Silent Communication: whispered speech-to-clear speech conversion

Viet-Anh Tran

INSTITUT POLYTECHNIQUE DE GRENOBLE

N° attribué par la bibliothèque

|__|__|__|__|__|__|__|__|__|__|

T H E S E

pour obtenir le grade de

DOCTEUR DE L'INP Grenoble

Spécialité : *Signal, Image, Parole, Télécoms*

préparée au : Département de Parole et Cognition du laboratoire GIPSA-Lab

dans le cadre de **l'Ecole Doctorale** : *Electronique, Electrotechnique, Automatique et Traitement du Signal*

présentée et soutenue publiquement

par

Viet Anh Tran

le 28 Janvier 2010

TITRE

Silent Communication : whispered speech-to-clear speech conversion

Communication silencieuse : conversion de la parole chuchotée en parole claire

DIRECTEUR DE THESE
Hélène Lœvenbruck
CO-DIRECTEUR DE THESE
Christian Jutten & Gérard Bailly

JURY

M.	Eric Moulines	, Président
Mme.	Régine André-Obrecht	, Rapporteure
M.	Gérard Chollet	, Rapporteur
Mme.	Hélène Lœvenbruck	, Directrice de thèse
M.	Christian Jutten	, Co-encadrant
M.	Gérard Bailly	, Co-encadrant
M.	Yves Laprie	, Examinateur

Acknowledgements

I would like to express my deepest appreciation to my advisors Dr. Hélène Lœvenbruck, Dr. Gérard Bailly and Prof. Christian Jutten, whose constant guidance, encouragement and support from the initial to the final level enabled me to develop an understanding of this research in GIPSA-Lab.

During my thesis, I had a great opportunity to work with Dr. Tomoki Toda, assistant professor at *Nara Institute of Science and Technology* (NAIST), Japan. I would like to thank Dr. Tomoki Toda for letting me use the NAM-to-speech system, developed at NAIST, and especially for his continuous support, valuable advice, discussions and collaboration throughout this research.

I would like to express my gratitude to the various members of the jury. I would like to acknowledge Prof Régine André-Obrecht and Dr. Gérard Chollet for having accepted to be examiners of my thesis and for their helpful comments and suggestions in the writing of the thesis. Special thanks go to Prof. Eric Moulines for accepting to be the president of the defense committee as well as Dr. Yves Laprie for his participation in the jury.

I am grateful to Coriandre Vilain, Christophe Savariaux, Lionel Granjon and Alain Arnal for their help in fixing the hardware problems and in data acquisition. Thanks also go to Pierre Badin for valuable advice about the presentation of my thesis.

I also wish to thank Dr. Panikos Heracleous with whom I was fortunate to discuss and collaborate on the Non-Audible Murmur recognition problem.

Finally, I would like to acknowledge my colleagues: Nino, Nadine, Mathieu, Oliver, Xavier, Anahita, Noureddine, Stephan, Atef, Amélie, Keigo, Yamato and others for many good times with them at Gipsa-Lab and at NAIST, and of course, my family and my friends for their endless encouragement and support.

Contents

List of figures

List of tables

Abstract

In recent years, advances in wireless communication technology have led to the widespread use of cellular phones. Because of noisy environmental conditions and competing surrounding conversations, users tend to speak loudly. As a consequence, private policies and public legislation tend to restrain the use of cellular phone in public places. Silent speech which can only be heard by a limited set of listeners close to the speaker is an attractive solution to this problem if it can effectively be used for quiet and private communication. The motivation of this research thesis was to investigate ways of improving the naturalness and the intelligibility of synthetic speech obtained from the conversion of silent or whispered speech. A Non-audible murmur (NAM) condenser microphone, together with signal-based Gaussian Mixture Model (GMM) mapping, were chosen because promising results were already obtained with this sensor and this approach, and because the size of the NAM sensor is well adapted to mobile communication technology. Several improvements to the speech conversion obtained with this sensor were considered.

A first set of improvement concerns characteristics of the voiced source. One of the features missing in whispered or silent speech with respect to loud or modal speech is F_0, which is crucial in conveying linguistic (question vs. statement, syntactic grouping, etc.) as well as paralinguistic (attitudes, emotions) information. The proposed estimation of voicing and F_0 for converted speech by separate predictors improves both predictions. The naturalness of the converted speech was then further improved by extending the context window of the input feature from phoneme size to syllable size and using a Linear Discriminant Analysis (LDA) instead of a Principal Component Analysis (PCA) for the dimension reduction of input feature vector. The objective positive influence of this new approach of the quality of the output converted speech was confirmed by perceptual tests.

Another approach investigated in this thesis consisted in integrating visual information as a complement to the acoustic information in both input and output data. Lip movements which significantly contribute to the intelligibility of visual speech in face-to-face human interaction were explored by using an accurate lip motion capture system from 3D positions of coloured beads glued on the speaker's face. The visual parameters are represented by 5 components related to the rotation of the jaw, to lip rounding, upper and lower lip vertical movements and movements

of the throat which is associated with the underlying movements of the larynx and hyoid bone. Including these visual features in the input data significantly improved the quality of the output converted speech, in terms of F_0 and spectral features. In addition, the audio output was replaced by an audio-visual output. Subjective perceptual tests confirmed that the investigation of the visual modality in either the input or output data or both, improves the intelligibility of the whispered speech conversion.

Both of these improvements are confirmed by subjective tests.

Finally, we investigated the technique using a phonetic pivot by combining Hidden Markov Model (HMM)-based speech recognition and HMM-based speech synthesis techniques to convert whispered speech data to audible one in order to compare the performance of the two state-of-the-art approaches. Audiovisual features were used in the input data and audiovisual speech was produced as an output. The objective performance of the HMM-based system was inferior to the direct signal-to-signal system based on a GMM. A few interpretations of this result were proposed together with future lines of research.

Résumé

Les avancées des technologies de communication sans fil ces dernières années ont mené à l'utilisation répandue des téléphones portables pour la communication privée. En raison des conditions environnementales bruyantes et des conversations environnantes concurrentes, les utilisateurs tendent à parler fort. Par conséquent, la législation publique tendent à limiter l'utilisation du téléphone mobile dans les lieux publics. La voix silencieuse qui peut seulement être entendue par un ensemble limité d'auditeurs entourant le locuteur est une solution attrayante à ce problème si elle peut effectivement être employée pour la communication privée, confidentielle. La motivation de cette thèse était d'étudier différentes façons d'améliorer le naturel et l'intelligibilité de la parole synthétique obtenus à partir de la conversion de la voix silencieuse ou chuchotée. Un microphone à condensateur NAM, utilisant le *mapping* direct signal-vers-signal basé sur un modèle GMM (Gaussian Mixture Model), a été choisi parce que des résultats prometteurs ont été déjà obtenus avec ce capteur et cette approche, et parce que la taille du capteur NAM est bien adaptée à la technologie de communication mobile. Différentes améliorations de la conversion de parole obtenue avec ce capteur ont été envisagées.

Un premier ensemble d'amélioration concerne les caractéristiques de la source voisée. Un des traits manquant dans la voix chuchotée ou silencieuse en ce qui concerne la parole modale est F_0, qui est crucial pour l'information linguistique (question ou affirmation, regroupement syntactique, etc.) aussi bien que l'information paralinguistique (attitudes, émotions). L'évaluation proposée du voisement et du F_0 pour la parole convertie en séparant les modules améliore les deux prédictions. Le naturel de la parole convertie a été alors encore amélioré en allongeant la fenêtre de contexte d'entrée, de la taille du phonème à la taille de la syllabe, et en employant une analyse LDA (Linear Discriminant Analysis) au lieu d'une PCA (Principal Component Analysis) pour la réduction de la dimension du vecteur d'entrée.

Une autre approche étudiée dans cette thèse a consisté à intégrer l'information visuelle comme complément à l'information acoustique dans les modalités d'entrée et de sortie. Les mouvements des lèvres qui contribuent de manière significative à l'intelligibilité de la parole visuelle dans l'interaction humaine face-à-face ont été intégrés en employant un système de capture précis des mouvements des lèvres à l'aide des positions 3D de billes collées sur le visage du locuteur. Les paramètres visuels ont été représentés par 5 composants liés à la rotation de la mâchoire, à l'arrondissement des lèvres, aux mouvements verticaux des lèvres supérieure et inférieure et aux mouvements de la gorge qui sont associés aux mouvements fondamentaux du larynx et de l'os hyoïde. L'inclusion de ces paramètres visuels dans les données d'entrée du système a amélioré de façon significative la qualité de la parole convertie en sortie, en termes de F_0 et de spectre. De plus, la sortie audio a été remplacée par une sortie audio-visuelle.

Des tests perceptifs subjectifs ont confirmé que l'intégration de la modalité visuelle soit dans les données d'entrée, soit dans celles de sortie, soit dans les deux, améliore significativement l'intelligibilité de la conversion de parole chuchotée.

Tous ces améliorations sont confirmées par les tests subjectifs.

Enfin, nous avons étudié la technique utilisant un pivot phonétique en combinant la reconnaissance de la parole et la synthèse de la parole basée sur un modèle de Markov caché (HMM) pour convertir les données chuchotées en parole claire afin de comparer la performance de ces deux approches « état de l'art ». Des paramètres audio-visuels ont été utilisés dans les données d'entrée et un signal de parole audiovisuel a été produit en sortie.

La performance objective de ce système à base de HMM était inférieure à celle du système d'appariement signal-vers-signal fondé sur un GMM. Plusieurs interprétations de ce résultat ont été proposés ainsi que des perspectives de recherche dans ce domaine.

Introduction

Motivation of research

Silent speech is the kind of speech defined as the articulated production of sound with little vibration of the vocal cords in the case of whisper or no vibration at all in the case of murmur, produced by the motions and interactions of speech organs such as tongue, palate, lips, etc., to avoid being overheard. This kind of speech is commonly used for private and confidential communication, especially useful in military environments, or can be used by laryngeal handicapped people who cannot speak normally.

Unfortunately, it is difficult to directly use silent speech as a medium for face-to-face communication as well as over mobile telephones because the linguistic content and paralinguistic information in the uttered message is degraded when the speaker murmurs or whispers. A recent direction of research considered by researchers and telecommunication industries is how to convert silent speech to modal voice in order to have a more intelligible and more familiar speech. If it could be done, potential outcomes such as "silent speech telephone" as well as robust "speaking-aid" applications for laryngeal handicaps would become feasible. My work in this thesis therefore concentrates on this direction.

Scope of thesis

Several silent speech devices have been explored in the literature including surface electromyography (sEMG) (Jorgensen *et al.*, 2003; Jorgensen and Binsted 2005; Jou, Schultz *et al.*, 2006, 2008; Walliczek *et al.*, 2006; Toth *et al.*, 2009), non-audible murmur (NAM) microphone (Nakajima, 2003,2006; Heracleous *et al.*, 2005, 2009; Toda and Shikano 2005), ultrasound and optical imagery (Denby and Stone, 2004; Denby *et al.*, 2006; Hueber *et al.*, 2007, 2008ab, 2009; Denby *et al.*, in press), Electromagnetic Articulography (EMA) (Fagan *et al.*, 2008) and Electro-encephalographic (EEG) (Suppes *et al.*,1997; Wester and Schultz 2006; Porbadnigk *et al.*, 2009). Among them, one that seems particularly interesting is the NAM microphone developed by researchers at Shikano Laboratory, NAIST, Japan because of its usability and its suitable size for mobile communication

technology. Nakajima *et al.*, 2003 proposed that it might be more efficient, in noisy environment, to analyze the original vibrations uttered as speech from inside the body, through the surface of the skin, instead of analyzing the sounds in the air, after being discharged through the mouth. They introduced a new communication interface which can capture acoustic vibration in the vocal tract from a stethoscopic sensor placed on the neck, below the ear. By using this microphone, Toda and Shikano (2005) proposed a NAM-to-Speech conversion system based on a GMM model in order to convert "non-audible" speech to audible voice. Although the segmental intelligibility of synthetic signals computed by this statistical feature mapping is quite acceptable, listeners have difficulty in chunking the speech continuum into meaningful words. This is mainly due to impoverished synthetic intonation estimated from unvoiced speech like NAM. The main contribution of our work in this thesis is to improve the performance of such a system.

The following is a summary of our contributions:

Signal-to-signal mapping

- **Pitch.** A first improvement proposed in this thesis is to better estimate the voice source for the converted speech. Several approaches are explored, including training on voiced segments only, separating the voicing and F_0 estimation in the synthesis process, optimizing the context window size of the input feature and using LDA (Linear Discriminant Analysis) instead of PCA (Principal Component Analysis) to reduce the input vector dimension.

