model. Each of the trajectory models succeeds in modelling f_0 contours that are consistent with the considered temporal domains. However, the ability of the trajectory model to account for long-term variations decreases when the temporal domain increases, due to the increase in complexity of the optimization process.

Furthermore, the evaluation of the *naturalness* of a speech prosody was questioned. In particular, the subjective evaluations revealed the difficulty for the participants to judge the difference between different alternatives of speech prosody, especially when the alternatives are perceived as equally likely. Additionally, the use of single sentences for the evaluation does not account for the variety of speech prosody across consecutive sentences - which is absolutely necessary to ensure the naturalness of synthetic speech (e.g., story telling). Finally, the evaluation of a speech prosody should be clearly reformulated by distinguishing *naturalness*, *variety*, and *liveliness* of speech prosody, and by designing specific evaluation procedures to account for them separately.

Modelling Speaking Style

The discrete/continuous modelling of speech prosody was extended to the modelling of speaking style. The issue of speaking style was theoretically introduced and related to the notion of Discourse Genre (DG). Firstly, a French speaking-style speech database was designed, comprising four speaking styles that correspond with specific situations of communication: church service (M), political speech (P), journalistic discourse (J), and sports commentary (S). A preliminary experiment was presented to assess the ability of listeners to identify a speaking style associated with a specific situation of communication, that was used as a reference for the evaluation of speaking style modelling.

In the first study, the ability of discrete/continuous HMM and HMM-based speech synthesis to model the symbolic/acoustic characteristics of various speaking styles was assessed. The acoustic description of a speaking style included timbre, voice quality, and prosodic and phonatory strategies. Incidentally, the robustness of the HMM-based speech synthesis was evaluated in the conditions of real-world applications. In the second study, the proposed discrete/continuous modelling of speech prosody was extended to the *speaker-independent* modelling of a speaking style and used to adapt the speaking style of a speaker in speech synthesis. Both methods were evaluated based on an identification experiment.

The preliminary experiment on natural speech provided evidence that a communicative situation relates to a speaking style that is shared among speakers and listeners. However, some of the speaking styles were substantially confused, for which a similarity in the situation of the discourse directly relates to a similarity in the speaking style. Additionally, the experiment demonstrated that the identification of a speaking style depends significantly on the language and/or the cultural background of the listener. The discrete/continuous modelling of speaking style was shown to consistently model the symbolic/acoustic characteristics of a speaking style, with an identification performance comparable to that obtained for natural speech. Comparison of the two methods suggests that a rich description of the acoustic speech characteristics and their local/global variations is required to model a speaking style accurately, while the conventional speech prosody characteristics (f_0, state duration) clearly do not suffice.

13.2 Further Directions

While the number of studies on the analysis and the modelling of speech prosody has dramatically increased in recent decades, the understanding and modelling of speech prosody remain ongoing "work-in-progress", due to the variety and complexity of speech prosody. The present study has raised a number of issues that remain to be solved in the analysis and statistical modelling of speech prosody.

Subjective Evaluation of Speech Prosody

Firstly, the formulation of a proper evaluation procedure has to be defined to evaluate the *naturalness* of a synthetic speech prosody. In particular, separate evaluation schemes should be employed to evaluate *correctness*, *variety*, and *liveliness* that all contribute in the perception of the *naturalness* of speech prosody. Additionally, it would be desirable to evaluate the adequacy/similarity of a speech prosody with a specific speaker, in particular with respect to his speaking style and strategies.

Description of Speech Prosody

Recent studies have argued for the hierarchical organization of speech prosody, and hierarchical models (Weighted Tree Automata) have recently been shown to model the prosodic structure grammar efficiently [Teppereman and Narayanan, 2008]. Thus, an explicit representation of the hierarchical organization of speech prosody, and the use of adequate statistical methods that can be used for context-dependent modelling would clearly improve modelling compared to conventional sequential models. Additionally, the description of speech prosody remains under debate both from the theoretical and the applicative standpoints, and the definition of the prosodic dimensions, the relevance of temporal domains used for the description of speech prosody variations, the appropriate stylization of prosodic contours, the precise phonological alphabet, and the adequate representation of prosodic structure would all contribute to the improvement of speech prosody modelling.

Trajectory Modelling

The trajectory modelling of speech prosody variations over various temporal domains is currently a popular trend in speech synthesis. In particular, the proposed trajectory modelling of short and long term speech prosody variations based on stylization has been shown to consistently model the variations associated with specific temporal domains. However, the proposed optimization of the joint-likelihood failed to accurately account for long-term variations when the temporal domain increases. The explicit formulation of the relationship that exists between syllable contours and long-term trajectories would alleviate the problem of long term trajectory modelling [Latorre and Akamine, 2008, Qian et al., 2009]. Then, trajectory modelling could be extended to any arbitrary number of temporal domains without a dramatic change in complexity during modelling and synthesis. In particular, the proposed trajectory modelling can be used to compare stylization methods and temporal domains that are used for trajectory modelling. Finally, a reformulation of the training procedure that is consistent with that used for the synthesis would improve the accuracy of the synthesized speech prosody.

Linguistic Context

The richness of the linguistic description of a text is a central issue in speech prosody modelling that is often underestimated in conventional speech synthesis systems. The refinement of the syntactic description and the integration of the higher linguistic description (e.g., semantic and discursive) would provide highly valuable information that could be used to refine the context-dependent modelling of speech prosody, and to improve the variety of the synthesized speech prosody. However, the derivation of a single context-dependent model that accounts simultaneously for the large range of linguistic levels and linguistic information is absolutely unrealistic. A reformulation of context-dependent modelling will be required in the case of very large vocabulary contexts. An appropriate formulation would probably consist of the derivation of several context-dependent models each associated with a specific linguistic dimension, and then to combine context-dependent models adequately during the synthesis of the speech prosody parameters.

Modelling Variability and Alternatives

The explicit modelling of speech prosody alternatives that correspond to the various strategies of a speaker would *de facto* improve the naturalness and variety of the speech prosody in speech synthesis [Bulyko and Ostendorf, 2001]. The statistical modelling and synthesis needs to be reformulated so as to provide a various alternatives instead of a single prosodic realization. Additionally, the reformulation would probably need to account simultaneously for short and long term variations. The statistical modelling of prosodic variability may be simply achieved with multi-modal distributions that may be combined during synthesis with more relaxed inference methods such as the *General Viterbi Algorithm* (GVA).

Unifying the Modelling of Speech Parameters

In most of the current speech synthesis systems, the inference of the speech parameters is achieved iteratively in a top-down process from the symbolic to the acoustic characteristics. For each of the levels, the optimal sequence of parameters is determined with respect to the considered level and the corresponding model. Thus, each of the levels is restricted to the parameters that are inherited from the higher-level, and does not benefit from their variability and the potential alternatives that may correspond to a more natural synthesized speech. A single method that could simultaneously model the symbolic and the acoustic characteristics and the potential alternatives would improve the quality and variety of the synthesized speech [Bulyko and Ostendorf, 2001]. This may additionally be used to vary the speech prosody of a speaker accurately in speech synthesis.

Modelling Speaking Style

Speaking style modelling can be extended to any arbitrary speaking styles associated with emotional states, situations, and sociological and geographical origins. However, a reformulation would

be required to manage a large range of para-linguistic and extra-linguistic contexts that are commonly observed in spontaneous speech. In particular, speech disfluencies (e.g., hesitations, reformulations) and para-linguistic non-verbal speech phenomena (e.g., laughter, sighs, inspiration, expiration) require specific processing that are not available in current speech synthesis systems. During the analysis, para-linguistic information has to be automatically labelled. During the training and synthesis, the location and the acoustic characteristics of the para-linguistic phenomena need to be modelled depending on the context. Additionally, the segmentation and the description of speech utterances into different types of narrative and/or discursive sequences (for instance, sports commentary significantly modifies the speech prosody characteristics depending on the more or less degree of implication of the speaker and the intensity of the action being commented on) would qualitatively improve the variety of the synthesized speaking style. Finally, more sophisticated methods have to be employed to adapt finely the characteristics of a speaker to those of a speaking style.

Appendices

Description of the Used Linguistic Features

In this appendix, a description of the linguistic characteristics that were extracted with the automatic linguistic processing chain (ALPAGE,FRMG) and used for the context-dependent modelling is provided.

parsing status

Description	:	parsing status
Type	:	symbolic
Unit	:	utterance
Alphabet	:	" full " " robust "

TAG operation (type)

Description	:	operation type used for derivation	
Type	:	symbolic	
Unit	:	form	
Alphabet	:	" adj "	adjunction
		" epsilon "	skips
		" lexical "	lexical
		" subst "	substitution

form lexical category (cat)

Description	:	syllable align form cat	
Type	:	symbolic	
Unit	:	form	
Alphabet	:	" adj "	adjective
		" adv "	adverb
		" advPref "	adverbial prefix
		" advneg "	negative adverb
		" aux "	auxiliary verb
		" ce "	sentential pronoun ("*ce*")
		" cla "	accusative clitic
		" cld "	dative clitic
		" clg "	genitive clitic
		" cll "	locative clitic
		" cln "	nominative clitic
		" clneg "	negative clitic
		" clr "	reflexive clitic
		" conj "	conjunction
		" coo "	coordinating conjunction
		" comp "	subject attribute
		" csu "	subordinating conjunction
		" det "	determinative
		" ilimp "	impersonal pronoun
		" nc "	common noun
		" ncpred "	predicate noun
		" np "	proper noun
		" number "	number
		" poncts "	sentence punctuation
		" ponctw "	form punctuation
		" predet "	pre-determinative
		" prel "	relative pronoun
		" prep "	preposition
		" pres "	presentative ("*Hélas*")
		" pri "	interrogative pronoun
		" pro "	pronoun
		" que "	sentential introducer ("*que*")
		" que-restr "	restrictive "*que*"
		" title "	title
		" v "	verb
		" xpro "	reflective pronoun

constituent category (xcat)

Description	:	syllable align form xcat	
Type	:	symbolic	
Unit	:	constituent	
Alphabet	:	" ArgComp "	verbal argument
		" CS "	subordinate phrase
		" Infl "	auxiliary verbal phrase
		" N "	nominal phrase
		" N2 "	nominal phrase
		" PP "	prepositional phrase
		" S "	sentence
		" V "	verbal phrase
		" V1 "	clitic verbal phrase
		" adj "	adjectival phrase
		" adjP "	adjectival phrase
		" adv "	adverbial phrase
		" advneg "	negative adverbial phrase
		" comp "	subject attribute phrase
		" coo "	coordinated phrase
		" det "	determinant phrase
		" nc "	common noun phrase
		" ncpred "	predicate noun phrase
		" number "	number phrase
		" prep "	prepositive phrase
		" pri "	interrogative phrase
		" pro "	pronominal phrase
		" v "	kernel verbal phrase

TAG operation (label)

Description	:	TAG operation	
Type	:	symbolic	
Unit	:	form	
Alphabet	:	" CS "	anchoring of a subordinate clause introduced by a coordinating conjunction
		" CleftQue "	que utilisé dans les constructions clivées
		" Comparative "	comparative adverb ("*Il est plus grand que Paul*")
		" Infl "	anchoring of an auxiliary verb on a verb
		" Modifier "	generic term used for an anchoring
		" Monsieur "	anchoring of a title on a noun
		" N "	anchoring of a preposed adjective on a noun
		" N2 "	anchoring of a nominal group
		" N2Rel "	anchoring of a nominal clause
		" N2app "	anchoring of a nominal apposition
		" Nc2 "	anchoring of a complex nominal construction "*c'est un mot-valise*"
		" PP "	anchoring of a prepositional group
		" Root "	generic term used for an anchoring on a tree root
		" S "	anchoring of a sentence
		" S2 "	specific anchoring of a sentence
		" SRel "	anchoring of a relative clause
		" SubS "	anchoring of a subordinate clause
		" V "	anchoring of a modal verb on a verb
		" V1 "	anchoring of a clitic on a verb
		" adj "	anchoring of an adjective
		" adv "	anchoring of an adverb
		" advneg "	anchoring of a negative adverb
		" audience "	anchoring of an audience ("*Je vous invite, chers Messieurs ...*")
		" causative-prep "	anchroging of a "à" in case of causative constructions ("*Je vais faire lire à Paul ce livre.*")
		" clg "	anchoring of a genitive clitic
		" cll "	anchoring of a locative clitic
		" clneg "	anchoring of a negative clitic
		" clr "	anchoring of reflexive clitic
		" comp "	anchoring of a subject attribute
		" coord, coord2, coord3 "	anchoring of a coordinating conjunction
		" csu "	anchoring of a subordinating conjunction
		" de "	anchoring of the lexical item "*de*"
		" det "	anchoring of a determinative
		" impsubj "	anchoring of a impersonal subject
		" nc "	anchoring of a common noun
		" ncpred "	anchoring of a predicative noun on a verb
		" ni "	special anchoring of the lexical item "*ni*"
		" np "	anchoring of a proper noun
		" number "	anchoring of a number
		" object "	anchoring of an object
		' person-mod "	anchoring on a personne ("*Les enfants, venez à table!*")
		" predet-ante "	anchoring of an ante-posed pre-determinant "*Tous ceux qui sont les exclus, les marginalisés.*"
		" predet-post "	anchoring of an post-posed pre-determinant "*A vous tous et à vous toutes, j'exprime les voeux de la République*".
		" prel "	anchoring of a relative pronoun
		" prep "	anchoring of a preposition
		" preparg "	anchoring of a verbal argument on a preposition ("*il donne un livre à Paul*")
		" pri "	anchoring of an interrogative pronoun
		" pro "	anchoring of a pronoun
		" que "	special anchoring of the lexical item "*que*"
		" quoted-S "	anchoring of a quotation
		" reference "	anchoring of a bibliographical reference
		" skip "	specific anchoring that manage spontaneous speech (hesitations, repetitions,...)
		" supermod "	anchoring on a superlative adverb
		" time-mod "	anchoring of a temporal adverb
		" vmod "	anchoring on a verb
		" void "	lexical anchoring
		" wh "	anchoring of an interogative
		" xcomp "	anchoring of a sentential verb

Related Projects

Rhapsodie: Reference Prosody Corpus of Spoken French [1]

Implications: prosodic transcription (responsible), corpus design, speech segmentation.

The project aims to constitute a reference corpus of spoken French subdivided into different representative discourse genres equipped with prosodic and syntactic semi-automatic annotations.

Since the beginning of the 1980s, a number of large-scale projects aiming to set up oral corpora for widely-spoken languages have been launched. More recently, various systems for sharing of resources and exchange were put in place at national level (see the Resource Centre for the Description of the Spoken Language (CRDO)). Three basic questions arise from these efforts to collect, exploit and store oral corpora: their subdivision into representative discourse genres, the transcription conventions adopted, the types of annotation made available (with the associated issue of standards of annotation - a major issue in connection with prosody, which taken overall remains the poor relation). In this context, our project aims to constitute a reference corpus of spoken French subdivided into different representative discourse genres equipped with prosodic and syntactic annotations that may be used in the analysis of the status of prosody in discourse as well as of its relations with syntax and information structure.

Participants: Laboratoire Modèles, Dynamiques, Corpus (Modyco, Nanterres), Institut de Recherche et Coordination Acoustique Musique (IRCAM, Paris), Laboratoire Langues, Textes, Traitements Informatiques, Cognition (LATTICE, Paris), Equipe de Recherche en Syntaxe et Sémantique (ERSS, Toulouse), Laboratoire Parole et Langage (LPL, Aix-en-Provence)

Support: Agence Nationale de la Recherche (ANR).

EMUS: Expressivity in MUsic and Speech [2]

Implications: organization, scientific committee.

Speech and music conceal a treasure of "expressive potential". In spite of semiotic differences, numerous aspects are common to music and speech because they share the same physical medium of communication (sound). If it is collectively agreed by both the scientific and the artistic communities that speech and music allow people to express, to perceive and to induce expressivity, the comparison between these two media has not been sufficiently productive . EMUS aims at gathering various communities with an interest for expressivity in speech andin music. These communities are numerous and are presented below (the list is not exhaustive):

expressivity (emotion) scientific domains: psychology, philosophy, neurosciences

speech (language) scientific domains: linguistics, psycho-phonetics, phonology, prosody, speech processing, cognitive sciences, neuro-linguistics

speech (artistic domains): theatre, poetry, cinema, numeric arts

[1] `http://rhapsodie.risc.cnrs.fr`
[2] `http://recherche.ircam.fr/equipes/analyse-synthese/EMUS`

music (scientific domain): musicology, performance, sound perception, cognitive sciences, neurosciences

music (artistic domains): composition, performance

The variety of the actors evoked above shows that the question is deeply multidisciplinary. To allow possible collaborations and to enrich our perspectives, EMUS is organizing four international conferences:

1. Prosody and Expressivity in Speech and Music (linguistics), Workshop of the International Conference on Speech Prosody, May 5th 2008, Campinas, Brazil.
2. Prosody of Expressivity in Music and Speech (performance/acoustic/music),AGORA Contemporary Music Festival, June 17th-18th 2008, IRCAM, Paris, France.
3. De la musique au langage : prosodie et babillage (production/acquisition), May 16th 2008, Ecole Normale Supérieur de Lyon, Lyon, France,
4. Microgenesis and semiotics of perceptual process (perception/semiotics), September 25th-26th 2008, RISC, Paris, France.

Organization: IRCAM (Grégory Beller, Nicolas Obin, Andrew Gerzso, Florence Quilliard, Xavier Rodet), University of Geneva, Linguistics Department (Antoine Auchlin), MODYCO - University of Nanterre (Anne Lacheret), ICAR, Ecole Normale Supérieure Lettres et Sciences Humaines (Aliyah Morgenstern).

Scientific Committee: Christophe d'Alessandro (LIMSI, Orsay), Antoine Auchlin (University of Geneva, Linguistics Department), Grégory Beller (IRCAM, Paris), Nick Campbell (ATR, Nara), Anne Lacheret (MODYCO, University of Nanterre), Sandra Madureira (PUC-SP), Aliyah Morgenstern (ICAR, Ecole Normale Supérieure Lettres et Sciences Humaines, Paris), Nicolas Obin (IRCAM, Paris).

HyperMusic: Prologue [3]

Implications: automatic segmentation and clustering of voice recordings (spoken and singing voice, vocalizations) for real-time control.

The project consists in the coordination of research and music for the composition and the creation of the opera: HyperMusic Prologue by composer Hector Parra. The objective of the research is the segmentation of different types of voice recordings (spoken and singing voice, various types of vocalization) into relevant acoustic segments, so as to design off-line acoustic databases that can be used for real-time processing. Segmentation was based on different topologies of hidden Markov models (HMM), either supervised by linguistic information extracted from text, or using clustering techniques to classify voice segments into acoustically similar clusters.

HyperMusic Prologue (2008-2009): A projective opera in seven planes, opéra de chambre pour deux voix, huit intrumentistes et électronique.

[3] `http://brahms.ircam.fr/works/work/23852/#program`

Composer: Hector Parra.
Booklet: Lisa Randall.

Genre: Vocal music with instruments.
Instrumentation: soloists : 1 soprano solo, 1 baryton solo, 1 flute, 1 clarinet, 1 French horn, 1 percussion, 1 violin, 1 alto, 1 cello, 1 contrabass, 1 real-time processing.

Electronics: Thomas Goepfer, Musical Assistant (IRCAM).

Research: Nicolas Obin, Pierre Lanchantin, Ashleigh Gonzales (Analysis-Synthesis Team, IRCAM).

Creation: 14 June 2009, Paris, Agora Festival, Centre-Pompidou, by Matthew Ritchie : scenography, Paul Desveaux : spatialization, Laurent Schneegans : lights, Charlotte Ellett : soprano, James Bobby : baritone, Ensemble intercontemporain, direction : Clément Power.

Command : Ensemble Intercontemporain and Ircam-Centre Pompidou, supported by the Catalan Ministry of Culture.

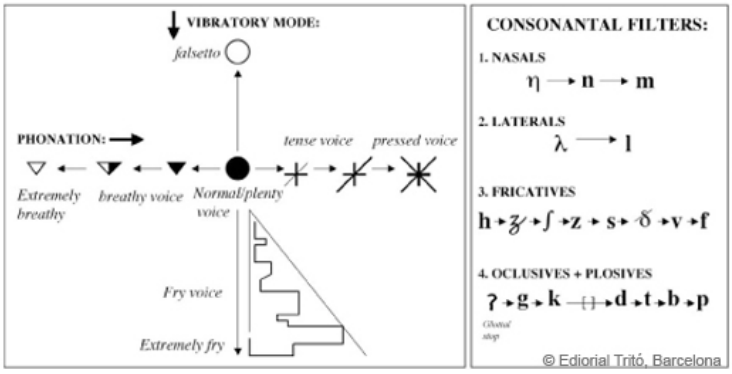


Figure 13.1: Illustration of Hypermusic Prologue.

Projet Exploratoire Pluridisciplinaire (PEPS)

Implications: organization, modelling speech prosody based on rich linguistic descriptions.

The project is a cross-disciplinary research on the modelling of speech prosody for speech synthesis. The objective of the project is the combination of statistical modelling and linguistics for the modelling of speech prosody. The major focus will be the integration of a rich description of speech prosody (transcription, representation) and text (surface/deep syntactic structure) characteristics into the statistical modelling of speech prosody. First, a method for the description of the symbolic and acoustic characteristics of speech prosody will be proposed. Second, the short and long term characteristics of speech prosody and the temporal domains over which relevant speech prosody variations occur will be investigated. Finally, a rich description of the linguistic characteristics of a text based on deep syntactic parsing will be used to improve the symbolic/acoustic modelling of speech prosody.

Participants: Nicolas Obin (IRCAM), Julie Glickman (MoDyCo Lab.).

Support: PEPS - ST2I 2008.

List of Publications

Book Chapters

[Rodet et al., 2009] Rodet, X., Beller, G., Bogaards, N., Degottex, G., Farner, S., Lanchantin, P., Obin, N., Roebel, A., Veaux, C., and Villavicencio, F. (2009). *Parole et musique*, chapter Transformation et synthèse de la voix parlée et de la voix chantée. Odile Jacob, Paris.