- **Audiovisual input/output.** Another solution explored in this thesis to improve the performance of the system is to integrate visual information as a complement to the acoustic information in both input and output data. Facial movements are estimated using an accurate motion capture device. The extracted facial parameters are then used to drive a the talking head system, a technique developed at the Speech and Cognition Department of GIPSA-lab.

HMM-based recognition-synthesis

Another approach to map silent speech to audible voice is combining speech recognition and synthesis techniques as proposed in (Hueber *et al.*, 2007, 2008ab, 2009). By introducing linguistic levels, both in recognition and synthesis, such a system can potentially compensate for the impoverished input by including linguistic knowledge into the recognition process. We compared this approach with

the previous direct signal-to-signal mapping, to explore a further solution to improve the quality of the whisper-to-modal speech conversion system.

Organization of thesis

The thesis is organized as follows.

Chapter 1 starts with the definition of different types of silent speech, followed by the presentation of several new communication interfaces and techniques that have been used recently to capture silent speech. These interfaces, as mentioned above, include Electromagnetic Articulography (EMA), surface Electromyography (sEMG), Non-Audible Murmur microphone (NAM), Encephalography (EEG), Ultrasound (US) and optical imagery of the tongue and lips. Finally, some techniques used to map silent speech to audible speech are presented. The first technique is direct signal-to-signal mapping using aligned corpora. This technique is derived from voice conversion, a technique frequently used in speech synthesis to generate different output voices with limited resources. The second method incorporates linguistic knowledge by chaining speech recognition and speech synthesis. Both approaches are experimented in this thesis.

Chapter 2 provides background information on whispered speech production and perception from previous research of different view points. This chapter starts with the physiological characteristics of whispered speech, then describes the acoustic differences between whispered speech and phonated speech. Next, this chapter discusses how listeners can identify “pitch” during whispered speech production even if there is no vibration of the vocal folds. Finally, visual cues are described as a complementary information which improves speech perception, especially in the case of silent speech or in conditions of lip-reading, where there is no acoustic information.

Chapter 3 focuses on the design and the acquisition of our data for different languages: French and Japanese. The chapter presents the construction of the French corpus which only includes audio data as well as the construction of a Japanese corpus with simultaneous audio and video modalities. The pre-processing of the video data is then explained. A guided PCA is applied to the video data in order to extract the main facial movement parameters related to speech articulation.

Chapter 4 describes my contribution to improve the intelligibility and the naturalness of the synthetic speech generated by a direct signal-to-signal mapping

technique. Different solutions are presented in this chapter. They include the improvement of the precision of voicing decision and F_0 estimation, by the extension of the context window of the input feature from the phoneme size to the syllable size and by the use of LDA instead of PCA to reduce the input feature dimension. Finally the use of visual parameters in both input and output of the conversion system is investigated. The positive influence of these solutions is assessed by objective tests and confirmed by subjective perceptual tests.

Chapter 5 on the other hand concentrates on the phonetic pivot HMM-based recognition-synthesis chain. After studying the impact of facial movement information on the performance of both recognition and synthesis, the objective performance of the two systems, i.e. the direct signal-to-signal mapping described in the previous chapter and the HMM-based conversion system, are compared.

Finally, the **conclusion** summarizes the contributions of this thesis and offers suggestions for future work.

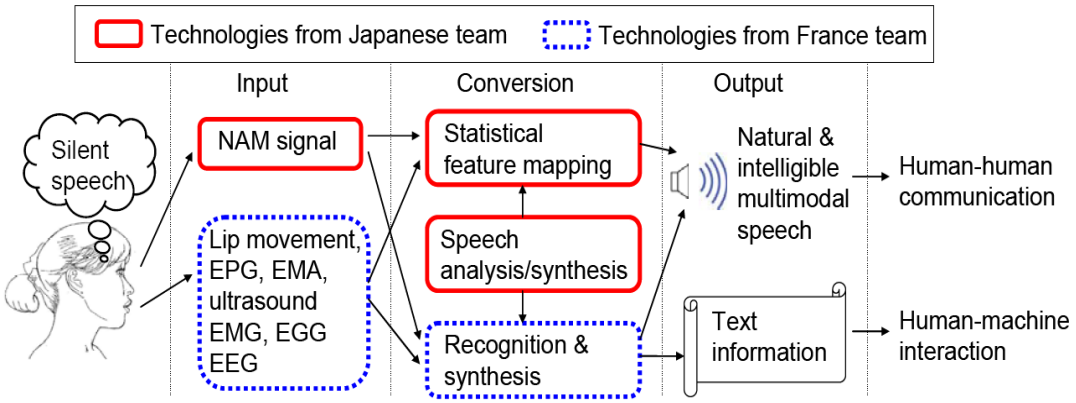


Figure 1. *Silent speech interfaces developed in the CASSIS project.*

Related project

The work presented in this thesis contributes to the CASSIS (*Computer-Assisted communication and Silent Speech InterfaceS*) project (figure 1) which involves the collaboration of GIPSA-Lab, ENST-Paris, ESPCI-Paris and NAIST. The aim of the CASSIS project is to enable confidential human-human or human-computer communication as well as to compensate for phonation loss in silent speech. The basic challenge is to convert multimodal signals (brain or muscular activities, orofacial movements, contact sounds, etc) gathered during silent or quasi-silent articulation into an audible signal or a phonological representation of what has

been said. The CASSIS project provides a scientific framework to share technologies, develop new interfaces and settle common data recordings and evaluation campaigns. Multilingual evaluation is crucial in this respect since performance is highly dependent on the phonological structure of the language (sound structure, syllable complexity, accentual structure, etc) and since language-specific knowledge is often injected implicitly or explicitly in the mapping process. Comparative assessment of such speech technologies will be performed on Japanese, English and French multimodal resources. The ultimate challenge of this project is also to consider real-time implementations of the proposed solutions so that usability studies can be conducted at the end of the project.

Chapter 1

Silent speech interfaces and silent-to-audible speech conversion

1.1 Introduction

Due to the limitations of traditional speech interfaces – i.e. limited robustness in the presence of ambient noise, lack of secure transmission of private and confidential information and interference of bystanders in concurrent environments – a novel approach is necessary to capture articulatory or cerebral causes rather than their acoustic consequences, i.e. silent speech. By acquiring multimodal data from the articulators, brain or muscular activities, orofacial movements, etc. gathered during silent articulation, a Silent Speech Interface (SSI) produces a representation of speech which can be synthesized directly and can provide an audible signal. (Denby *et al.*, in press).

Potential applications for SSIs can be found more and more. First of all, SSIs are intuitively used to provide privacy for conversations over cellular phones. An SSI, if non-invasive and small enough to be incorporated into a cell phone, could allow users to communicate more silently, privately and confidentially. Based on their natural non-air-borne speech cues, SSIs are robust against environmental background noise and therefore are also adapted to speech processing in noisy environments (Denby *et al.*, in press). Moreover, SSIs are gradually used to replace current speech pathology aids, such as the electrolarynx for the laryngectomy handicaps, thanks to their potentially natural sounding.

This chapter presents the definition of silent speech in section 1.2. A panorama of technologies used to capture silent speech signatures follows in section 1.3. Section 1.4 concentrates on two main conversion techniques to synthesize audible voice from silent speech: direct mapping signal-to-signal and coupling a speech synthesis system and a non-audible speech recognizer. Finally, section 1.5 presents some conclusions for this chapter.

1.2 Silent speech

Silent speech is a common, natural way that speakers use to reduce speech perceptibility. For example when they are asked to speak softly so as not to disturb others in a school library, in a conference ..., when they are too weak to speak normally or when they tell private, confidential information. It seems that silent speech is the most effective and efficient vocal communication when only a few people around the speaker (or only him/herself when s/he is speaking on the phone) should hear the message.

Depending on the level of "silence" or audibility, we can define 5 categories:

1. **Inner speech:** It is also called imagined speech, covert speech or verbal thought. It refers to the silent production of words in one's mind. Inner speech can be considered as mental simulation of speech. Some researchers suggest that inner speech is the same as overt speech, except that execution is blocked, i.e. overt speech equals inner speech plus a motor execution process. This 'continuity hypothesis' predicts that physiological correlates of inner speech (such as duration, muscular activity, heart rate, respiration rate and neuronal activity) should bear resemblance to those of overt speech. This hypothesis implies that speech motor activities may exist during inner speech, resulting from motor planning, but that these activities do not result in effective muscular recruitment. The hypothesis is given credit by orofacial electromyographic (EMG) recording during inner speech. EMG activity has been detected in the speech musculature during verbal mental imagery and covert rehearsal (Jacobson, 1931; Sokolov, 1972). McGuigan and Dollins (1989) conducted EMG recording showing that the lips are active when silently reading the letter "P", but not when reading "T". Reciprocally, the tongue was active only for silently reading "T". Livesay *et al.* (1996) observed lip EMG activity in an inner speech recitation task, but not in a visualization task. Fadiga *et al.* (2002) have shown, using Transcranial Magnetic Stimulation (TMS), that during listening of words which involve tongue movements, there is an increase of motor-evoked potentials recorded from the listeners' tongue muscles.

2. **Subvocal invisible speech** : This speech mode is articulated very softly so that it cannot be heard, but the speech articulators (tongue, lips, perhaps jaw) may slightly move. It corresponds to articulation without phonation and with very little air emission. Subvocal speech is usually **hypo-articulated**, such as when a person silently reads or recites to him/herself.

The articulatory movements are so small that they may not be noticed by surrounding viewers. This kind of speech is what is studied by researchers from a group at NASA Ames Research Center, who are working on the decoding of facial EMG signals arising from subvocal speech (Bett and Jorgensen 2005; Jorgensen and Binsted 2005). By sticking EMG sensors under a person's chin and on either side of the Adam's apple, this group claim they can recover EMG activities from laryngeal, mandibular and lingual muscles and translate them into words.

3. **Subvocal visible speech** : This speech mode corresponds to silent speech mouthing. Speech is normally articulated but without air emission. The articulators move, as if one wanted to be seen but not heard. Lip, jaw, cheek and chin motions can thus be tracked to decipher speech. This speech mode is used by Tanja Schultz, Alex Waibel and colleagues (at Carnagie Mellon University, CMU and at the Interactive Systems Laboratories of Universität Karlsruhe) to derive audible speech from facial EMG recordings (e.g. Jou *et al.*, 2007). Another research group, involved in the *Ouisper*[1] project, also try to synthesize audible speech from the ultrasound motion of the tongue movement during the production of this speech mode (Denby and Stone, 2004; Denby *et al.*, 2006; Hueber *et al.*, 2007, 2008ab, 2009).

4. **Non-audible murmur** (NAM): This speech mode is softly whispered so that a nearby person would not be able to hear it. It is defined as the articulated production of respiratory sound without recourse to vocal-fold vibration, produced by the motions and interactions of speech organs such as the tongue, palate, lips etc. (Nakajima *et al.*, 2003ab). It can be viewed as subvocal invisible speech with air emission.

5. **Whispered speech**: although it is difficult to define precisely the acoustic differences between "whisper" and "NAM", the term "whisper" implies that limited nearby listeners can hear the content of the speech, and that it can be recorded by an external microphone through transmission in the air (Nakajima *et al.*, 2003ab).

The following sections will present some current technological attempts in the literature, designed to capture silent speech as well as several approaches to *silent speech-to-audible speech* conversion, aiming at obtaining a more natural and intelligible voice.

[1] http://www.neurones.espci.fr/ouisper/

1.3 Silent speech interfaces (SSI)

The speech production process can be regarded as producing a set of coherent multimodal signals, such as the electrical cerebral signals, the myoelectrical signals observable from the muscles, the movements of the orofacial articulators that are visible or not, or acoustic signals (figure 1.1). The aim of this section is to overview the spectrum of available technologies that can be used to record useful signals characterizing articulation and phonation of silent speech.