Journal Papers

[Obin et al., 2011e] Obin, N., Lanchantin, P., Lacheret, A., and Rodet, X. (2011e). Symbolic Modelling of Speech Prosody: From Linguistics to Statistical Modelling. Submitted to *IEEE Transactions on Audio, Speech, and Language Processing*.

[Avanzi et al., 2011] Avanzi, M., Lacheret-Dujour, A., Obin, N., and Victorri, B. (2011). Vers une modélisation continue de la structure prosodique. Le cas des proéminences accentuelles. *Journal of French Language Studies*, 21(1):53–71.

[Beller et al., 2009] Beller, G., Veaux, C., Degottex, G., Obin, N., and Lanchantin, P. Rodet, X. (2009). Ircamcorpustools : Plateforme pour les corpus de parole. *Traitement Automatique des Langues*, 49(3).

International Conference Proceedings

[Obin et al., 2011a] Obin, N., Avanzi, M., and Lacheret, A. (2011a). Transcription of French Prosody in Discourse: the Rhapsodie Protocole. In *Interface Discours Prosodie*, Manchester, U.K.

[Obin et al., 2011b] Obin, N., Lacheret, A., and Rodet, X. (2011b). Stylization and Trajectory Modelling of Short and Long Term Speech Prosody Variations. In *Interspeech*, pages 2029–2032, Florence, Italy.

[Obin et al., 2011c] Obin, N., Lanchantin, P., Lacheret, A., and Rodet, X. (2011c). Discrete/Continuous Modelling of Speaking Style in HMM-based Speech Synthesis: Design and Evaluation. In *Interspeech*, pages 2785–2788, Florence, Italy.

[Obin et al., 2011d] Obin, N., Lanchantin, P., Lacheret, A., and Rodet, X. (2011d). Reformulating Prosodic Break Model into Segmental HMMs and Information Fusion. In *Interspeech*, pages 1829–1832, Florence, Italy.

[Avanzi et al., 2011a] Avanzi, M., Bordal, G., and Obin, N. (2011a). Typological Variations in the Realization of French Accentual Phrase. In *International Congress of Phonetic Sciences*, pages 268–271, Hong Kong, China.

[Lanchantin et al., 2011a] Lanchantin, P., Farner, S., Veaux, C., Degottex, G., Obin, N., Beller, G., Villavicencio, F., Hueber, T., Schwartz, D., Huber, S., Peeters, G., Roebel, A., and Rodet, X. (2011a). Vivos Voco: A Survey of Recent Research on Voice Transformations at IRCAM. In *International Conference on Digital Audio Effects (DAFx)*, pages 277–285, Paris, France.

[Lanchantin et al., 2011b] Lanchantin, P., Obin, N., and Rodet, X. (2011b). Extended Conditional GMM and Covariance Matrix Correction for Real-Time Spectral Voice Conversion. Submitted to *Interspeech*, Florence, Italy.

[Obin et al., 2010a] Obin, N., Lacheret, A., and Rodet, X. (2010a). Expectations for Speaking Style Identification: a Prosodic Study. In *Interspeech*, pages 3070–3073, Makuhari, Japan.

[Obin et al., 2010b] Obin, N., Lacheret, A., and Rodet, X. (2010b). HMM-based Prosodic Structure Model using Rich Linguistic Context. In *Interspeech*, pages 1133–1136, Makuhari, Japan.

[Obin et al., 2010c] Obin, N., Lanchantin, P., Lacheret, A., and Rodet, X. (2010c). Towards Improved HMM-based Speech Synthesis using High-Level Syntactical Features. In *Speech Prosody*, Chicago, U.S.A.

[Lacheret et al., 2010] Lacheret, A., Obin, N., and Avanzi, M. (2010). Design and Evaluation of Shared Prosodic Annotation for Spontaneous French Speech: From Expert Knowledge to Non-Expert Annotation. In *Linguistic Annotation Workshop*, pages 265–273, Uppsala, Sweden.

[Obin et al., 2009a] Obin, N., Rodet, X., and Lacheret-Dujour, A. (2009a). A Multi-Level Context-Dependent Prosodic Model Applied To Durational Modeling. In *Interspeech*, pages 512–515, Brighton, U.K.

[Obin et al., 2009b] Obin, N., Rodet, X., and Lacheret-Dujour, A. (2009b). A Syllable-Based Prominence Model Based On Discriminant Analysis And Context-Dependency. In *International Conference on Speech and Computer*, pages 97–100, St-Petersburg, Russia.

[Obin et al., 2008b] Obin, N., Lacheret, A., Veaux, C., Rodet, X., and Simon, A.-C. (2008b). A Method for Automatic and Dynamic Estimation of Discourse Genre Typology with Prosodic Features. In *Interspeech*, pages 1204–1207, Brisbane, Australia.

[Obin et al., 2008c] Obin, N., Rodet, X., and Lacheret-Dujour, A. (2008c). French Prominence: a Probabilistic Framework. In *International Conference on Audio, Speech, and Signal Processing*, pages 3993–3996, Las Vegas, U.S.A.

[Beller et al., 2008] Beller, G., Obin, N., and Rodet, X. (2008). Articulation Degree as a Prosodic Dimension of Expressive Speech. In *Speech Prosody*, Campinas, Brazil.

National Conference Proceedings

[Obin, 2010] Obin, N. (2010). Modélisation du Style en Synthèse de la Parole. In *Journées Jeunes Chercheurs en Audition, Acoustique musicale et Signal audio*, Paris, France.

[Avanzi, 2010] Avanzi, M., Obin, N., Lacheret, A. (2010). Vers une Modélisation Continue de la Structure Prosodique du Français: le Cas des Proéminences Accentuelles. In *Journées Conscila*, ENS ULM, Paris, France.

[Obin et al., 2008a] Obin, N., Goldman, J.-P., Avanzi, M., and Lacheret-Dujour, A. (2008a). Comparaison de trois outils de détection automatique de proéminences en français parlé. In *Journées d'Etude de la Parole*, pages 85–88, Avignon, France.

[Obin et al., 2008d] Obin, N., Rodet, X., and Lacheret-Dujour, A. (2008d). Un modèle de durée des syllabes fondé sur les propriétés syllabiques intrinsèques et les variations locales de débit. In *Journées d'Etude de la Parole*, pages 333–336, Avignon, France.

Bibliography

[Abeillé, 1988] Abeillé, A. (1988). Parsing french with tree adjoining grammar: some linguistic accounts. In *International Conference on Computational Linguistics*, pages 7–12, Budapest,Hungary.

[Abney, 1992] Abney, S. (1992). Prosodic structure, performance structure and phrase structure. In *Human Langage Technology: Proceedings of the workshop on Speech and Natural Language*, pages 425–428, Morristown, NJ, USA. Association for Computational Linguistics.

[Acapela Group, 2010] Acapela Group (2010). Acapela Speech Synthesis System. http://www.acapela-group.com/text-to-speech-interactive-demo.html.

[Anastasakos et al., 1997] Anastasakos, T., McDonough, J., and Makhoul, J. (1997). Speaker adaptive training : A maximum likelihood approach to speaker normalization. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 1043–1046, Munich, Germany.

[Andersen, 1999] Andersen, P. (1999). *Nonverbal Communication: Forms and Functions*. McGraw-Hill.

[Aristotle, 0 BC] Aristotle (350 BC). *Poetics*.

[Artaud, 1938] Artaud, A. (1938). *Le Théâtre et son double*. Gallimard, Paris.

[Atkinson, 1978] Atkinson, J. (1978). Correlation analysis of the physiological factors controlling fundamental voice frequency. *Journal of the Acoustic Society of America*, 63(1):211–222.

[AT&T Labs Natural Voices, 2010] AT&T Labs Natural Voices (2010). Natural Voices Speech Synthesis System. http://www2.research.att.com/ ttsweb/tts/demo.php.

[Atterer and Klein, 2002] Atterer, M. and Klein, E. (2002). Integrating linguistic and performance-based constraints for assigning phrase breaks. In *International Conference on Computational Linguistics*, pages 995–998, Taipei, Taiwan.

[Aubergé, 1991] Aubergé, V. (1991). *La synthèse de la parole: "des règles aux lexiques"*. PhD. thesis, Université Pierre Mendès-France, Grenoble, France.

[Austin, 1962] Austin, J. L. (1962). *How to Do Things With Words*. Oxford University Press, Oxford.

[Avanzi et al., 2011a] Avanzi, M., Bordal, G., and Obin, N. (2011a). Typological Variations in the Realization of French Accentual Phrase. In *International Congress of Phonetic Sciences*, pages 268–271, Hong Kong, China.

[Avanzi et al., 2007] Avanzi, M., Goldman, J.-P., Lacheret-Dujour, A., Simon, A.-C., and Auchlin, A. (2007). Méthodologie et algorithmes pour la détection automatique des syllabes proéminentes dans les corpus de français parlé. *Cahiers of French Language Studies*, 13(2):2–30.

[Avanzi et al., 2011b] Avanzi, M., Lacheret-Dujour, A., Obin, N., and Victorri, B. (2011b). Vers une modélisation continue de la structure prosodique. le cas des proéminences accentuelles. *Journal of French Language Studies*, 21(1):53–71.

[Avanzi et al., 2008] Avanzi, M., Lacheret-Dujour, A., and Victorri, B. (2008). Analor: A Tool for Semi-Automatic Annotation of French Prosodic Structure. In *Speech Prosody*, pages 119–122, Campinas,Brazil.

[Bachorowski, 1999] Bachorowski, J. A. (1999). Vocal expression and perception of emotion. *Current Directions in Psychological Science*, 8:53–57.

[Bahl et al., 1983] Bahl, L., Jelinek, F., and Mercer, R. (1983). A maximum likelihood approach to continuous speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5(2):179–190.

[Bailly, 1989] Bailly, G. (1989). Integration of rhythmic and syntactic constraints in a model of generation of french prosody. *Speech Communication*, 8(2):137–146.

[Bakhtin, 1984] Bakhtin, M. (1984). *Esthétique de la création verbale (The aesthetics of verbal creation)*. Gallimard, Paris.

[Barbosa, 2004] Barbosa, P. (2004). *Caractérisation et génération automatique de la structure rythmique du français*. PhD. Thesis, Institut de la Communication Parlée, Grenoble.

[Barbosa, 2006] Barbosa, P. (2006). A dynamical model for generating prosodic structure. In *Speech Prosody*, pages 366–369, Dresden, Germany.

[Barnlund, 1968] Barnlund, D. C. (1968). *Interpersonal Communication: Survey and Studies*. Houghton Mifflin.

[Baum and Petrie, 1966] Baum, L. E. and Petrie, T. (1966). Statistical inference for probabilistic functions of finite state markov chains. *Annals of Mathematical Statistics*, 37(6):1554–1563.

[Baum et al., 1970] Baum, L. E., Petrie, T., Soules, G., , and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *Annals of Mathematical Statistics*, 41(1):164–171.

[Baumann et al., 2000] Baumann, S., Grice, M., and Benzmueller, R. (2000). GToBI - a phonological system for the transcription of german intonation. In *Prosody 2000: Speech Recognition and Synthesis*, pages 21–28, Kraków , Poland.

[Beaugendre, 1992] Beaugendre, F. (1992). A perceptual study of French intonation. In *International Conference on Spoken Language Processing*, pages 739–742, Alberta, Canada.

[Beaver et al., 2007] Beaver, D., Zack Clarck, B., Flemming, E., Jaeger T., F., and Wolters, M. (2007). When semantics meets phonetics : Acoustical studies of second-occurrence focus. *Journal of the Linguistic Society of America*, 83(2):245–276.

[Beckman and Ayers, 1997] Beckman, M. and Ayers, G. (1997). Guidelines for ToBI labelling. Technical report, Linguistics Department, Ohio State University.

[Beckman and Jun, 1996] Beckman, M. and Jun, S.-A. (1996). K-ToBI (Korean ToBI) labelling convention. Technical report, Ohio State University and UCLA. http://www.linguistics.ucla.edu/people/jun/ktobi/ktobi3-2.pdf.

[Beckmann, 1986] Beckmann, M. (1986). *Stress and non-stress accent*. Foris, Dordrecht.

[Bell et al., 2006] Bell, P., Burrows, T., and Taylor, P. (2006). Adaptation of prosodic phrasing models. In *Speech Prosody*, Dresden, Germany.

[Beller et al., 2008] Beller, G., Obin, N., and Rodet, X. (2008). Articulation Degree As a Prosodic Dimension of Expressive Speech. In *Speech Prosody*, Campinas, Brazil.

[Beller et al., 2009] Beller, G., Veaux, C., Degottex, G., Obin, N., and Lanchantin, P. Rodet, X. (2009). IrcamCorpusTools : Plateforme Pour Les Corpus de Parole. *Traitement Automatique des Langues*, 49(3).

[Bennett and Rodet, 1989] Bennett, G. and Rodet, X. (1989). *Current directions in computer music research*, chapter Synthesis of the singing voice, pages 19–44. MIT Press.

[Benvéniste, 1966] Benvéniste, E. (1966). *Problème de linguistique générale*. Gallimard, Paris.

[Biber, 1988] Biber, D. (1988). *Variation Across Speech and Writing*. Cambridge University Press, Cambridge.

[Black and Taylor, 1994] Black, A. and Taylor, P. (1994). Assigning intonation elements and prosodic phrasing for english speech synthesis from high level linguistic input. In *International Conference on Spoken Language Processing*, pages 715–718, Yokohama, Japan.

[Black and Taylor, 1997a] Black, A. W. and Taylor, P. (1997a). Assigning Phrase Breaks from Part-of-Speech sequences. In *European Conference on Speech Communication and Technology*, pages 995–998, Rhodes, Greece.

[Black and Taylor, 1997b] Black, A. W. and Taylor, P. A. (1997b). The Festival Speech Synthesis System: System documentation. Technical Report HCRC/TR-83, Human Communciation Research Centre, University of Edinburgh, Scotland, UK.

[Bogaards and Roebel, 2005] Bogaards, N. and Roebel, A. (2005). An interface for analysis-driven sound processing. In *Convention of the Audio Engineering Society*, New York, USA.

[Bonafonte and Agüero, 2004] Bonafonte, A. and Agüero, P. D. (2004). Phrase break prediction using a finite state transducer. In *Advances in Speech Technology Workshop*, pages 1275–1278, Maribor, Slovenia.

[Borg and Groenen, 2005] Borg, I. and Groenen, P. (2005). *Modern Multidimensional Scaling: theory and applications*. Springer-Verlag.

[Boula de Mareuil et al., 2008] Boula de Mareuil, P., Rilliard, A., and Allauzen, A. (2008). A diachronic study of prosody through french audio archives. In *Speech Prosody*, page 531–534.

[Boula de Mareüil, 1997] Boula de Mareüil, P. (1997). *Etude linguistique appliquée à la synthèse de la parole à partir du texte*. PhD. thesis, Université de Paris XI, Orsay.

[Boulez, 2005] Boulez, P. (2005). *Points de repère III*, chapter Automatisme et décision. Christian Bourgois, Paris.

[Box et al., 1978] Box, G. E. P., Hunter, W. G., and Hunter, J. S. (1978). *Statistics for Experimenters*. John Wiley Sons.

[Breiman et al., 1984] Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and regression trees*. Wadsworth & Brooks.

[Broth et al., 2005] Broth, M., Forsgren, M.and Norén, C., and Sullet-Nylander, F. (2005). Le français parlé des médias. In *Le Français parlé des médias*, pages 63–83, Stockholm, Sweden.

[Broyden, 1970] Broyden, C. (1970). The convergence of a class of double-rank minimization algorithms. *Journal of the Institute of Mathematics and Its Applications*, 6:76–90.

[Bulyko and Ostendorf, 2001] Bulyko, I. and Ostendorf, M. (2001). Joint Prosody Prediction And Unit Selection For Concatenative Speech Synthesis. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 781–784, Salt Lake City, USA.

[Béchet, 2001] Béchet, F. (2001). Lia_Phon : un système complet de phonétisation de textes. *Traitement Automatique des Langues*, pages 47–67.

[Bühler, 1934] Bühler, K. (1934). *Sprachtheorie. Die Darstellungsfunktion der Sprache*. Verlag von Gustav Fischer, Jena.

[Campbell, 1992] Campbell, N. (1992). Prosodic encoding of english speech. In *International Conference on Spoken Language Processing*, pages 663–666, Edmonton, Canada.

[Campbell, 2000] Campbell, N. (2000). *Prosody: Theory and experiment*, chapter Timing in speech: A multi-level process, pages 281–334. Kluwer Academic Publishers.

[Campbell and Erickson, 2004] Campbell, N. and Erickson, D. (2004). What do people hear? a study of the perception of non-verbal affective information in conversational speech. *Journal of the Phonetic Society of Japan*, 7(4):9–28.

[Campbell and Mokhtari, 2003] Campbell, N. and Mokhtari, P. (2003). Voice quality: the 4th prosodic dimension. In *International Congress of Phonetic Sciences*, pages 2417–2420, Barcelona, Spain.

[Campione et al., 2000] Campione, E., Hirst, D., and Véronis, J. (2000). *Automatic stylisation and symbolic coding of F0: implementations of the INTSINT model*, chapter Intonation. Research and Applications. Kluwer, Dordrecht.

[Cepstral, 2010] Cepstral (2010). Cepstral Speech Synthesis System. http://www.cepstral.com/demos/.

[Chen et al., 2003] Chen, S.-H., Lai, W.-H., and Wang, Y.-R. (2003). A new duration modeling approach for mandarin speech. *IEEE Transactions on Speech and Audio Processing*, 11(4):308–320.

[Cohen et al., 2001] Cohen, H., Douaire, J., and Elsabbagh, M. (2001). The role of prosody in discourse processing. *Brain & Cognition*, 46:73–82.

[Cohen, 1960] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

[Cohen, 1968] Cohen, J. (1968). Weighted Kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220.

[Condon, 1971] Condon, W. S. (1971). Speech and Body Motion Synchrony of the Speaker-Hearer. *Perception of Language*, page 150–173.

[Cont, 2010] Cont, A. (2010). A coupled duration-focused architecture for realtime music to score alignment. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 32(6):974–987.

[Cooper and Paccia-Cooper, 1980] Cooper, W. E. and Paccia-Cooper, J. (1980). *Syntax and speech*. Harvard University Press, Cambridge.

[Cutler, 1997] Cutler, A. (1997). *Computing prosody: Computational models for processing spontaneous speech*, chapter Prosody and the structure of the message, pages 63–66. Springer Verlag.

[d'Alessandro and Mertens, 1995] d'Alessandro, C. and Mertens, P. (1995). Automatic pitch contour stylization using a model of tonal perception. *Computer Speech and Language*, 9(3):257–288.

[de Boor, 1978] de Boor, C. (1978). *A Pratical Guide to Splines*. Springer Verlag.

[De Jaegher and Di Paolo, 2007] De Jaegher, H. and Di Paolo, E. (2007). Participatory sense-making. An enactive approach to social cognition. *Phenomenology the Cognitive Sciences*, 6:485–507.

[Degand and Simon, 2009] Degand, L. and Simon, A.-C. (2009). Mapping prosody and syntax as discourse strategies: how basic discourse units vary across genres. *Where Prosody Meets Pragmatics*, pages 81–107.

[Degottex et al., 2010] Degottex, G., Roebel, A., and Rodet, X. (2010). Phase minimization for glottal model estimation. *IEEE Transactions on Audio, Speech, and Language Processing*. Accepted Publication.

[Delais, 1994] Delais, E. (1994). Rythme et structure prosodique en français. *French Generative Phonology : Retrospectives and Perspectives*, pages 131–150.

[Delais-Roussarie, 2000] Delais-Roussarie, E. (2000). Vers une nouvelle approche de la structure prosodique. *Langue Française*, 126.

[Delais-Roussarie and Durand, 2003] Delais-Roussarie, E. and Durand, J. (2003). *Corpus et variation en phonologie du français: méthodes et analyses*. Presses Universitaires du Mirail, Toulouse.

[Delais-Roussarie et al., 2006] Delais-Roussarie, E., Post, B., and Portes, C. (2006). Annotation prosodique et typologie. *Travaux interdisciplinaires du Laboratoire parole et langage d'Aix-en-Provence*, 25:61–95.

[Delattre, 1938] Delattre, P. (1938). L'accent final en français : accent d'intensité, accent de hauteur, accent de durée. *French Review*, 12(2):141–145.

[Delattre, 1966] Delattre, P. (1966). Les dix intonations de base du français. *The French Review*, 40(1):1–14.

[Delattre, 1969] Delattre, P. (1969). L'intonation par les oppositions. *Le français dans le monde*, 64:6–13.

[Deleuze and Guattari, 1975] Deleuze, G. and Guattari, F. (1975). *Kafka. Pour une Littérature Mineure*. Editions de Minuit, Paris.

[Dell, 1984] Dell, F. (1984). L'accentuation dans les phrases en français. *Forme sonore du langage: structure des représentation en phonologie*, pages 65–122.

[Depalle et al., 1994] Depalle, P., Garcia, G., and Rodet, X. (1994). A virtual castrato(!?). In *International Computer Music Conference*, pages 357–360, Aarhus, Denmark.

[Deulofeu, 1998] Deulofeu, J. (1998). Les commentaires sportifs constituent-t-ils un genre ? au sens linguistique du terme ? *Colloque Questions de méthode dans la linguistique sur corpus*.

[Devijver and Kittler, 1982] Devijver, P. A. and Kittler, J. (1982). *Pattern Recognition: A Statistical Approach*. Prentice-Hall.

[Di Cristo, 1985] Di Cristo, A. (1985). *De la Microprosodie à l'Intonosyntaxe*. Publications de l'Université d'Aix-en-Provence, Aix-en-Provence.

[Di Cristo, 2004] Di Cristo, A. (2004). La prosodie au carrefour de la phonétique, de la phonologie et de l'articulation formes-fonctions. *Travaux Interdisciplinaires du Laboratoire Parole et Langage*, 23:67–211.

[Dittmar, 1996] Dittmar, N. (1996). *Theoretical linguistics and grammatical description*, chapter Explorations in 'idiolects', pages 109–128. Benjamins, Amsterdam.