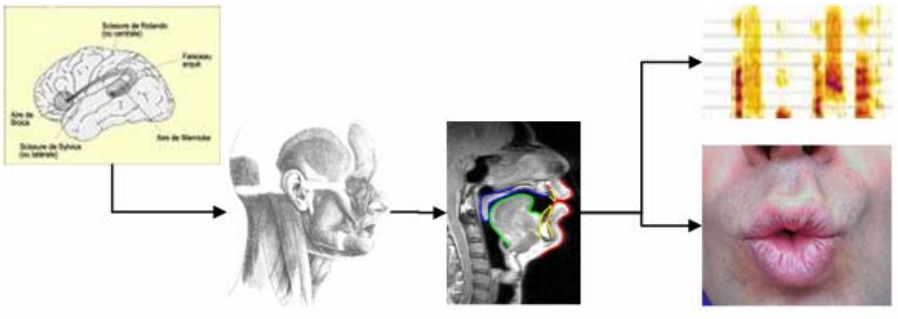


Figure 1.1. *Multimodal signals during the speech production process*

1.3.1 Electro-encephalographic (EEG) sensors

It is well known that several brain areas are activated in the production of speech. *Broca's* area has been shown to be involved in the planning and decision process during speech production while *Wernicke's* area has been shown to be active during speech comprehension (Callies, 2006; Wester and Schultz, 2006).

To exploit the electromagnetic waves created by these cerebral activities, non-invasive electroencephalography or EEG devices are frequently used as a brain-to-computer interface (BCI) (Wolpaw *et al.*, 2002).

By using this type of sensor for silent speech, Suppes *et al.* (1997) have shown that isolated words can be recognized based on ElectroEcephaloGraphy (EEG) and MagnetoEncephaloGraphy (MEG) recordings. In one of their experimental conditions called internal speech, the subjects were shown one out of 12 words on a screen and asked to utter this word ’silently’ without using any articulatory muscles (this corresponds to what we defined as “inner speech” above). Recognition rates, based on a least-squares criterion, varied widely, but were significantly different from chance. The two best scores were above 90%. These results show that brain waves carry substantial information about the word being processed under experimental conditions of conscious awareness.

Figure 1.2. *EEG-based recognition system for unspoken speech (Wester and Schultz, 2006)*

Another work by Wester and Schultz (2006) investigated a similar approach which directly recognizes "*unspoken speech*" in brain activity measured by EEG signals (figure 1.2). "Unspoken speech" here refers to what we defined as "inner speech" above. In this study, the authors used 16 channel EEG data recorded using the International 10-20 system (Jasper, 1958) in five different modalities: normal speech, whispered speech, silent speech, mumbled speech and unspoken or inner speech. They concluded that speech recognition on EEG brain waves is possible with a word accuracy four to five times higher than chance for vocabularies of up to ten words. The same results were found for the other modalities. Unspoken speech was slightly worse than the other modalities (Wester and Schultz, 2006). Their experiments also showed that the important EEG recording regions for unspoken speech recognition seem to be the motor cortex, Broca's area and Wernicke's area as shown in figure 1.3.

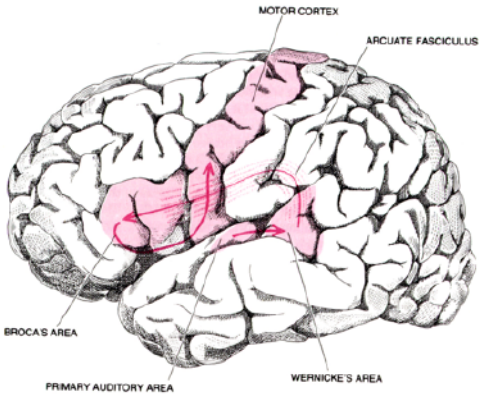


Figure 1.3. *The motor cortex, Broca's area and Wernicke's area seem to be recruited for inner speech (Wester and Schultz, 2006)*

However, subsequent investigations lead to the hypothesis that the good recognition performance might in fact have resulted from temporal correlated artifacts in the brain waves since the words were presented in blocks. Due to these artifacts, the recognition results reported in (Wester and Schultz, 2006) might be overestimated. A recent study by Porbadnigk *et al.* (2009) tried to prove this hypothesis. In their experiments, each of the first five words of the international radio-telephony spelling alphabet (alpha, bravo, charlie, delta, echo) was repeated 20 times by 21 subjects. Each session had the same word list (length 100) but the word order was varied: blockwise, sequential mode or random mode (see Porbadnigk *et al.* 2009 for more details). The interested signals, 16 EEG channels using a 128 cap montage, were recognized by a HMM-based classifier. The authors discovered that the average recognition rate of 45.5% in block mode drops to chance level for all other modes. This means that temporally correlated brain activities tend to superimpose the signal of interest. The authors also found that cross-session training (within subjects) yields recognition rates only at chance level.

Dasalla and colleagues (Dasalla *et al.*, 2009) proposed another scheme for a silent speech BCI using neural activities associated with vowel speech imagery. EEG was recorded in three healthy subjects for three tasks, imaginary speech of the English vowels */a/* and */u/*, and a no-action state as control. 50 trials were performed for each task, with each trial containing two seconds of task-specific activity. Trial averages revealed readiness potentials at 200 ms after stimulus and speech related potentials peaking after 350 ms. The authors then designed spatial filters using the common spatial patterns (CSP) method, which, when applied to the EEG data, produce new time series with variances that are maximally discriminative. After spatially filtering the EEG data, the authors trained a nonlinear support vector

machine (SVM) for the classification task. Overall classification accuracies ranged from 68% to 78%. Results indicate significant potential for the use of the proposed system for EEG-based silent speech interfaces.

1.3.2 Surface Electromyography (sEMG)

The surface Electromyography (sEMG) method measures muscular electric potential with a set of electrodes attached to the skin where the articulatory muscles underlie (Jou and Schultz, 2008). In the speech production chain (as presented in the figure 1.1), neural signals generated from the brain drive articulatory muscles. The articulatory muscles then contract and relax accordingly to shape the geometry of the vocal tract and produce appropriate sounds. The muscle activity alters the small electrical currents through the body tissue whose resistance creates potential differences, and the sEMG method can pick up this kind of potential change for further signal processing, e.g., speech recognition or speech synthesis. Figure 1.4 shows the surface muscles involved in speech production that can be used for an EMG-based system (Maier-Hein L., 2005). The motivation here is that the sEMG method is inherently robust to ambient noise because the sEMG electrodes are directly in contact with the human tissue without the air-transmission channel. Therefore the sEMG method makes it possible to recognize completely silent speech, which means mouthing words without making any sound (Jou and Schultz, 2008).

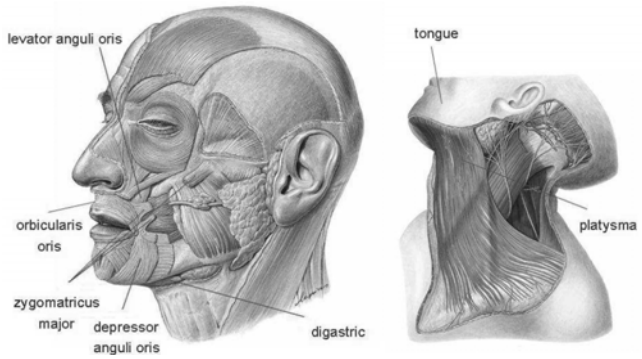


Figure 1.4. *Some of the surface muscles involved in speech production (Maier-Hein L., 2005)*

EMG was earlier used to develop a speech prosthesis that functions using the myo-electric signal as its only input (Morse and O'brien, 1986; Susie and Tsunoda, 1985). In (Morse and O'Brien, 1986), time domain recognition of speech from myoelectric signals using maximum likelihood pattern recognition was first studied. The authors investigated the EMG signals from three muscles of the neck,

near the vocal tract and one on the left temple. Recognition accuracy as high as 97% was attained on a two-word vocabulary. However, for larger vocabularies, the recognition accuracy deteriorated, falling below 70% accuracy for ten-word vocabularies and 35% for 17-word vocabularies in the following studies of theses authors (Morse *et al.*, 1989, 1990, 1991). In another work by Susie and Tsunoda (1985), the target muscles were three muscles around the mouth: the *digastricus*, the *zygomaticus major* and the *orbicularis* (as can be seen in the figure 1.4). The words were classified using a finite automation. An average recognition accuracy of 64% was attained in discriminating the five Japanese vowels. This speech recognition accuracy using EMG signals would be considered poor by today's conventional speech recognition standards; however, the accuracy of these systems is above the random guessing which promises the presence of speech information within the myoelectric signals from the muscles of articulation.

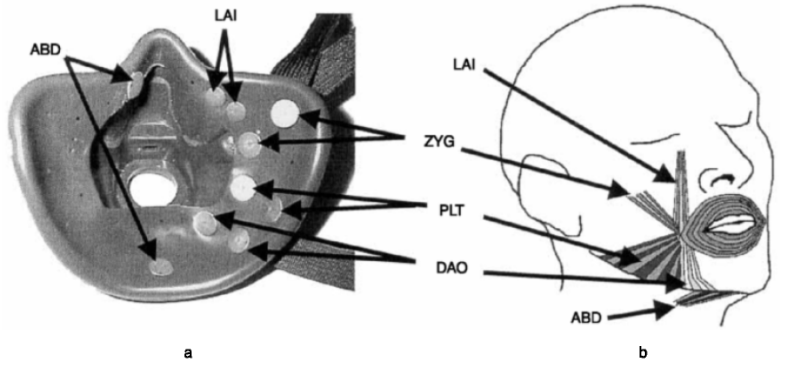


Figure 1.5. *(a) Pilot oxygen mask with electrodes embedded in rubber lining: (b) anatomical diagram indicating location of target muscles (Chan et al., 2002).*

The research of Chan and colleagues (2002) continued in this direction by experimenting automatic speech recognition using sEMG. They proposed to supplement voiced speech with EMG in the context of noisy aircraft pilot communication. In their work, they studied the feasibility of augmenting auditory speech information with EMG signals recorded from primary facial muscles using sensors embedded in a pilot oxygen mask. They used five surface signal sites (using Ag-AgCl button electrodes), from five facial muscles: the levator anguli oris (LAI), the zygomaticus major (ZYG), the platysma (PLT), the depressor anguli oris (DAO) and the anterior belly of the digastric (ABD) (Figure 1.5). They recorded the EMG activity of these muscles during the vocalized pronunciation of the digits zero to nine, in parallel with an additional acoustic channel to segment the signals. A Linear Discriminant Analysis (LDA) classifier utilized on a set of wavelet transform features (reduced by Principle Component Analysis (PCA) yielded a word accuracy of 93%. The authors also claimed that the performance was very

sensitive to the pre-triggering value for sEMG signals. This phenomenon occurs because the onset of muscle activity precedes the acoustic voice signal. Since the temporal position of articulation signal relative to the acoustic signal varies with the speaking rate, the authors proposed to use a Hidden Markov Model (HMM) (Chan *et al.*, 2002). Even though the HMM classifier yielded worse maximum recognition rates (86%) than the LDA classifier for the same data, it was much less susceptible to temporal misalignment, that is, there was no dramatic decrease in performance when the pre-trigger value used for the training set was slightly different from the one used in the test set.