[Donovan and Woodland, 1995] Donovan, R. and Woodland, P. (1995). Automatic speech synthesizer parameter estimation using HMMs. In *International Conference on Audio, Speech, and Signal Processing*, pages 640–643, Detroit, Michigan.

[Dudley et al., 1939] Dudley, H., Riesz, R., and Watkins, S. (1939). A synthetic speaker. *Journal of the Franklin Institute*, 227(6):739–764.

[Dusterhoff et al., 1999] Dusterhoff, K. E., Black, A. W., and Taylor, P. (1999). Using decision trees within the Tilt intonation model to predict F0 contours. In *European Conference on Speech Communication and Technology*, pages 1627–630.

[d'Alessandro and Castellengo, 1994] d'Alessandro, C. and Castellengo, M. (1994). The pitch of short-duration vibrato tones. *Journal of the Acoustical Society of America*, 95:1617–1630.

[Eskenazi, 1993] Eskenazi, M. (1993). Trends in speaking styles research. In *European Conference on Speech Communication and Technology*, pages 501–509, Berlin, Germany.

[Fant, 1953] Fant, G. (1953). Speech communication research. *Royal Swedish Academy of Engineering Sciences*, 2:331–337.

[Farges and Clements, 1988] Farges, E. P. and Clements, M. A. (1988). An analysis-synthesis hidden Markov model of speech. In *International Conference on Audio, Speech, and Signal Processing*, pages 323–326, New-York, USA.

[Ferreira, 1988] Ferreira, F. (1988). *Planning and timing in sentence production: The syntax-to-phonology conversion*. Phd. thesis, University of Massachusetts.

[Festival, 2010] Festival (2010). Festival Speech Synthesis System. http://www.cstr.ed.ac.uk/projects/festival/.

[Fineberg, 2006] Fineberg, J. (2006). Lolita. http://brahms.ircam.fr/works/work/18304/.

[Fletcher and Munson, 1933] Fletcher, H. and Munson, W. A. (1933). Loudness, its definition, measurement, and calculation. *Journal of the Acoustical Society of America*, 5:82–108.

[Fonagy, 1983] Fonagy, I. (1983). *La vive voix: Essais de psycho-phonétique*. Payot, Paris.

[Forney, 1973] Forney, D. (1973). The Viterbi algorithm. *Proceeding of the IEEE*, 61(3):268–278.

[Fougeron, 1998] Fougeron, C. (1998). *Variations articulatoires en début de constituants prosodiques de différents niveaux en français*. PhD. thesis, Paris III, Sorbonne Nouvelle.

[Fraisse, 1974] Fraisse, P. (1974). *Psychologie du rythme*. Presses Universitaires de France.

[Fujisaki, 1981] Fujisaki, H. (1981). Dynamic characteristics of voice fundamental frequency in speech and singing. acoustical analysis and physiological interpretations. Technical report, K.T.H. Quarterly Progress Report and Status Progress. Departement for Speech, Music and Hearing.

[Fujisaki, 1983] Fujisaki, H. (1983). *The Production of Speech*, chapter Dynamic characteristics of voice fundamental frequency in speech and singing, pages 39–55. Springer, New York.

[Fukada et al., 1994] Fukada, T., Komori, Y., Aso, T., and Ohora, Y. (1994). A study of pitch pattern generation using HMM-based statistical information. In *International Conference on Spoken Language Processing*, pages 723–726, Genove, Italy.

[Gales and Young, 1993] Gales, M. and Young, S. (1993). Segmental HMMs for speech recognition. In *European Conference on Speech Communication and Technology*, pages 1579–1582, Berlin, Germany.

[Gallese, 2003] Gallese, V. (2003). The Roots of Empathy: The Shared Manifold Hypothesis and the Neural Basis of Intersubjectivity. *Psychopathology*, 36:171–180.

[Gao et al., 2008] Gao, B., Qian, Y., Wu, Z., and Soong, F. (2008). Duration refinement by jointly optimizing state and longer unit likelihood. In *Interspeech*, pages 2266–2269, Brisbane, Australia.

[Gee and Grosjean, 1983] Gee, J. and Grosjean, F. (1983). Performance structures: a psycholinguistic and linguistic appraisal. *Cognitive Psychology*, 15:411–458.

[Geman and Geman, 1984] Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741.

[Genette et al., 1986] Genette, G., Todorov, T., Jaussand, H.-R., Schaeffer, J.-M., Scholes, R., Stempel, W. D., and Vitor, K. (1986). *Théorie des Genres*. Seuil, Paris.

[Gervasoni, 2008] Gervasoni, S. (2007-2008). Com que voz. http://brahms.ircam.fr/works/work/19824/.

[Gilles and Peters, 2004] Gilles, P. and Peters, J. (2004). *Regional Variations in Intonation*. Tbingen.

[Giustiniani and Pierucci, 1991] Giustiniani, M. and Pierucci, P. (1991). Phonetic ergodic HMM for speech synthesis. In *European Conference on Speech Communication and Technology*, pages 349–352, Genove, Italy.

[Grabe, 1998] Grabe, E. (1998). *Comparative intonational phonology: English and German*. Phd. thesis, Planck Institut fur Psycholinguistik, Nijmegen, and University of Nijmegen.

[Grabe et al., 1994] Grabe, E., Kochanski, G., and Coleman, J. (1994). Quantitative modelling of intonational variation. In *Speech Analysis and Recognition in Technology, Linguistics and Medicine*, pages 1–23.

[Grabe et al., 2001] Grabe, E., Post, B., and Nolan, F. (2001). Modelling intonational variation in english: The IViE system. In *Prosody 2000: Speech Recognition and Synthesis*, pages 51–58, Poznan , Poland.

[Grefenstette and Tapanainen, 1994] Grefenstette, G. and Tapanainen, P. (1994). What is a word, what is a sentence? problems of tokenization. In *Conference on Computational Lexicography and Text Research*, pages 79–87, Budapest, Hungary.

[Gussenhoven, 1984] Gussenhoven, C. (1984). *On the grammar and semantics of sentence accents*. Foris, Dordrecht.

[Gussenhoven, 2004] Gussenhoven, C. (2004). *The phonology of tone and intonation*. Cambridge University Press, Cambridge.

[Halliday, 1985] Halliday, M. (1985). *Spoken and written Language*. Oxford University Press, Oxford.

[Hamon et al., 1989] Hamon, C., Mouline, E., , and Charpentier, F. (1989). A diphone synthesis system based on time-domain prosodic modifications of speech. In *International Conference on Acoustics, Speech, and Signal Processing*, page 238–241, Glasgow , UK.

[Hashimoto, 1987] Hashimoto, T. (1987). A list-type reduced-constraint generalization of the Viterbi algorithm. *IEEE Transactions on Information Theory*, 33(6):866–876.

[Hegel, 1835] Hegel, G. W. F. (1835). *Lectures on Aesthetics*. Heinrich Gustav Hotho, Berlin.

[Hermes, 1987] Hermes, D. (1987). Vowel-onset detection. *IPO-APR*, 22:15–24.

[Hirschberg, 1991] Hirschberg, J. (1991). Using text analysis to predict intonational boundaries. In *European Conference on Speech Communication and Technology*, pages 1275–1278, Genova, Italy.

[Hirst and Di Cristo, 1998] Hirst, D. and Di Cristo, A. (1998). *Intonation Systems: a survey of twenty languages*. Cambridge University Press, Cambridge.

[Hirst et al., 2000] Hirst, D., Di Cristo, A., and Espresser, R. (2000). *Prosody: Theory and Experiments*, chapter Levels of representation and levels of analysis for the description of intonation systems. M. Horne.

[Hirst and Espesser, 1993] Hirst, D. and Espesser, R. (1993). Automatic modelling of fundamental frequency using a quadratic spline function. In *Travaux de l'Institut de Phonétique d'Aix*, volume 15, pages 71–85.

[Holm, 2003] Holm, B. (2003). *SFC : un modèle de superposition de contours multiparamétriques pour la génération automatique de la prosodie - Apprentissage automatique et application l'énonciation de formules mathématiques*. PhD. thesis, Institut de la Communication Parlée, Grenoble.

[House, 1990] House, D. (1990). *Tonal Perception in Speech*. MIT Press, Lund University Press.

[House, 1995] House, D. (1995). The influence of silence on perceiving the preceding tonal contour. In *International Congress of Phonetic Sciences*, pages 122–125, Stockholm, Sweden.

[HTS, 2010] HTS (2010). HMM-based Speech Synthesis System (HTS). http://www.text2speech.com.

[Hunt and Black, 1996] Hunt, A. J. and Black, A. W. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. In *International Conference on Acoustics, Speech, and Signal Processing*, page 373–376, Atlanta, USA.

[Imai, 1983] Imai, S. (1983). Cepstral analysis synthesis on the mel frequency scale. In *International Conference on Speech, Signal, and Audio Processing*, pages 93–96, Boston, USA.

[Ingulfen et al., 2005] Ingulfen, T., Burrows, T., and Buchholz, S. (2005). Influence of syntax on prosodic boundary prediction. In *Interspeech*, pages 1817–1820, Lisboa, Portugal.

[International Telecommunication Union, 1996] International Telecommunication Union (1996). *Methods for subjective determination of transmission quality*. Geneva, Switzerland.

[Joshi et al., 1975] Joshi, A., Levy, L., and Takahashi, M. (1975). Tree adjunct grammars. *Journal of the Computer and System Sciences*, 10(1):136–163.

[Jun, 1993] Jun, S.-A. (1993). *The Phonetics and Phonology of Korean Prosody*. PhD thesis, Ohio State University.

[Jun, 2005] Jun, S.-A. (2005). *Prosodic Typology: The Phonology of Intonation and Phrasing: The Phonology of Intonation and Phrasing*. Oxford University Press, Oxford.

[Kabal, 1999] Kabal, P. (1999). Measuring speech activity. Technical report, MMSP Lab Technical Report, Dept. Electrical & Computer Engineering, McGill University.

[Kawahara et al., 1999a] Kawahara, H., Katayose, H., De Cheveigné, A., and Patterson, R. D. (1999a). Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of f0 and periodicity. In *European Conference on Speech Communication and Technology*, pages 2781–2784, Budapest, Hungary.

[Kawahara et al., 1999b] Kawahara, H., Masuda-Katsuse, I., and de Cheveigné, A. (1999b). Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. *Speech Communication*, 27:187–207.

[Klatt, 1987] Klatt, D. (1987). Review of Text-to-Speech Conversion for English. *Journal of the Acoustic Society of America*, 82(3):737–793.

[Kloker, 1976] Kloker, D. (1976). A technique for the automatic location and description of pitch contours. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 55–58, Philadelphia, USA.

[Koch and Oesterreicher, 2001] Koch, P. and Oesterreicher, W. (2001). *Lexikon der Romanistischen Linguistik*, chapter Langage parlé et langage écrit, pages 584–627. Niemeyer, Tbingen.

[Kratzer, 1981] Kratzer, A. (1981). *Words, Worlds and Contexts*, chapter The notional category of modality, pages 38–74. Walter de Gruyter.

[Krstulović et al., 2007] Krstulović, S., Hunecke, A., and Schröder, M. (2007). An HMM-based speech synthesis system applied to german and its adaptation to a limited set of expressive football announcements. In *Interspeech*.

[Kundu and Bahl, 1988] Kundu, A. and Bahl, P. (1988). Recognition of Handwritten Script: A Hidden Markov Model Based Approach. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 928–931, New-York, USA.

[Labov, 1972] Labov, W. (1972). *Sociolinguistic Patterns*. University of Pennsylvania Press, Philadelphia.

[Labov et al., 1968] Labov, W., Uriel, W., and Herzog, M. (1968). *Directions for Historical Linguistics*, chapter Empirical foundations for a theory of language change, pages 95–189. University of Texas Press, Austin.

[Lacheret et al., 2010] Lacheret, A., Obin, N., and Avanzi, M. (2010). Design and Evaluation of Shared Prosodic Annotation for Spontaneous French Speech: From Expert Knowledge to Non-Expert Annotation. In *Linguistic Annotation Workshop*, pages 265–273, Uppsala, Sweden.

[Lacheret et al., 2009] Lacheret, A., Victorri, B., and Avanzi, M. (2009). Schématisation discursive et schématisation intonative : question de genre ? To be published.

[Lacheret-Dujour and Beaugendre, 1999] Lacheret-Dujour, A. and Beaugendre, F. (1999). *La prosodie du français*. CNRS.

[Ladd, 1983] Ladd, D. (1983). Phonological features of intonational peaks. *Language*, 59:721–759.

[Ladd, 1996] Ladd, R. D. (1996). *Intonational Phonology*. Cambridge University Press, Cambridge.

[Ladd and Campbell, 1991] Ladd, R. D. and Campbell, N. (1991). Theories of prosodic structure: evidence from syllable duration. In *International Congress of Phonetic Sciences*, pages 290–293, Aix-en-Provence, France.

[Lanchantin et al., 2011] Lanchantin, P., Farner, S., Veaux, C., Degottex, G., Obin, N., Beller, G., Villavicencio, F., Hueber, T., Schwartz, D., Huber, S., Peeters, G., Roebel, A., and Rodet, X. (2011). Vivos Voco: A Survey of Recent Research on Voice Transformations at IRCAM. In *International Conference on Digital Audio Effects (DAFx)*, pages 277–285, Paris, France.

[Lanchantin et al., 2008] Lanchantin, P., Morris, A., Rodet, X., and Veaux, C. (2008). Automatic phoneme segmentation with relaxed textual constraints. In *International Conference on Language Resources and Evaluation*, pages 2403–2407, Marrakech, Morroco.

[Lanza and Pasquet, 2009] Lanza, M. and Pasquet, O. (2009). Häxan, la sorcellerie à travers les âges. http://brahms.ircam.fr/works/work/23986/.

[Latorre and Akamine, 2008] Latorre, J. and Akamine, M. (2008). Multilevel parametric-base F0 model for speech synthesis. In *Interspeech*, pages 2274–2277, Brisbane, Australia.

[Lee, 1988] Lee, K.-F. (1988). On large-vocabulary speaker-independent continuous speech recognition. *Speech Communication*, 7(4):375–379.

[Leggetter and Woodland, 1995] Leggetter, C. J. and Woodland, P. C. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer Speech & Language*, 9(2):171–185.

[Levinson, 1986] Levinson, S. (1986). Continuously variable duration hidden Markov models for automatic speech recognition. *Computer Speech and Language*, 1:29–45.

[Liberman, 1975] Liberman, M. (1975). *The Intonation System of English*. PhD. thesis, Massachusetts Institute of Technology.

[Lindblom, 1983] Lindblom, B. (1983). *The Production of Speech*, chapter Economy of Speech Gestures. Spinger-Verlag.

[Ljolje and Fallside, 1986] Ljolje, A. and Fallside, F. (1986). Synthesis of natural sounding pitch contous in isolated utterances using hidden Markov models. In *IEEE Transactions on Acoustic, Speech, and Signal Processing*, volume ASSP-34, pages 1074–1080.

[Lolive et al., 2006] Lolive, D., Barbot, N., and Boëffard, O. (2006). Melodic contour estimation with B-spline models using a MDL criterion. In *International Conference on Speech and Computer*, pages 333–338, Saint Petersburg, Russia.

[Loquendo, 2010] Loquendo (2010). Loquendo Speech Synthesis System. http://www.loquendo.com/en/demo-center/interactive-tts-demo.

[Léon, 1993] Léon, P. (1993). *Précis de Phonostylistique - Parole et Expressivité*. Nathan, Paris.

[Maeda, 1974] Maeda, S. (1974). A characterization of fundamental frequency contours of speech. Technical report, MIT Research Laboratory of Electronics. Quarterly Progress Report.

[Manning and Schütze, 1999] Manning, C. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge.

[Martin, 1975] Martin, P. (1975). Analyse phonologique de la phrase francaise. *Linguistics*, 146:35–67.

[Martin, 1987] Martin, P. (1987). Prosodic and rhythmic structures in french. *Linguistics*, 25(5):925–950.

[Martin, 2010] Martin, P. (2010). Prosodic structure revisited: a cognitive approach. In *Speech Prosody*, Chicago, USA.

[Martinet, 1956] Martinet, A. (1956). *La description phonologique*. Droz, Paris-Geneve.

[Matsui and Furui, 1992] Matsui, T. and Furui, S. (1992). Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMMs. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 157–160, San Francisco, USA.

[Mehrabian, 1972] Mehrabian, A. (1972). *Nonverbal Communication*. Walter De Gruyter, New-York.

[Mertens, 1987] Mertens, P. (1987). *L'intonation du français. De la description linguistique à la reconnaissance automatique.* PhD. thesis, Université de Leuven.

[Mertens, 2004a] Mertens, P. (2004a). The prosogram : Semi-automatic transcription of prosody based on a tonal perception model. In *Speech Prosody*, pages 549–552, Nara, Japan.

[Mertens, 2004b] Mertens, P. (2004b). Quelques allers-retours entre la prosodie et son traitement automatique. *Le français moderne*, 72(1):39–57.

[Mishra et al., 2006] Mishra, T., Van Santen, J., , and Klabbers, E. (2006). Decomposition of Pitch Curves in the General Superpositional Intonation Model. In *Speech Prosody*, Dresden, Germany.

[Mokhtari and Campbell, 2003] Mokhtari, P. and Campbell, N. (2003). Automatic measurement of pressed/breathy phonation at acoustic centres of reliability in continuous speech. *IEICE Transactions on Information and Systems*, 86(3):574–582.

[Monnin and Grosjean, 1993] Monnin, P. and Grosjean, F. (1993). Les structures de performance en français : caractérisation et prédiction. *L'année psychologique*, 93:9–30.

[Morlec, 1997] Morlec, Y. (1997). *Génération multiparamétrique de la prosodie du français par apprentissage automatique*. PhD. thesis, Institut de la Communication Parlée, Grenoble.

[Morton et al., 1976] Morton, J., Marcus, S., and Frankish, C. (1976). Perceptual centers (p-centers). *Psychological Review*, 83:405–408.

[Nicholson et al., 2003] Nicholson, H., Bard, E. G., Lickley, R., Anderson, A. H., Mullin, J., Kenicer, D., and Smallwood, L. (2003). The intentionality of disfluency: Findings from feedback and timing. In *Disfluency in Spontaneous Speech*, pages 17–20.

[Nogueiras et al., 2001] Nogueiras, A., Moreno, A., Bonafante, A., , and Maririo, J. (2001). Speech Emotion Recognition Using Hidden Markov Models. In *European Conference on Speech Communication and Technology*, pages 2679–2682, Genova, Italy.

[Nuance, 2010] Nuance (2010). RealSpeak Speech Synthesis System. http://www.nuance.com/realspeak/demo/.

[Nwe et al., 2003] Nwe, T., Foo, S., and Silva, L. (2003). Speech emotion recognition using hidden Markov models. *Speech communication*, 41(4):603–623.

[Obin, 2010] Obin, N. (2010). Modélisation du Style en Synthèse de la Parole. In *Journées Jeunes Chercheurs en Audition, Acoustique musicale et Signal audio*, Paris, France.

[Obin et al., 2011a] Obin, N., Avanzi, M., and Lacheret, A. (2011a). Transcription of French Prosody in Discourse: the Rhapsodie Protocole. In *Interface Discours Prosodie*, Manchester, U.K.

[Obin et al., 2008a] Obin, N., Goldman, J.-P., Avanzi, M., and Lacheret-Dujour, A. (2008a). Comparaison de trois outils de détection automatique de proéminences en français parlé. In *Journées d'Etude de la Parole*, pages 85–88, Avignon, France.

[Obin et al., 2010a] Obin, N., Lacheret, A., and Rodet, X. (2010a). HMM-based Prosodic Structure Model Using Rich Linguistic Context. In *Interspeech*, pages 1133–1136, Makuhari, Japan.

[Obin et al., 2011b] Obin, N., Lacheret, A., and Rodet, X. (2011b). Stylization and Trajectory Modelling of Short and Long Term Speech Prosody Variations. In *Interspeech*, pages 2029–2032, Florence, Italy.

[Obin et al., 2008b] Obin, N., Lacheret, A., Veaux, C., Rodet, X., and Simon, A.-C. (2008b). A Method for Automatic and Dynamic Estimation of Discourse Genre Typology with Prosodic Features. In *Interspeech*, pages 1204–1207, Brisbane, Australia.

[Obin et al., 2010b] Obin, N., Lanchantin, P., Lacheret, A., and Rodet, X. (2010b). Towards Improved HMM-based Speech Synthesis Using High-Level Syntactical Features. In *Speech Prosody*, Chicago, U.S.A.

[Obin et al., 2011c] Obin, N., Lanchantin, P., Lacheret, A., and Rodet, X. (2011c). Discrete/Continuous Modelling of Speaking Style in HMM-based Speech Synthesis: Design and Evaluation. In *Interspeech*, pages 2785–2788, Florence, Italy.

[Obin et al., 2011d] Obin, N., Lanchantin, P., Lacheret, A., and Rodet, X. (2011d). Reformulating Prosodic Break Model into Segmental HMMs and Information Fusion. In *Interspeech*, pages 1829–1832, Florence, Italy.

[Obin et al., 2011e] Obin, N., Lanchantin, P., Lacheret, A., and Rodet, X. (2011e). Symbolic Modelling of Speech Prosody: From Linguistics to Statistical Modelling. *IEEE Transactions on Audio, Speech, and Language Processing*, page Submitted.

[Obin et al., 2008c] Obin, N., Rodet, X., and Lacheret-Dujour, A. (2008c). French Prominence: a Probabilistic Framework. In *International Conference on Audio, Speech, and Signal Processing*, pages 3993–3996, Las Vegas, U.S.A.

[Obin et al., 2008d] Obin, N., Rodet, X., and Lacheret-Dujour, A. (2008d). Un modèle de durée des syllabes fondé sur les propriétés syllabiques intrinsèques et les variations locales de débit. In *Journées d'Etude de la Parole*, pages 333–336, Avignon, France.

[Obin et al., 2009a] Obin, N., Rodet, X., and Lacheret-Dujour, A. (2009a). A Multi-Level Context-Dependent Prosodic Model Applied To Durational Modeling. In *Interspeech*, pages 512–515, Brighton, U.K.