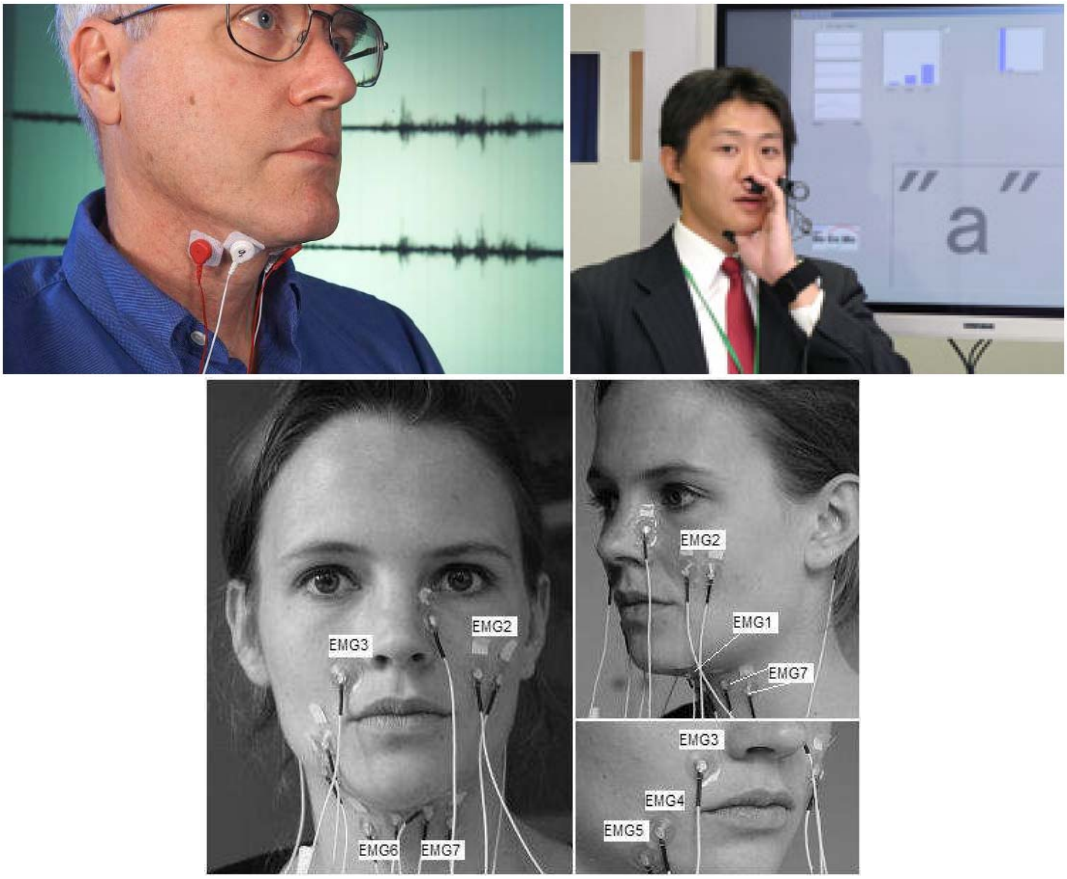


Figure 1.6. *(a) the subvocal EMG system proposed by Jorgensen et al. (reference). (b) the subvocal EMG system proposed by DoCOMO Research center. (c) electrode positioning by the CMU group.*

For silent speech recognition with EMG, Manabe *et al.* (2003), in the DOCOMO project (Figure 1.6b), showed that it is possible to recognize five Japanese vowels using surface EMG signals recorded with electrodes pressed on the facial skin. The electrodes were placed under the chin, on the cheek, and on the lips, corresponding to the digastrics muscle , the zygomaticus major, and the orbicularis oris (Manabe

et al., 2003). Using an artificial neural network (ANN), the recognition accuracy was over 90%. In their later work on ten Japanese digits recognition (Manabe and Zhang, 2004), experiments on speaker-dependent recognition on 10 isolated Japanese digits with a multi-stream HMM indicated that it is effective to give different weights to different sEMG channels so that the corresponding muscles can contribute to a different extent to the classification. The authors reported a maximum recognition rate of 65% using delta filterbank coefficients and spectral subtraction.

Chuck Jorgensen and colleagues (Jorgensen *et al.*, 2003; Jorgensen and Binsted 2005) from the NASA Ames Resarch Center proposed an EMG system that captures subvocal speech. Note that this speech mode is articulated very softly so that it cannot be heard, but the speech articulators may slightly move. Their aim is to track hypoarticulated subvocal speech, with very limited articulatory movement and no phonation, nor air emission. Their idea is to intercept nervous control signals sent to speech muscles using surface EMG electrodes placed on the larynx and sublingual areas below the jaw (Figure 1.6b). The authors reported recognition rates of 92% on a set of six control words using a neural network classifier (Jorgensen *et al.*, 2003). They examined various feature extraction methods, including STFT coefficients, wavelets, and Linear Predictive Coding (LPC) coefficients and reported a maximum word accuracy for dual tree wavelets (92%) followed by Fourier coefficients (91%). In 2005, the authors extended the original six word vocabulary to the ten English digits and achieved a word accuracy of 73% (Jorgensen and Binsted, 2005).

Several important issues in sEMG based non-audible speech recognition concern repositioning electrodes between recording sessions, environmental temperature changes, and skin tissue properties of the speaker. In order to reduce the impact of these factors, Maier-Hein, Schultz, Waibel and colleagues have investigated a variety of signal normalization and model adaptation methods (Maier-Hein *et al.*, 2005). By experimenting on a vocabulary consisting of the ten English digits “zero” to “nine”, the authors noted that sharing training data across sessions and applying methods based on Variance Normalization and Maximum Likelihood adaptation improve across-sessions performance. They achieved an average word accuracy of 97.3% for within-session testing using seven EMG channels as presented in figure 1.6c. Across-sessions testing without any adaptation yielded an average of 76.2%. By applying the normalization and adaptation methods they were able to bring recognition rates back up to 87.1%.

A recent work by Lee (2008) proposed a method to use global control variables to model correlations among the EMG channels. The system treats each EMG as a speech signal and thus uses Mel-scale spectral coefficients (MFCC) as main features, with delta and delta-delta for dynamics modeling. EMG channels are obtained from three articulatory muscles of the face: the *levator anguli oris*, the *zygomaticus major*, and the *depressor anguli oris*. The sequence of EMG signals for each word is modelled by a HMM framework. Their aim is building a model for state observation density when multi-channel observation sequences are given. The proposed model reflects the dependencies between each of the EMG signals, which are described by introducing a global control variable. In their preliminary study, 60 Korean isolated words were used as recognition variables. The findings indicated that such a system may achieve an accuracy of up to 87%, which is superior to the independent probabilistic model.

However, these pioneering studies are limited to small vocabulary sizes ranging from five to around forty isolated words. The main reason of this limitation is that the classification unit is set to a whole utterance, instead of a restrained to a phone, a smaller and more flexible unit that is frequently used in large vocabulary continuous speech recognition. Such a phone-based continuous speech recognition system has first proposed by Jou *et al.* (2006). It attained the accuracy of 70% for a 100-word corpus pronounced by a single speaker.

The EMG method is interesting not only in speech recognition but also in speech synthesis. In (Lam *et al.*, 2005), the authors showed that EMG features extracted from 2 channels captured from the cheek and the chin can be converted into speech signals in a frame-by-frame basis. The conversion is done via a two-layer feed-forward backpropagation neural network. The network was trained to map the short-time Fourier transformation parameters of the sEMG signals to one of six possible speech codebook indices. These indices are further decoded as Linear Predictive Coding (LPC) coefficients, pitch, and energy information to reconstruct the speech signals. The authors noted that the classification rate of the network is similar for different sEMG frame sizes on silence vectors, ranging from 83.7% to 86.4%. Classification rates of phonemes */æ/*, */i/*, */ʃ/* become higher when the sEMG frame size increases. For */æ/*, they range from 84.4% to 96.0%, for /*i*/, from 87.4% to 93.7%, and from 72.5% to 90.1% for */ʃ/*. The average classification rate is also higher for larger sEMG frame sizes and reaches 90.5% when this size is 112.5 ms. Because the conversion was performed at the phoneme level, the proposed methodology had the potential to synthesize an unlimited vocabulary size, in continuous speech. Another conversion method based on a Gaussian

Mixture Model (GMM) on the sentences pronounced by a Taiwanese was recently studied in (Toth *et al.*, 2009) with the same positions of EMG channels proposed in (Maier-Hein *et al.*, 2005). In their framework, the converted speech from EMG signals was then decoded by a speech recognizer. The best result, using an optimal EMG recognizer based on bundled phonetic features, was 18.0% Word Error Rate. The authors also claimed that there are still a number of difficulties that need to be overcome. The biggest barriers to a silent speech interface with sEMG appear to be the production of an adequate excitation signal and the different characteristics of EMG signals produced during audible and silent speech.

1.3.3 Tongue displays

Several devices are able to provide information on the movements of inner speech organs. Apart from very sophisticated medical imaging techniques such as cineradiography or real-time MRI that provide complete tongue displays, several other techniques such as ElectroMagnetic Articulography (EMA) or ultrasound (US) imaging provide partial information of the inner organs in motion: EMA provides 2D or 3D movements of a few coils glued on the tongue and possibly the velum with high precision, ultrasound imaging provides partial 2D or 3D surface of the tongue. When these data are regularized with deformable shape models acquired on the same subject using medical imaging techniques or adapted from a generic tongue model, very convincing tongue articulation, and possibly sound, can be estimated.

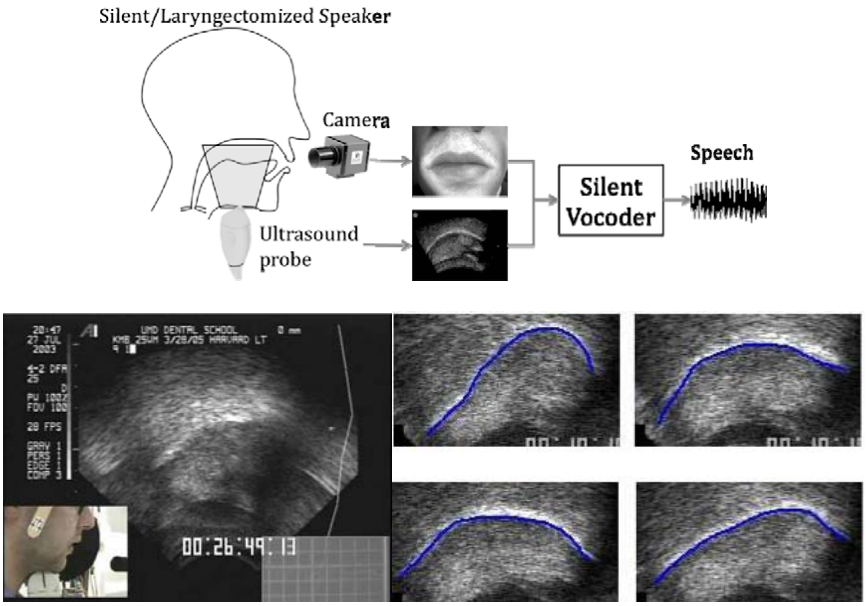


Figure 1.7. *Ultrasound-based SSI (schematic) (Denby et al., 2009)*

1.3.3.1 Ultrasound

In the SSI developed in the *Ouisper* project (Denby and Stone, 2004; Denby *et al.*, 2006, Denby *et al.*, in press; Hueber *et al.*, 2007, 2008ab, 2009), an ultrasound imaging system of the tongue is coupled with a standard video camera placed in front of the speaker's lips. Non-acoustic features, derived exclusively from spatial configurations of these two articulators, are used to drive a speech synthesizer, as illustrated in figure 1.7 (Denby *et al.*, in press).

A first version of a "visuo-acoustic" mapping task was proposed in (Denby and Stone, 2004; Denby *et al.* 2006) by using a multilayer perceptrons network. By using this network, the tongue contours and lip profiles extracted from a 2 minutes long ultrasound dataset were mapped either onto GSM codec parameters (the 13 kbit/sec codec transforms blocks of 160, 13-bit speech samples (20 ms at 8000 samples/second) into 260 bits of coded information as described in (Denby *et al.*, 2004) or onto Line Spectral Frequencies (LSF). LSF are used to represent Linear Prediction Coefficients (LPC) for transmission over a channel. LSF have several properties, e.g. smaller sensitivity to quantization noise, that make them superior to direct quantization of LPC. In a later version, extraction and parameterization of the tongue contour were replaced by *EigenTongues* decomposition which projects each ultrasound image into a representative space of "standard vocal tract configurations" (Hueber *et al.*, 2007). All these approaches, however, only predict spectral features, and thus only permit LPC-based speech synthesis, without any prescription on the excitation signal (Denby *et al.*, in press). Hueber et al (Hueber *et al.* 2007; Hueber, Chollet *et al.* 2008ab) then proposed a new framework in which a "visuo-phonetic decoding stage" was developed and combined with a subsequent concatenative synthesis procedure to produce speech. In this framework, a large audio-visual unit dictionary is constructed, which associates a visual realization with an acoustic one for each diphone. In the training stage, visual feature sequences are modeled for each phonetic class by a context-independent continuous Hidden Markov Model (HMM). In the test stage, the given sequence of visual features is decoded as a set of phonetic targets. However, the speech quality of this approach strongly depends of the phone recognition performance. The visuo-phonetic decoder is currently able to correctly predict about 60% of the phonetic target sequences and therefore the system is not able to systematically provide an intelligible output speech. Thus, improvement of the visuo-phonetic decoding stage remains a critical issue. (Hueber *et al.*, 2008b) recorded a larger audio-visual speech database with a new acquisition system which is able to record two video streams (front and profile), along with the

acoustic signal, at more than 60 frames per second (fps) instead of 30 fps for the earlier baseline acquisition system. Also, in (Hueber *et al.*, 2009), the modelling of tongue and lips feature sequences using context-dependent multi-stream HMMs has been proposed in order to improve the visuo-phonetic decoding stage. Moreover, they mention that as any speech recognition system, performance can be further improved by using a language model that can constrain the phonological lattice. The results showed that the performance improvement is about 8% higher than that of the system proposed in (Hueber *et al.*, 2008ab).

1.3.3.2 Electromagnetic Articulography (EMA)

Another method to get information on the movements of the inner organs during speech production is using electromagnetic articulography (EMA). EMA is a motion tracking method that tracks the Cartesian coordinates of sensor coils which can be fixed on specific places of the lips, the teeth, the jaw, and the velum of the subject. Usually, there are two trajectories for each coil, one for the movement in the front-back direction of the head, and one for the top-bottom direction (Toutios and Margaritis, 2005).

Fagan *et al.* (2008) proposed a silent speech recognition system based on an EMA, This system consisting of permanent magnets attached at a set of points in the vocal apparatus, coupling with magnetic sensors positioned around the user's head provides a greater flexibility in terms of placement and use. The magnets were glued to the user's tongue, lips and teeth, and six dual axis magnetic sensors were mounted on a pair of glasses (figure 1.8). The aim of this work was to assess the potential of indirect measurement of movement of the vocal apparatus as a means to determine the intended speech of a subject, rather than to develop a complete speech recognition system. As such the results show considerable promise. Based on a simple template matching algorithm, Dynamic Time Warping (DTW), it has been shown that it is possible to classify a subset of phonemes and a small number of words with degree of accuracy which is similar to that achieved for speech recognition based on acoustic information – albeit with a significantly smaller vocabulary data set. The subject was asked to repeat a set of 9 words and 13 phonemes to provide training data while 10 repetitions of each word/phone were compared to the training set template. With this condition, the recognition accuracy attained 97% for words and 94% for phonemes. The authors also noted that the processing is still able to correctly identify the best fit where the discrimination is less clear, for instance between labial phonemes (*/b/-/m/-/p/-/f/*) and velars (*/g/-/k/*), even where the difference is between voiced and unvoiced versions of the same phoneme (e.g. */g/-/k/* and */b/-/p/*). With these preliminary results, further

development of the sensing and processing systems may be possible to achieve acceptable recognition for larger vocabularies.

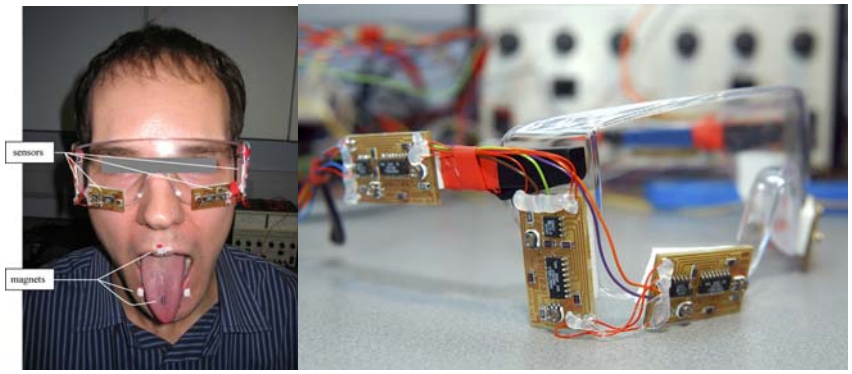


Figure 1.8. *Placement of magnets and magnetic sensors for an EMA-based SSI (Fagan et al., 2008)*

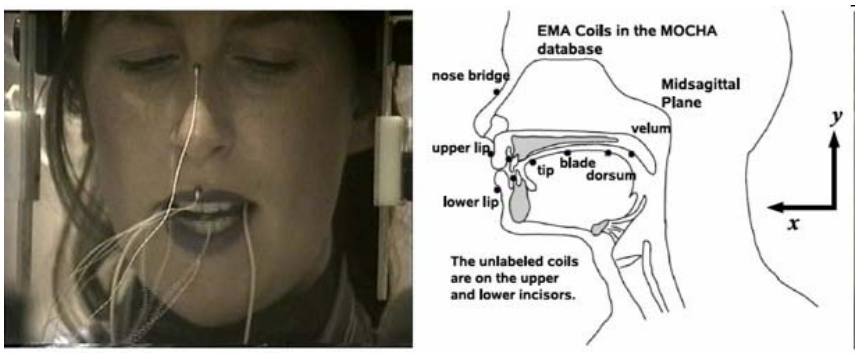


Figure 1.9. *EMA positions in MOCHA-TIMIT[2] database*

The EMA method has been used also in speech synthesis, especially for articulatory movements-to-speech spectrum conversion or inversion. Toda and colleagues (Toda, Black and Tokuda, 2008) proposed a signal-to-signal mapping technique inspired by voice conversion systems (Stylianou *et al.*, 1998, Kain and Macon, 1998abc) which does not compute any intermediate phonetic representation of the spoken message. The authors used articulatory movement data extracted from MOCHA-TIMIT database captured by EMA sensors attached on seven articulators: upper lip, lower lip, lower incisor, tongue tip, tongue body, tongue dorsum, and velum) and two reference points (the bridge of the nose and the upper incisor). The data were sampled in the midsagittal plane at 500 Hz. Each articulatory location is represented by *x*- and *y*-coordinates as shown in figure 1.9. The training corpus contains 414 sentences while 46 sentences were used for the test corpus. The authors used a GMM to model the joint probability density of an

[2] http://www.cstr.ed.ac.uk/research/projects/artic/mocha.html

articulatory parameter and an acoustic parameter. The parameters of this model are determined thanks to the training corpus by a Maximum Likelihood Estimation (MLE) within dynamic features instead of a Minimum Mean-Square Error (MMSE). They demonstrated that the converted speech by MMSE reached only 44% preference score while 47.8% was obtained for MLE without dynamic features and 58.3% for MLE with dynamic features.

1.3.4 Non-Audible Murmur (NAM) microphone

In (Nakajima *et al*, 2003ab), the author proposed that "speech is one of the actions that originate inside the human body. One of the best methods of examining what is happening in the human body is to touch it, as medical doctors have always to do first. It might be more efficient to analyze the original vibrations uttered as human speech from inside the body, through the surface of the skin, instead of analyzing the sounds in the air, after being discharged through the mouth." If this is possible, then we can put non-audible murmur into use as a new speech communication interface.

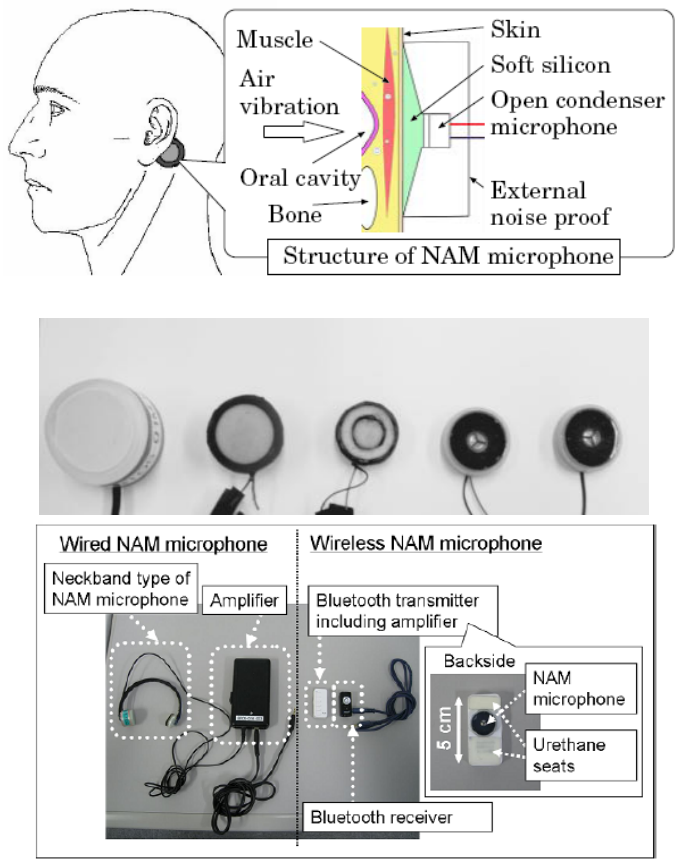


Figure 1.10. *Anatomy of NAM microphone (Toda et al., 2009; Shimizu et al., 2009)*

1.3.4.1 Anatomy of the tissue-conductive microphone

The NAM microphone was developed at Nara Institute of Science and Technology (NAIST), Japan. A NAM microphone is an electret condenser microphone (ECM) covered with a soft polymer material, which provides a better impedance matching with the soft tissue of the neck (figure 1.10). As presented in (Shimizu *et al.*, 2009), there are 2 types of NAM microphone depending what type of soft polymer material is used: soft silicone (SS) or urethane-elastomer (UE). In the soft silicone (SS) manufactured by Mitsumi Electric Co., Ltd, an ECM whose diaphragm is exposed is covered with soft silicone and placed in a rigid cylindrical case of 30mm diameter x 20mm height. The distance between the surface and the ECM diaphragm is about 1 mm. Another type, urethane elastomer (UE) made by Nakajima and colleagues has an ECM whose diaphragm is exposed and covered with urethane elastomer and placed in a rigid cylindrical case of 20mm diameter x 10mm height. The urethane elastomer is adhesive, which facilitates the attachment of the NAM microphone to the skin.

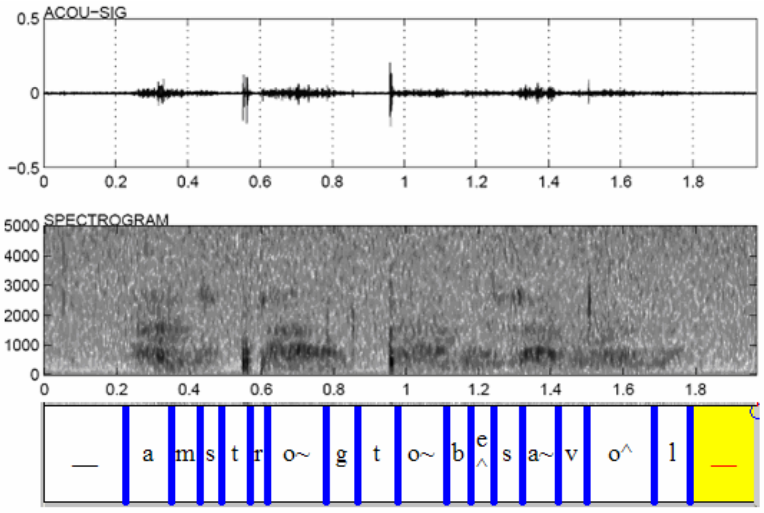


Figure 1.11. *Whispered speech captured by NAM sensor for the utterance:*

"Armstrong tombe et s'envole"

1.3.4.2 Acoustic characteristics

The NAM microphone can be placed on the skin, below the ear to capture acoustic vibrations in the vocal tract as shown in figure 1.11. This position allows a high quality recording of various types of body transmitted speech such as normal speech, whisper and NAM. Body tissue and lack of lip radiation act as a low-pass filter and the high frequency components are attenuated. However, the non-audible

murmur spectral components still provide sufficient information to distinguish and recognize sound accurately (Heracleous *et al.*, 2005). Currently, the SS-type NAM microphone can record sound with frequency components up to 3 kHz.This frequency has recently been extended to 6kHz for the UE-type (Shimizu *et al.*, 2009). Although this microphone is intrinsically not sensitive to ambient noise, when using simulated noise, its performance decreases in real noise environment because of the modified articulatory strategies due to the Lombard reflex effect (Heracleous *et al.*, 2005). Figure 1.11 shows an example of an utterance captured by this microphone. Note that the signal delivered by the NAM microphone is highly sensitive to the contacts between organs such as seen in occlusions.

The NAM signals were first exploited by the speech group at NAIST for NAM recognition (Nakajima *et al.*, 2003; Heracleous, Nakajima *et al.* 2005). The authors showed that this microphone could be used as a noise-robust sensor and could achieve the same performance as conventional speech recognition without noise. In their experiments, they achieved 93.8 % word accuracy in clean environment, and 93.1 % word accuracy in noisy environment. Moreover, by using maximum likelihood linear regression adaptation (MLLR) technique to create acoustic models for non-audible murmur, the NAM recognition system reached a very promising performance (92.1 %).

Another interesting line of research using a NAM microphone is voice generation from silent speech. In (Toda and Shikano 2005), the authors proposed to use a signal-to-signal mapping technique to convert NAM to phonated speech. It was shown that this system effectively works but its performance is still insufficient, especially in the naturalness of the converted speech. This is partly due to the poor F_0 estimation from unvoiced speech. The authors claimed that a good F_0 estimate is necessary to improve the performance of NAM-to-Speech systems. (Nakagiri *et al.*, 2006) therefore proposed another system which converts NAM to whisper. F_0 values do not need to be estimated for converted whispered speech because whisper is another type of unvoiced speech, just like NAM, but more intelligible. (Nakamura *et al.*, 2006) also used this trick in order to get a more natural and more intelligible speech to realize a speaking-aid system for total laryngectomees.