[Obin et al., 2009b] Obin, N., Rodet, X., and Lacheret-Dujour, A. (2009b). A Syllable-Based Prominence Model Based On Discriminant Analysis And Context-Dependency. In *International Conference on Speech and Computer*, pages 97–100, St-Petersburg, Russia.

[Obin et al., 2010c] Obin, N. Dellwo, V., Lacheret, A., and Rodet, X. (2010c). Expectations for Discourse Genre Identification: a Prosodic Study. In *Interspeech*, pages 3070–3073, Makuhari, Japan.

[O'Dell, 1995] O'Dell, J. (1995). *The Use of Context in Large Vocabulary Speech Recognition*. PhD. thesis, Cambridge University.

[Ohala, 1996] Ohala, J. (1996). Ethological theory and the expression of emotion in the voice. In *International Conference on Spoken Language Processing*, pages 1812–1815, Philadelphia, USA.

[Ohala, 1993] Ohala, J. J. (1993). *Historical linguistics: problems and perspectives.*, chapter The Phonetics of Sound Change, pages 237–278. Longman, London.

[Ohno and Fujisaki, 1995] Ohno, S. and Fujisaki, H. (1995). A method for quantitative analysis of the local speech rate. In *European Conference on Speech Communication and Technology*, pages 421–424, Madrid, Spain.

[Orange Labs, 2010] Orange Labs (2010). Orange Labs Speech Synthesis System. http://tts.elibel.tm.fr/tts.

[Ostendorf et al., 1996] Ostendorf, M., Digalakis, V., and Kimball, O. (1996). From hmm's to segment models: a unified view of stochastic modeling for speech recognition. *IEEE Transactions on Speech and Audio Processing*, 4(5):360–378.

[Ostendorf and Veilleux, 1994] Ostendorf, M. and Veilleux, N. (1994). A hierarchical stochastic model for automatic prediction of prosodic boundary location. *Journal of Computational Linguistics*, 20(1):27–54.

[Paroubek et al., 2008] Paroubek, P., Robba, I., Vilnat, A., and Ayache, C. (2008). EASY, Evaluation of Parsers of French: what are the results? In *International Conference on Language Resources and Evaluation*, pages 2480–2486, Marrakech, Marocco.

[Parra, 2009] Parra, H. (2008-2009). Hypermusic: Prologue. http://brahms.ircam.fr/works/work/23852/.

[Pasdeloup, 1992] Pasdeloup, V. (1992). *Talking Machines. Theories, Models, and Designs*, chapter A prosodic model for French text-to-speech synthesis: A psycholinguistic approach., pages 335–348. Elsevier Science Publishers.

[Peeters, 2001] Peeters, G. (2001). *Modèles et modélisation du signal sonore adaptés à ses caractéristiques locales*. PhD. thesis, Université PAris VI.

[Peeters, 2002] Peeters, G. (2002). Pourquoi Gérard Depardieu parle anglais sans accent. *La Recherche*, 358:98–99.

[Peeters, 2003] Peeters, G. (2003). Automatic classification of large musical instrument databases using hierarchical classifiers with inertia ratio maximization. In *Audio Engineering Society Convention*, volume 115, pages 5959–5963, New-York, USA.

[Perrault, 1697] Perrault, C. (1697). *Le Petit Poucet*. Barbin.

[Pfitzinger, 1998] Pfitzinger, H. (1998). Local speech rate as a combination of syllable and phone rate. In *International Conference on Spoken Language Processing*, pages 1087–1090, Sydney, Australia.

[Pfitzinger, 2006] Pfitzinger, H. R. (2006). Five dimensions of prosody: intensity, intonation, timing, voice quality, and degree of reduction. In *Speech Prosody*, Dresden, Germany. Keynote.

[Pierrehumbert, 1980] Pierrehumbert, J. (1980). *The phonology and phonetics of English Intonation*. Phd. thesis, Massachusett Institute of Technology.

[Pierrehumbert and Hirschberg, 1990] Pierrehumbert, J. and Hirschberg, J. (1990). *Intention in communication*, chapter The meaning of intonation in the interpretation of discourse, page 271–311. MIT Press.

[Plato, 0 BC] Plato (380 BC). *The Republic*.

[Post, 2000] Post, B. (2000). *Tonal and phrasal structures in French intonation*. Academic Graphics, The Hague.

[Post et al., 2006] Post, B., Delais-Roussarie, E., and Simon, A.-C. (2006). IVTS, un système de transcription pour la variation prosodique. *Bulletin de la Phonologie du Français Contemporain*, 6:51–68.

[Price et al., 1991] Price, P. J., Ostendorf, M., Shattuck-Hufnagel, S., and Fong, G. (1991). The use of prosody in syntactic disambiguation. *Journal of the Acoustical Society of America*, 90:2956–2970.

[Proust, 1913] Proust, M. (1913). *Du côté de chez Swann*. Grasset.

[Proust, 1927] Proust, M. (1927). *Le Temps Retrouvé*. La Nouvelle Revue Française.

[Qian et al., 2009] Qian, Y., Wu, Z., and Soong, F. K. (2009). Improved prosody generation by maximizing joint likelihood of state and longer units. In *International Conference on Acoustics, Speech and Signal Processing*, pages 3781–3784, Taipei, Taiwan.

[Qin et al., 2009] Qin, L., Wu, Y.-J., Ling, Z.-H., Wang, R.-H., and Dai, L.-R. (2009). Minimum generation error criterion considering global/local variance for HMM-based speech synthesis. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 4621 – 4624, Las Vegas, USA.

[Rabiner, 1989] Rabiner, L. R. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, 77(62):257–286.

[Rangarajan Sridhar et al., 2008] Rangarajan Sridhar, V., Bangalore, S., and Narayanan, S. (2008). Exploiting acoustic and syntactic features for automatic prosody labeling in a maximum entropy framework. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(4):797–811.

[Rastier, 1989] Rastier, F. (1989). *Sens et textualité*. Hachette, Paris.

[Reynolds and Carlson, 1995] Reynolds, D. and Carlson, B. (1995). Text-dependent speaker verification using decoupled and integrated speaker and speech recognizers. In *European Conference on Speech Communication and Technology*, pages 157–160, Madrid, Spain.

[Rissanen, 1984] Rissanen, J. (1984). Universal coding, information, prediction, and estimation. *IEEE transaction on Information Theory*, 30(4):629–636.

[Rodet, 1977] Rodet, X. (1977). *Analyse du signal vocal dans sa représentation amplitude-temps. Synthèse de la parole par règles*. PhD. thesis, Université Paris VI.

[Rodet et al., 2009] Rodet, X., Beller, G., Bogaards, N., Degottex, G., Farner, S., Lanchantin, P., Obin, N., Roebel, A., Veaux, C., and Villavicencio, F. (2009). *Parole et musique*, chapter Transformation et synthèse de la voix parlée et de la voix chantée. Odile Jacob.

[Rodet et al., 1984] Rodet, X., Potard, Y., and Barrière, J. (1984). The CHANT Project: From the Synthesis of the Singing Voice to Synthesis in General. *Computer Music Journal*, 8(3):15–31.

[Rohmer, 2007] Rohmer, E. (2007). Les Amours d'Astrée et de Céladon. http://www.imdb.fr/title/tt0823240/.

[Ross and Ostendorf, 1996] Ross, K. and Ostendorf, M. (1996). Prediction of abstract prosodic labels for speech synthesis. *Computer Speech and Langage*, 10:155–185.

[Rousseau, 1781] Rousseau, J.-J. (1781). *Essai sur l'Origine des Langues*.

[Russel and Moore, 1985] Russel, M. and Moore, R. (1985). Explicit modeling of state occupancy in hidden Markov models for automatic speech recognition. In *International Conference on Acoustic, Speech, and Signal Processing*, pages 2376–2379, Tempa, USA.

[Sagot, 2010] Sagot, B. (2010). The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French. In *International Conference on Language Ressources and Evaluation*, pages 2744–2751, Valletta, Malte.

[Sagot and Boullier, 2005] Sagot, B. and Boullier, P. (2005). From Raw Corpus to Word Lattices: Robust Pre-parsing Processing with SxPipe. 15(4):653–662.

[Sagot et al., 2006] Sagot, B., Clément, L., Villemonte de La Clergerie, E., and Boullier, P. (2006). The Lefff 2 syntactic lexicon for French: architecture, acquisition, use. In *International Conference on Language Ressources and Evaluation*, pages 1348–1351, Genova, Italy.

[Scherer et al., 1991] Scherer, K. R., Banse, R., Wallbott, H. G., and Goldbeck, T. (1991). Vocal cues in emotion encoding and decoding. *Motivation and Emotion*, 15:123–148.

[Schmid and Atterer, 2004] Schmid, H. and Atterer, M. (2004). New statistical methods for phrase break prediction. In *International Conference On Computational Linguistics*, pages 659–665, Geneva, Switzerland.

[Schober and Brennan, 2001] Schober, M. F. and Brennan, S. E. (2001). Disfluency rate in conversation: effects of age, relationship, topic, role and gender. *Language and Speech*, 44:123–147.

[Schramm, 1954] Schramm, W. (1954). *The Process and Effects of Communication*, chapter How Communication Works, pages 3–26. University of Illinois Press.

[Schröder and Trouvain, 2003] Schröder, M. and Trouvain, J. (2003). The german text-to-speech synthesis system MARY: A tool for research, development and teaching. *International Journal of Speech Technology*, 6:365–377.

[Schwarz, 2003] Schwarz, D. (2003). The Caterpillar System for Data-Driven Concatenative Sound Synthesis. In *Digital Audio Effects (DAFx)*, pages 135–140.

[Schwarz, 1978] Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464.

[Scott, 1993] Scott, S. (1993). *P-centers in Speech: An Acoustic Analysis*. PhD. thesis, University College of London.

[Searle, 1969] Searle, J. (1969). *Speech Acts*. Cambridge University Press.

[Selrik, 1981] Selrik, E. (1981). On prosodic structure and its relation to syntactic structure. In *Nordic Prosody II*, pages 111–140, Trondheim, Norway.

[Selrik, 1984] Selrik, E. (1984). *Phonology and Syntax: The Relation between Sound and Structure*. MIT Press, Cambridge.

[Shafer, 1976] Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton University Press.

[Shannon, 1948] Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423.

[Shattuck-Hufnagel and Turk, 1996] Shattuck-Hufnagel, S. and Turk, A. E. (1996). A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguistic Research*, 25:193–247.

[Shichiri et al., 2002] Shichiri, K., Sawabe, A., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. (2002). Eigenvoices for HMM-based speech synthesis. In *International Conference on Spoken Language Processing*, pages 1269–1272, Denver, Colorado.

[Shinoda and Watanabe, 2000] Shinoda, K. and Watanabe, T. (2000). MDL-based context-dependent subword modeling for speech recognition. *Journal of the Acoustical Society of Japan*, 21(2):79–86.

[Shochi et al., 2009] Shochi, T., Rilliard, A., Aubergé, V., and Erickson, D. (2009). *The role of prosody in Affective Speech*, chapter Intercultural Perception of English, French and Japanese Social Affective Prosody, pages 31–59. Peter Lang.