1.4 Conversion of silent speech to audible speech

When the speaker murmurs or whispers, the acoustic characteristics of the signal are modified (the details will be presented in chapter 2), and thus the linguistic

contents of the message can be strongly degraded, as well as the paralinguistic parameters (i.e. speaker emotions, attitudes and his/her identity). A crucial issue then is developing methods to capture, characterize and convert silent speech to audible voice, to generate a more understandable sound. This section presents two state-of-the-art approaches: direct signal-to-signal mapping and HMM-based recognition-synthesis.

1.4.1 Direct signal-to-signal mapping

The speech waveform carries a variety of information: segmental, supra-segmental, para-linguistic, etc. Among them, the linguistic content of the message being uttered is of greatest interest to most leading speech technologies today. However, non-linguistic information such as the speaker's mood, individuality, emotion or position with respect to what he/she says also plays a crucial part in oral communication (Moulines and Sagisaka, 1995). Voice individuality, in particular, is important not only because it helps us identify the person to whom we are talking, but also because it enriches our daily life with variety (Kuwabara and Sagisaka, 1995). This information is, of course, related to the physiological and the behavioral characteristics of the speaker. These characteristics exist both in the short-term spectral envelope (vocal tract characteristics) and in the supra-segmental features of speech (voice source characteristics).

Voice conversion is a generic term for the techniques that, from speech signal uttered by a source speaker, aim at transforming the characteristics of the speech signal in such a way that a human naturally perceives the characteristics of another target speaker in the transformed speech (Moulines and Sagisaka, 1995).

The potential applications of such techniques are numerous. First of all, voice conversion will be an essential component in text-to-speech systems based on the selection and concatenation of acoustical units. The synthesis database for such systems contains an organised collection of carefully recorded speech, and the speaker identity of the synthesis output bears resemblance to the original speaker identity of the database speaker. The creation of a synthesis database for a new synthesis voice requires a significant recording and labelling effort and a significant amount of computational resources. Voice conversion would be therefore a simple and efficient way to have the desired variety to the spoken messages while avoiding the extensive recording of hours of speech by different speakers. Only a few minutes of speech are normally sufficient for the conversion. Another application concerns interpreting telephony, which would make the communication between foreigners easier by first recognizing the speech sentences

uttered by each speaker, and then translating and synthesizing them in a different language. In recent years, voice conversion has been also used in the context of speaking aids for the speech handicaps (Nakamura *et al.*, 2006) and future "silent speech communication" in order to use silent speech as a communication medium (Toda et Shikano, 2005).

This section will review some studies carried out on voice conversion using different statistical frameworks in the last two decades, such as Vector Quantization (Able *et al.*, 1988), Linear Multivariate Regression (LMR) (Valbret *et al.*, 1992), Dynamic Frequency Warping (DFW) (Valbret *et al.*, 1992), speaker interpolation approach (Iwahashi and Sagisaka, 1994, 1995), continuous transformations based on Gaussian Mixture Model (GMM) (Stylianou *et al.,* 1998; Kain *et al.*, 1998abc, Toda *et al.*, 2003) as well as Artificial Neural Networks (ANN) and Radial Basis Function (RBF) (Narendranath *et al.*, 1995; Watanabe *et al.*, 2002).

1.4.1.1 General framework of voice conversion system

Voice conversion systems all share three main components:

- **Feature extraction**: the features represent the speaker specific characteristics in the speech waveforms such as spectral envelope, intensity, F_0 and aperiodic components which are used at each instant *t,* to represent both vocal tract and source excitation characteristics. For silent speech, i.e NAM or whisper, source excitation information is not included because it is uttered without the vibration of the vocal folds, but multi-frame representations of spectral envelope and intensity can be used to characterize silent speech signal to take into account context variation.

- **Transformation**: (a) In the training phase, the system uses aligned samples of source and target features to estimate a transformation function. (b) In the transformation phase, the estimated function is used to map the source acoustical characteristic to the target one. This can be seen as a quantization or optimization process that estimates the most probable speech signal given the source signals and an *a priori* joint model of the source and target features.

- **Speech synthesis**: i.e. a speech vocoder which converts acoustical characteristics obtained by the mapping function to speech samples.

Figure 1.12 shows a general framework for voice conversion with basic building blocks.

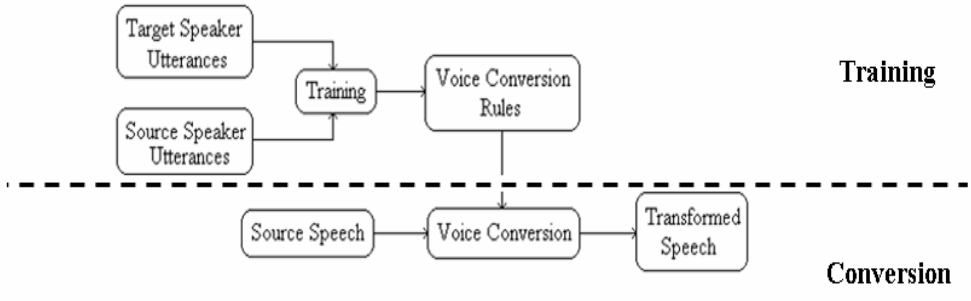


Figure 1.12. *Voice conversion framework.*

1.4.1.2 Vector quantization based voice conversion

One of the first voice conversion system based on *mapping codebooks* or *vector quantization (VQ)* was proposed by Abe *et al.* (1988). In this approach, the codevectors of a source codebook (speaker A) is mapped to the correspondent codevectors of a target codebook (speaker B). The mapping of the codebooks is constructed by first vector-quantifying frame by frame the source and target features from a learning word set. A Dynamic Time Warping (DTW) algorithm is then used to determine the correspondence between vectors of the same words for the two speakers. Finally, these vector correspondences are accumulated as histograms. The mapping codebook is defined as the linear combination of the target codevectors, using the histogram as a weighting function.

By using this approach with LPC parameters as the spectral characteristics and by using a LPC vocoder, Abe *et al.* (1988) showed that the distortion between the converted vector and the target vector decreased by 23% compared to non conversion for a female-to-female conversion task, by 45% for a male-to-male and by 64% for a male-to-female conversion. However, a basic problem with this approach lies in the discontinuities in the speech signal because the parameter space of the converted spectral envelope is discretized. Several solutions were investigated to overcome this shortcoming of the VQ-based approach, for example fuzzy VQ (Kuwabara and Sagisaka, 1995; Arslan, 1999). In this variation, the input vector is represented as a combination of several neighbouring codevectors instead of being only the nearest one. Thanks to this representation, the discontinuities in the feature space reduce and therefore the quality of the synthesized speech improves.

1.4.1.3 Voice conversion using Linear Multivariate Regression

Another approach to derive spectral transformations, borrowed from the speech recognition domain, is Linear Multivariate Regression (LMR), first proposed by

(Valbret *et al.*, 1992). The basic idea of this method is that an optimal transformation should depend on the acoustical characteristics of the sound to be converted. In this method, first, the extracted spectral envelopes, i.e. cepstral coefficients, from the source and target signals are first time-aligned by Dynamic Time Warping (DTW) technique. The next step is to partition the acoustic space of the source speaker into non-overlapping classes by means of a standard vector quantization in order to decrease the mapping complexity. The overall mapping is approximated by a finite set of elementary transforms, each of them being associated with a class. The LMR consists therefore in finding the optimal linear transformation, that is, for each class q, the matrix P_q which minimizes a "mean square" error between the set of source vectors and the set of target vectors.

$$P_q = \arg\min \sum_{n=1}^{M_q} \| (y_{n,q} - C_q^t) - P_q (x_{n,q} - C_q^s) \|^2 \tag{1.1}$$

where C_q^s and C_q^t denote the source and target empirical means of the qth class. M_q is the total number of vectors in the qth class, $x_{n,q}$ and $y_{n,q}$ with $n = 1, \ldots, M_q$ are the set of source and target vectors belonging to the qth class. For any source vector $x_{n,q}$ belonging to the class q, the transformed vector is given by:

$$\hat{y}_{n,q} = C_q^t + P_q (x_{n,q} - C_p^s) \tag{1.2}$$

This technique tries to move the formants from their initial positions towards their positions in the target space, but their amplitudes and their band-width are not well preserved. Moreover, the "hard" classification in this method introduces undesired discontinuities in the converted spectra.

1.4.1.4 Voice conversion using Dynamic Frequency Warping

The same transformation method as LMR is Dynamic Frequency Warping, proposed in the same study by Valbret *et al.* (1992). While LMR operates on the vector of cepstral coefficients, on the other hand, DFW directly operates on the spectral envelope. The aim of this method is to obtain an optimal non-linear warping function of the frequency axis to simulate changes of speaker characteristics. The DFW is therefore closely related to the acoustic theory of speech production, in the sense that changes in vocal-tract length produce a non-linear transformation of formant frequencies (Valbret *et al.*, 1992). Each pair of source and target log-magnitude spectra is first computed. Their spectral tilts are estimated to eliminate the glottal effects by fitting the envelope with a linear function of frequency, using a least-square regression line. Then, a dynamic

frequency warping algorithm is applied between each source and target residuals, the log-magnitude spectrum minus the spectral tilt. The number of warping functions is equal to the number of pairs of source-target spectral vectors within the class. In (Valbret *et al.*, 1992), the authors showed that the warping functions obtained for vectors belonging to the same class look rather similar and very few paths deviate from the main "beam". To avoid artifacts, they worked out a median warping function. The experimental results showed that LMR performs slightly better than DFW. The spectral envelope transformed by LMR is much closer to the target envelope than the one transformed by DFW. In particular, DFW can only move formant position without modifying their amplitudes, whereas LMR is able to cope with both formant frequencies and amplitudes. Also, the LMR converted speech is most often judged closer to the target speaker than the DFW converted speech (Valbret *et al.*, 1992).

1.4.1.5 Voice conversion by speaker interpolation

Iwahashi and Sagisaka (1994, 1995) stated that suitable constraints on speech spectrum consistency were not used in voice conversion based on spectral mapping techniques, such as VQ codebook mapping (Abe *et al.*, 1988), linear multivariate regression and dynamic frequency warping (Valbret *et al.*, 1992). Moreover, a large amount of spectrum data from the target speaker is needed. In the case when it is not possible to extract such large amounts of data, these methods do not work well. They proposed another technique for spectrum transformation by using speaker adaptation method which only needs a small amount of spectrum data to model the target speaker's spectral characteristics. The spectrum data of utterances spoken by a moderate number of speakers is pre-stored. In this system, after the spectrum sequences of the same utterance by multiple speakers are time-aligned by DTW, the interpolation is carried out between these time-aligned spectrum sequences by the following transformation:

$$Y_{ij} = \sum_{k=1}^{M} w_k . x_{kij} \tag{1.3}$$

under the constraint

$$\sum_{k=1}^{M} w_k = 1 \tag{1.4}$$

Here, x_{kij} represents the jth spectral parameter of the ith frame in a DTWed utterance of the kth speaker. M is the number of pre-stored speakers, the w_k is the interpolation ratio for kth speaker. Y_{ij} represents the jth spectral parameter of the ith

frame of the generated spectrum. The optimal interpolation coefficients w_1, w_2, …, w_M between pre-stored speakers are determined by minimization of the function:

$$Q(w_1, w_2, ..., w_M) = \sum_{i=1}^{N} (Y_i - y_i)^2 \tag{1.5}$$

where y_i represents the spectral vector of the ith frame of the spectrum of the target speaker. N is the number of the training samples. This minimization is done by solving normal equations.

Their experiment was carried out by using this adaptation method. The number of pre-stored speakers was 4 (2 females and 2 males). The interpolation ratio was determined by only one Japanese word, spoken by the 16 target speakers (8 females and 8 males). The distance between the spectrum of the target speaker and the one generated by the interpolation was calculated for 64 words. The authors reported that the distortion between the interpolated spectrum and the target one is reduced by about 35% compared with the distance between the spectrum of the target speaker and the spectrum of the pre-stored speaker closest to the target. Moreover, to obtain more precise adaptation by using a larger amount of training data, the transformation was represented by multiple interpolating functions in their later work (Iwahashi and Sagisaka, 1995). The multiple functions' outputs are weighted-summed, using weighting values given by a Radial Basis Function network (RBF). Using 10 training words, the reduction rate increased to 48% by this multi-functional transformation.

A derived method of the speaker interpolation, called Eigenvoices Conversion (EVC), was recently proposed for loosening the parallel training data constraint in conventional voice conversion system (Toda *et al.*,2006, 2007; Ohtani *et al.*, 2009). EVC may be used to convert a source speaker's voice to arbitrary target speakers' voices. A canonical eigenvoice GMM is first trained by using multiple parallel data sets consisting of utterance pairs of the source and multiple pre-stored target speakers. Then the speaker individuality of the converted speech can be flexibly controlled thanks to this conversion model by manually setting weight parameters. The optimum weight set for a specific target speaker is estimated using only speech data of the target speaker without any linguistic restrictions. In their experiments (Toda *et al.*, 2006), the authors found that EVC outperforms the conventional GMM-based system trained by parallel data when using a small amount of training utterances because EVC effectively uses the information extracted from a lot of pre-stored speakers as a prior knowledge. However, the performance of EVC is not significantly changed when the number of training utterances increases due to the

constant model complexity. Consequently, the performance of the conventional GMM-based VC is better than the one of the EVC when using dozens of target utterances.

1.4.1.6 GMM-based voice conversion

Gaussian mixture model

As explained in (Stylianou *et al.*, 1995), a Gaussian mixture density is a weighted sum of Q components densities modelled by normal distribution (Gaussian), and given by the equation

$$p_{GMM}(x;\alpha,\mu,\Sigma)=\sum_{q=1}^{Q}N(x;\mu_q,\Sigma_q),\quad \sum_{q}^{Q}\alpha_q=1,\ \alpha_q\geq 0 \tag{1.6}$$

where α_q denotes the prior probabilities of x having been generated by component q, and $N(x;\mu_q,\Sigma_q)$ denotes the n-dimensional normal distribution with mean vector μ_q and covariance matrix Σ_q given by

$$N(x;\mu_q,\Sigma_q)=\frac{1}{(2\pi)^{n/2}\sqrt{|\Sigma_q|}}\exp\left(-\frac{1}{2}\left(x-\mu_q\right)^T\Sigma_q^{-1}\left(x-\mu_q\right)\right) \tag{1.7}$$

The conditional probability of a GMM class q given x is derived by direct application of Bayes's rule

$$\Pr\left(q\,|\,x\right)=\frac{\alpha_q N(x;\mu_q,\Sigma_q)}{\sum_{p=1}^{Q}\alpha_p N(x;\mu_p,\Sigma_p)} \tag{1.8}$$

The Gaussian mixture model (GMM) is a classic parametric model used in many pattern-recognition techniques (Duda *et al.*, 2000) and other speech application such as speaker recognition (Reynolds and Rose, 1995) or source separation (Zhang *et al.*, 2005). In the GMM context, a speaker's voice is characterized by Q acoustic classes representing some broad phonetic events, such as vowels, nasal or fricatives. The probabilistic modelling of an acoustic class is important since there is variability in features coming from the same class due to variations in pronunciation and co-articulation. Thus, the mean vector μ_q represents the average features for the acoustic class w_q, and the covariance matrix Σ_q models the variability of features within the acoustic class (Stylianou *et al*., 1998).

The GMM parameters $\{\alpha, \mu, \Sigma\}$ are estimated by Expectation Maximization (EM) algorithm (Dempster *et al.*, 1977). EM is an iterative method for parameters estimation by maximizing the likelihood function. In the case of Gaussian mixture densities, each iteration implies two successive stages:

Expectation:

$$\Pr(q \mid x_t) = \frac{\alpha_q \mid \Sigma_q \mid^{-1/2} \exp\left[-\frac{1}{2}\left(x_t - \mu_q\right)^T \Sigma_q^{-1}(x_t - \mu_q)\right]}{\sum_{i=1}^{Q} \alpha_i \mid \Sigma_i \mid^{-1/2} \exp\left[-\frac{1}{2}\left(x_t - \mu_i\right)^T \Sigma_i^{-1}(x_t - \mu_i)\right]} \tag{1.9}$$

Calculate the conditional probability that each observation x_t is generated by class q.

Maximization:

The prior probabilities $\hat{\alpha}_q$ are re-estimated in each iteration by the mean of the posterior probabilities $\Pr(q \mid x_t)$:

$$\hat{\alpha}_q = \frac{1}{N} \sum_{t=1}^{N} \Pr(q \mid x_t) \tag{1.10}$$

The means and the covariance matrices are respectively re-estimated by the means and the covariance matrices of the observations balanced by the posterior probabilities.

$$\hat{\mu}_q = \frac{\sum_{t=1}^{N} \Pr(q \mid x_t) x_t}{\sum_{t=1}^{N} \Pr(q \mid x_t)} \tag{1.11}$$

$$\hat{\Sigma}_q = \frac{\sum_{t=1}^{N} \Pr(q \mid x_t)(x_t - \mu_q)(x_t - \mu_q)^T}{\sum_{t=1}^{N} \Pr(q \mid x_t)} \tag{1.12}$$

In these equations, N indicates the number of observations in the training data.

The widespread use of the EM algorithm stems from the facts that it guarantees a non-decreasing likelihood function after each iteration and that it provides a general yet powerful framework capable of dealing with many complicated

estimation problems (Redner and Walker, 1984). However, the EM algorithm may converge towards one local optimum. In practice, the initialization of the EM algorithm is particularly important and vector quantization is usually chosen for this task. The α_i are proportionally initialized with the number of vectors in each class while μ_i and Σ_i are initialized by the means and the empirical variances of the vectors in each class.

Using Gaussian mixture model for conversion

Voice conversion is considered as a regression process between two voices. Regression analysis is a statistical tool for the investigation of relationships between variables. Usually, the investigator seeks to ascertain the causal effect of one variable upon another. To explore such issues, the investigator assembles data on the underlying variables of interest and employs regression to estimate the quantitative effect of the causal variables upon the variable that they influence.

The goal in conversion is to estimate a mapping function which makes it possible to establish the relation between source and target feature vectors. In (Stylianou et al., 1995), the source feature space was "softly" classified by a GMM distribution. Then, parameters of this model were estimated by solving normal equations for a least-squares problem based on the correspondence between source and target features. By using cepstral distortion between the converted speech and target speech as a objective criterion, the authors demonstrated that a GMM is more efficient and robust than a VQ-based technique. In their later study (Baudoin and Stylianou, 1996), the authors concluded that a GMM is as good as or better than other approaches, such as ANN, VQ and Linear Multivariate Regression (LMR).

Kain and Macon then proposed to improve the GMM method by studying the use of GMMs for regression (Kain and Macon, 1998a,b,c). The authors stated that modeling the joint density rather than only the source density can lead to a more judicious allocation of mixture components and avoids certain numerical problems when inverting large and possibly poorly conditioned matrices (Kain, 2001).

Although the performance of GMM-based method dominates other approaches, Toda *et al.* (2005) stated that there are two important issues must be further considered. The first issue, related to the spectral discontinuities, was alleviated by considering the correlation between frames by applying a parameter generation algorithm with dynamic features. This generation algorithm was first proposed by Tokuda *et al.* (2000) for the HMM system and could be easily adapted to the GMM-based mapping (Toda *et al.*, 2005). The second issue is the over-smoothing

of converted spectra, which is inevitable in the conventional ML-based parameter estimation. The authors stated that removed variance features are regarded as a noise in modelling acoustic probability density. This smoothing causes error reduction of the spectral conversion but also causes the degradation of the converted speech quality because those removed features are still necessary for synthesizing high-quality speech. This problem was addressed by taking into account of the global variance of the converted spectra in each utterance during the training by a normal distribution (Toda *et al.*, 2005).

1.4.1.7 Voice conversion by neural networks

Vocal tract shape between two speakers is non linear and hence, non-linear Artificial Neural Networks (ANN) based method could be used for this mapping. Narendranath *et al.* (1995) proposed a system where transformation function of formants is modeled by a neural network with one input layer (3 nodes), two hidden layers (8 nodes) and an output layer (3 nodes). Five pairs of speakers are considered in their experiment, each pair consisted of a male and a female speaker. Only speech data of 5 vowels */i/, /e/, /a/, /o/, /u/* in isolated utterances from each of 5 pairs of speakers is collected. The first three formants were then extracted using a method based on minimum phase group delay functions (Murthy and Yegnanarayana, 1991). For training, 50 sentences with nearly 500 pairs of formants were used. The formant values (F1-F3) corresponding to the source speaker (male) were given as the input of this network. The desired output was the formants extracted from the corresponding frame of speech of the target speaker (female). The weights were adjusted using the backpropagation algorithm until the weights converge. The results showed that the reduction rate of the distortion between generated formants and target formants depends on the vowel. A reduction rate of 67% was found for the first formant for vowel */i/* and */a/* whereas only 35% for the vowel */o/*. The reduction rate for the second formant was 68% for the vowel */e/* and 25% for the vowel /o/ and /a/. A reduction rate for the third formant was also achieved 76% for vowel /u/ and 55% for /a/.

In another work of Watanabe *et al.* (2002), the authors proposed to use a Radial Basis Function networks to map the spectral envelope from one speaker to another. 16 LPC coefficients of a phoneme uttered by the source and target speaker are chosen as the input vector and target vector for training their system. In the conversion phase, the source speech signal is analyzed into both the LPC spectral envelopes and the LPC residual signals. The transformed LPC spectral envelopes by the trained RBF is combined with residual signals. The converted speech, matched the average of F_0 to that of target's F_0 using TD-PSOLA technique

(Moulines and Laroche, 1995) is finally produced. The experiment results of the five Japanese vowels *(/a/, /i/, /u/, /e/,* and */o/*) showed that the proposed system reduces by around 87% average cepstrum distances between the converted speech and the target speech comparing with the one between source and target speech. The listeners in their subjective test also considered the converted speech closer to the target speaker than to the source speaker around 78% times for male-to-male and 76% times for female-to-female conversion respectively.

In the recent work of Desai *et al.* (2009), the authors stated that the techniques proposed by Narendranath *et al.* (1995) and Watanabe *et al.* (2002) needed to used a tedious task of carefully preparing training data which involved manual selection of vowels or syllable regions from both the source and target speaker. In their proposed system, the spectral envelopes of the source and target are characterized by 25 mel-cepstral coefficients. These two parameters are automatically time-aligned by a DTW. Various ANN architectures are experimented to find an optimal one. They concluded that the one with 25 nodes for input and output and two hidden layers (50 nodes for each one) revealed the best results. Moreover, both objective test and subjective test showed their system is better than broadly-used GMM-based system.

1.4.2 Combination of speech recognition and speech synthesis

The combination consists in plugging a speech synthesis system to a speech recognizer. The generation is quite straight forward: the recognizer segments the speech flow into phonemic units using both signal-dependent information and a more or less sophisticated language model. A standard speech synthesis system then converts this phonetic string into a synthetic voice either using the pre-recorded modal voice of the speaker or built-in available resources. The performance of such a system is mainly dependent on the recognition performance: correct recognition will result in a perfect reconstructed speech while recognition failures or inadequate language models result in drastic degradations (Denby *et al.,* 2009).

In this framework, concatenative synthesis or HMM-based speech synthesis can be used to generate speech sound (Hueber *et al.*, 2008, 2009). But in our work, HMM-based speech synthesis as described in (Tokuda *et al.*, 2000) is chosen because this system allows the voice characteristics, speaking styles, emotions can be easily, parametrically controlled and could bring a more intimate coupling between speech recognition and synthesis components than the diphone-based concatenative

system. We concentrate here on some studies of such synthesis system after a brief introduction of HMM.

1.4.2.1 Hidden Markov Models

As described in (Rabiner, 1993), a Hidden Markov Model (HMM) is a stochastic technique for the study of the complete-incomplete data problems associated with time series. It is well suited to the incorporation of temporal information. The HMM based system is supposed to be a Markov process with unknown parameters. The challenge is thus to determine these parameters starting from other observable parameters. Then, the extracted parameters can be employed for pattern recognition. Historically, the Hidden Markov Models were introduced in 1960-70 by Baum and collaborators. They were then used in speech recognition systems as from the 80s and were applied afterwards in other fields such as the bio-informatics, artificial intelligence, gesture recognition …

Definition of a Hidden Markov Model

As described in (Huang X.D. *et al.*, 1990), a HMM model is defined as a set of states, each one of them associated with a probability distribution. The transitions between the states are controlled by a set of probabilities called transition probabilities. In a particular state, observations can be generated/observed in accordance with the associated probability distribution. In opposition to a regular Markov model where the state is directly visible by an external observer, in a HMM model, the state is not directly visible, but state-dependent observations are visible. The state sequence through which the model passes is thus hidden.