[Shuang et al., 2009] Shuang, Z., Kang, S., Shi, Q., Qin, Y., and Cai, L. (2009). Syllable HMM based Mandarin TTS and Comparison with Concatenative TTS. In *Interspeech*, pages 1767–1771, Brighton, UK.

[Silverman et al., 1992] Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., and Hirschberg, J. (1992). ToBI: a standard for labeling english prosody. In *International Conference of Spoken Language Processing*, pages 867–870, Banff, Canada.

[Simon et al., 2009] Simon, A.-C., Auchlin, A., Avanzi, M., and Goldman, J.-P. (2009). *Les voix des Français*, chapter Les phonostyles: une description prosodique des styles de parole en français. Peter Lang.

[SoftVoice, 2010] SoftVoice (2010). SoftVoice Speech Synthesis System. http://www.text2speech.com.

[Sreenivasa Rao and Yegnanarayana, 2007] Sreenivasa Rao, K. and Yegnanarayana, B. (2007). Modeling durations of syllables using neural networks. *Computer Speech and Language*, 21(2):282–295.

[Sun and Applebaum, 2001] Sun, X. and Applebaum, T. H. (2001). Intonational Phrase Break Prediction Using Decision Tree and N-Gram Model. In *European Conference on Speech Communication and Technology*, pages 3–7, Aalborg, Denmark.

[SVOX, 2010] SVOX (2010). SVOX Speech Synthesis System. http://www.tik.ee.ethz.ch/cgi-bin/w3svox.

[Syrdal and McGory, 2000] Syrdal, A. K. and McGory, J. (2000). Inter-transcriber reliability of ToBI prosodic labeling. In *International Conference on Spoken Language Processing*, pages 235–238, Beijing, China.

[Syrdal et al., 2001] Syrdal, A., K., Hirschberg, J., McGory, J., and Beckman, M. (2001). Automatic ToBI prediction and alignment to speed manual labeling of prosody. *Speech Communication*, 33(1-2):135–151.

['t Hart, 1991] 't Hart, J. (1991). F0 stylization in speech : straight lines versus parabolas. *Journal of the Acoustical Society of America*, 90(6):3368–3370.

['t Hart et al., 1990] 't Hart, J., Collier, R., and Cohen, A. (1990). *A perceptual study of intonation*. MIT Press, Cambridge University Press.

[Tachibana et al., 2005] Tachibana, M., Yamagishi, J., Masuko, T., and Kobayashi, T. (2005). Speech synthesis with various emotional expressions and speaking styles by style interpolation and morphing. *IEICE Transaction on Information. and Systems*, E88-D(11):2484–2491.

[Taylor, 1994] Taylor, P. (1994). The Rise/Fall/Connection Model of intonation. *Speech Communication*, 15:169–186.

[Taylor, 1998] Taylor, P. (1998). The TILT intonation model. In *International Conference on Spoken Language Processing*, pages 1383–1386, Sydney, Australia.

[Taylor, 2000] Taylor, P. (2000). Analysis and synthesis of intonation using the TILT model. *Journal of the Acoustic Society of America*, 107:1697–1714.

[Teppereman and Narayanan, 2008] Teppereman, J. and Narayanan, S. (2008). Tree grammars as models of prosodic structure. In *Interspeech*, pages 2286–2289, Brisbane, Australia.

[Teutenberg et al., 2008] Teutenberg, J., Watson, C., and Riddle, P. (2008). Modelling and Synthesising F0 contours with the Discrete Cosine Transform. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 3973–3976, Las Vegas, U.S.A.

[The Euphonia, 1846] The Euphonia (1846). The Euphonia, or Speaking Automaton. *Illustrated London News*, 9:59.

[Toda and Tokuda, 2007] Toda, T. and Tokuda, K. (2007). A speech parameter generation algorithm considering global variance for HMM-based speech synthesis. *IEICE Transactions on Information and Systems*, 90(5):816–824.

[Toda and Young, 2009] Toda, T. and Young, S. (2009). Trajectory training considering global variance for HMM-based speech synthesis. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 4025–4028, Taipei, Taiwan.

[Todorov, 1978] Todorov, T. (1978). *Les genres de discours*. Seuil, Paris.

[Tokuda et al., 1995] Tokuda, K., Kobayashi, T., and Imai, S. (1995). Speech parameter generation from HMM using dynamic features. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 660–663, Detroit, USA.

[Tokuda et al., 1999] Tokuda, K., Masuko, T., Miyazaki, N., and Kobayashi, T. (1999). Hidden Markov models based on multi-space probability distribution for pitch pattern modeling. In *International Conference on Audio, Speech, and Signal Processing*, pages 229–232, Phoenix, Arizona.

[Tokuda et al., 2000] Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., and Kitamura, T. (2000). Speech parameter generation algorithms for HMM-based speech synthesis. In *International Conference on Audio, Speech, and Signal Processing*, pages 1315–1318, Istanbul, Turkey.

[Tokuda et al., 2003] Tokuda, K., Zen, H., and Kitamura, T. (2003). Trajectory modeling based on HMMs with the explicit relationship between static and dynamic features. In *European Conference on Speech Communication and Technology*, pages 865–868, Geneva, Switzerland.

[Tseng and Lee, 2004] Tseng, C. and Lee, Y. (2004). Intensity in relation to prosody organization. In *International Symposium on Chinese Spoken Language Processing*, pages 217–220, Hong-Kong, China.

[Van Rijsbergen, 1979] Van Rijsbergen, C. (1979). *Information retrieval*. Butterworths, London.

[Van Santen and Moebius, 1999] Van Santen, J. and Moebius, B. (1999). *Intonation Analysis, Modelling and Technology*, chapter A quantitative model of f0 generation and alignment, pages 269–288. Kluwer Academic, Netherlands.

[Veilleux et al., 1990] Veilleux, N., Ostendorf, M., Price, P., and Shattuck-Hufnagel, S. (1990). Markov modeling of prosodic phrase structure. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 777–780, Albuquerque, USA.

[Venditti, 1995] Venditti, J. (1995). Japanese ToBI labelling guidelines. Technical report, Ohio State University. http://web.mac.com/jen.venditti/iWeb/Site/Japanese%20ToBI_files/jtobi_oct_1.pdf.

[Vettin and Todt, 2004] Vettin, J. and Todt, D. (2004). Laughter in conversation: Features of occurrence and acoustic structure. *Journal of Nonverbal Behavior*, 28(2):93–115.

[Villemonte de La Clergerie, 2005a] Villemonte de La Clergerie, E. (2005a). DyALog: a Tabular Logic Programming based environment for NLP. In *Workshop on Constraint Satisfaction for Language Processing*, Barcelona, Spain.

[Villemonte de La Clergerie, 2005b] Villemonte de La Clergerie, E. (2005b). From metagrammars to factorized TAG/TIG parsers. In *International Workshop On Parsing Technology*, pages 190–191, Vancouver, Canada.

[Villemonte de La Clergerie, 2010] Villemonte de La Clergerie, E. (2010). Convertir des dérivations TAG en dépendances. In *Traitement Automatique des Langues Naturelles*, Montréal, Canada.

[Villemonte de La Clergerie et al., 2008] Villemonte de La Clergerie, E., Hamon, O., Mostefa, D., Ayache, C., Paroubek, P., and Vilnat, A. (2008). PASSAGE : from French Parser Evaluation to Large Sized Treebank. In *International Conference on Language Resources and Evaluation*, pages 3570–3577, Marrakech, Morocco.

[Vogel et al., 1996] Vogel, S., Ney, H., and Tillmann, C. (1996). HMM based Word Alignment in Statistical Translation. In *International Conference on Computational Linguistics*, pages 836–841, Copenhagen, Denmark.

[von Kempelen, 1791] von Kempelen, W. (1791,). *Mechanismus der menschlichen Sprache nebst der Beschreibung seiner sprechenden Maschine*. J. B. Degen, Wien.

[Wagner, 2005] Wagner, P. (2005). Great expectations - introspective vs. perceptual prominence ratings and their acoustic correlates. In *Interspeech*, pages 2381–2384, Lisbon, Portugal.

[Walter, 1982] Walter, H. (1982). *Enquête phonologique et variétés régionales du français*. Presses Universitaires de France, Paris.

[Watson and Gibson, 2004] Watson, D. and Gibson, E. (2004). The relationship between intonational phrasing and syntactic structure in language production. *Language and Cognitive Processes*, 6:713–755.

[Wennerstrom, 2001] Wennerstrom, A. (2001). *The Music of Everyday Speech: Prosody and Discourse Analysis*. Oxford University Press, Oxford.

[Wightman, 2002] Wightman, C. (2002). ToBI or not ToBI? In *Speech Prosody*, pages 25–29, Aix-en-Provence, France.

[Wouters and Macon, 2001] Wouters, J. and Macon, M. (2001). Control of spectral dynamics in concatenative speech synthesis. *IEEE Transactions on Speech and Audio Processing*, 9(1):30–38.

[Yamagishi, 2006] Yamagishi, J. (2006). *Average-Voice-Based Speech Synthesis*. PhD. thesis, Tokyo Institute of Technology.

[Yamagishi, 2007] Yamagishi, J. (2007). Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training. *IEICE Transactions on Information and Systems*, 90(2):533–543.

[Yamagishi et al., 2008] Yamagishi, J., Kawai, H., and Kobayashi, T. (2008). Phone duration modeling using gradient tree boosting. *Speech Communication*, 50(5):405–415.

[Yamagishi et al., 2004] Yamagishi, J., Masuko, T., and Kobayashi, T. (2004). HMM-based expressive speech synthesis - Towards TTS with arbitrary speaking styles and emotions. In *Special Workshop in Maui*, Maui, Hawaï.

[Yan et al., 2009] Yan, Z.-J., Qian, Y., and Soong, F. K. (2009). Rich Context Modeling for High Quality HMM-based TTS. In *Interspeech*, pages 4025–4028, Brighton, UK.

[Yoshimura et al., 1997] Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. (1997). Speaker interpolation in HMM-based speech synthesis system. In *European Conference on Speech Communication and Technology*, pages 2523–2526, Rhodes, Greece.

[Yoshimura et al., 1998] Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. (1998). Duration modeling for HMM-based speech synthesis. In *International Conference on Spoken Language Processing*, pages 29–32, Sydney, Australia.

[Yoshimura et al., 1999] Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. (1999). Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In *European Conference on Speech Communication and Technology*, pages 2347–2350, Budapest, Hungary.

[Young et al., 2002] Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., Valtchev, V., and Woodland, P. (2002). *The HTK Book*. Cambridge University Press, Cambridge.

[Zen et al., 2007] Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A., and Tokuda, K. (2007). The HMM-based speech synthesis system version 2.0. In *Speech Synthesis Workshop*, pages 294–299, Bonn, Germany.

[Zen et al., 2009] Zen, H., Tokuda, K., and Black, A. (2009). Statistical parametric speech synthesis. *Speech Communication*, 51(11):1039–1064.

[Zen et al., 2004] Zen, H., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. (2004). Hidden semi-Markov model based speech synthesis. In *International Conference on Spoken Language Processing*, pages 1397–1400, Jeju Island, Korea.