An HMM can be defined by the following elements:

- T = length of the observation sequence, $O_1, O_2, ..., O_T$
- N = number of states in the model.
- L = number of observation symbols. In the case where the observations are continuous, L is infinite. In this notation, $v = \{v_1, v_2, ..., v_L\}$ is a discrete set of symbol observations. O_t belongs to one such observation symbol.
- S = {s}, a set of states (A state can be considered to possess some measurable, distinctive properties of events). For simplicity, state i at time t may be denoted by $s_t = i$ when ambiguity does not exist.

- $A = \{a_{ij} \mid a_{ij} = \Pr(s_{t+1} = j \mid s_t = i)\}$, $a_{ij} \geq 0 \ \forall i, j$, $\sum_{j=1}^{N} a_{ij} = 1 \ \forall i$, state transition probability distribution, where a_{ij} denotes the transition probability from state i to state j.
- $B = \{b_j(O_t) \mid b_j(O_t) = \Pr(O_t \mid s_t = j)\}$, For each state, there is a corresponding output probability; and all of these output probabilities represent random variables or stochastic processes to be modelled.
- $\pi = \{\pi_i \mid \pi_i = \Pr(s_1 = i)\}$, initial state distribution.

An HMM thus can be represented by using the compact notation $\lambda = (A, B, \pi)$. Specification of an HMM involves the choice of the number of states, N, the number of discrete symbols L, and specification of three probability densities with matrix form A, B and π.

In the HMM theory, three assumptions are made for a simple mathematical model:

- Markov chain hypothesis: concerning the definition of the transition matrix A, the probability dependence is truncated to just the preceding state.
- Stationary hypothesis: the state-transition probabilities is independent of time, that is

$$\Pr(s_{t_1+1} = j \mid s_{t_1} = i) = \Pr(s_{t_2+1} = j \mid s_{t_2} = i) \quad \forall t_1, t_2 \tag{1.13}$$

- Independence hypothesis of output observations: the current observation is statistically independent of the preceding observations. Mathematically, this assumption can be formulated for a HMM λ by:

$$\Pr(O \mid s_1, s_2, ..., s_T, \lambda) = \prod_{t=1}^{T} \Pr(o_t \mid s_t, \lambda) \tag{1.14}$$

Basic algorithms for HMMs

Given the definition of HMMs, there are three key problems (Rabiner, 1989).

- Problem 1: Compute the probability $\Pr(O \mid \lambda)$ that the observed sequence $O = O_1, O_2, ..., O_T$ was produced by the model $\lambda = (A, B, \pi)$.
- Problem 2: given the observation sequence O and the model $\lambda = (A, B, \pi)$, determine what is the most likely state sequence $S = s_1, s_2, ..., s_T$ according

to some optimality criterion. This problem can be solved by *Viterbi* algorithm.

– Problem 3: given the observation sequence O, how do we adjust the model parameters $\lambda = (A, B, \pi)$ to maximize $\Pr(O \mid \lambda)$.

Problem 1 is the *evaluation* problem. Given a model and a sequence of observations, how do we compute the probability that the observed sequence was produced by the model or we can also view the problem as one of scoring how well a given model matches a given observation sequence. This problem can be solved by *Forward-backward* algorithm.

Problem 2 is the *decoding* problem in which we attempt to uncover the hidden part of the model, i.e., to find the as "correct" as possible state sequence. The choice of criterion for this problem is a strong function of the intended use for the uncovered state sequence. The decoding problem is solved by the *Viterbi* algorithm (see annex)

Problem 3 is the *estimation* problem in which we attempt to optimize the model parameters so as to best describe how a given observation sequence comes about. The observation sequence used to adjust the model parameters is called a training sequence. The training problem is the crucial one for most applications of HMMs, since it allows us to optimally adapt model parameters to observed training data – i.e., to create best models for real phenomena. This problem is solved by the *Baum-Welch* method or equivalently by the EM method (Dempster *et al.*, 1977).

1.4.2.2 HMM-based speech synthesis

Speech synthesis research aims to tackle both the two characteristics: *naturalness* and *intelligibility* of the generated speech. Naturalness describes how closely the output sounds like human speech while intelligibility is the ease to understand of the output signal. Building such a system has evolved in the last two decades from a knowledge-based (or rules-based) approach to a data-driven one (Zen, Tokuda and Black, 2009). High-quality synthetic voices may be built from sufficiently single speaker databases of natural speech rather than using each phonetic unit and its applicable contexts. Such system first proposed by Moulines and Charpentier, (1990), called diphone system. The diphone synthesis uses a minimal speech database containing all the diphones (one example for each diphone) occurring in a language and the number of diphones depends on the phonotactics of each language. Then, more general, but more resource consuming, techniques of unit-selection synthesis where appropriate sub-word units (individual phones, diphones,

syllables, words, phrases and sentences) are automatically selected from large databases of natural speech were proposed. Although these techniques have evolved to become the dominant approach, provide the greatest naturalness because they apply only a small amount of digital signal processing, there are still some limitations. When a required sentence needs phonetic and prosodic contexts that are under-represented in a database, the quality of the synthesizer can be severely degraded. These events occur frequently and a single bad join in an utterance can destroy the listeners' flow. Maximum naturalness typically require unit-selection speech databases to be very large, ranging into the gigabytes of recorded data, representing dozens of hours of speech (Kominek and Black, 2003). Moreover, as there is usually very few modifications to the selected pieces of natural speech , the output speech has the same style as the original recordings. With the need for more control over speech variations, larger databases containing examples of different styles are required. Unfortunately, this task is very difficult and costly (Black, 2003).

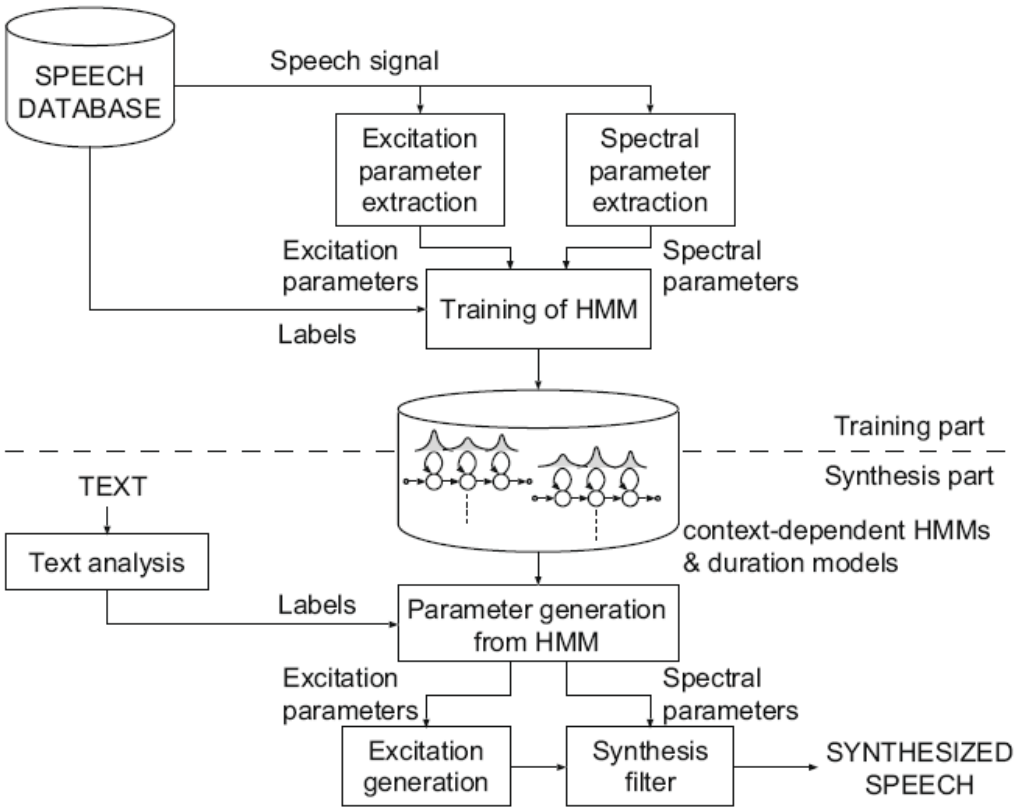


Figure 1.13. *HMM-based speech synthesis system (HTS)*

In contrast to the unit-selection synthesis which try to preserve the natural speech units from a pre-recorded database as much as possible, HMM-based speech

synthesis generates the *average* of some sets of similarly sounding speech segments. In (Zen *et al.*, 2009), the authors stated that although the best examples of unit-selection synthesis seem to be better than the best examples of statistical parametric synthesis, it appears that the quality of statistical parametric synthesis has already reached a level where it can stand in its own right. Moreover, parametric models offer other benefits related to their flexibility in changing voice characteristics, speaking styles and emotions of output speech. Figure 1.13 shows the schematic overview of a HMM-based speech synthesis system.

In a statistical parametric synthesis system, parametric representations of speech including spectral and excitation parameters are extracted from a speech corpus and then are modelled by using a set of generative models (e.g., HMMs). The parameters of these models are usually estimated by a Maximum Likelihood (ML) criterion as presented following:

$$\hat{\lambda} = \arg\max_{\lambda} \{p(O \mid W, \lambda)\} \tag{1.15}$$

where λ is a set of model parameters, O is a set of training data, and W is a set of word sequences corresponding to O. Speech parameters o, are then generated for a given word sequence to be synthesized, w, from the set of estimated models, $\hat{\lambda}$, to maximized their output probabilities as

$$\hat{o} = \arg\max_{o} \{p(o \mid w, \hat{\lambda})\} \tag{1.16}$$

Finally, a speech waveform is generated from the parametric representations of speech. HMMs have been widely used for these representations. A HMM-based speech synthesis consists of training and synthesis parts (Zen *et al.*, 2004, 2009) as presented in the figure 1.13.

- The training part performs the maximum likelihood estimation of Eq. (1.15) by using EM algorithm (Dempster *et al.*, 1977). The main difference between this training part and a speech recognition is the input features. In fact, in addition to the spectrum (e.g., mel-cepstral coefficients and their dynamic features) parameters, the excitation (e.g., $\log F_0$ and its dynamic features), usually ignored in the speech recognizer, is also extracted from a database of natural speech. Both of these parameters are modelled by a set of multi-stream (Young *et al.*, 2006) context-dependent HMMs. Another difference is that linguistic and prosodic contexts are really taken into account in addition to phonetic ones (Tokuda *et al.*, 2008).

– The synthesis part performs the maximization of Eq. (1.16). This can be viewed as the inverse operation of speech recognition. First, a given word sequence is converted into a context-dependent label sequence, and then the utterance HMM is constructed by concatenating the context-dependent HMMs according to the label sequence. Second, the speech parameter generation algorithm proposed by Tokuda *et al.* (2000) generates the sequences of spectral and excitation parameters from the utterance HMM. Finally, a speech waveform is synthesized from the generated spectral and excitation parameters using excitation generation and a speech synthesis filter.

Although the current performance of HMM-based speech synthesis using MLE criterion is quite good, two important issues are needed to solve for the training phase. The first issue is the inconsistency between the training and the synthesis phases. The MLE criterion only evaluates the model's pertinence to the data in the likelihood sense which does not reflect the final distance between the generated parameters and the target vectors. The second issue is the mutual constraint between static and dynamic features. These constraints are considered in only parameter generation while they are ignored in the training phase. The two issues above were both resolved by a trajectory model proposed by Tokuda *et al.* (2004) which can explicitly model the inter-frame dependencies and hidden dynamics in speech signal. Although this proposition implied the minimization of the error between training and generated data, the HMM training is still under the MLE framework, which cannot actually resolve the first issue. Wu *et al.* (2006) then proposed a reformulated training procedure for HMM parameters estimation by using Minimum Generation Error (MGE) criterion. In their experiments, the author showed that the synthesized speech by using MGE criterion is more natural than the one generated by MLE-based system (85.7% and 14.7% respectively in preference score).

1.5 Summary

Silent speech is a natural way that speakers use to tell private, confidential information. However, traditional speech interfaces seem to be inefficient or even incapable to capture this type of speech. As a result, a number of Silent Speech Interfaces (SSI) have been proposed to capture different type of multimodal signals delivered during the silent speech production such as EEG for cerebral activities, EMG for muscle activities, EMA, Ultrasound for articulatory movements or NAM

microphone for speech transmitted through the soft tissue of the speaker's face. From the signals captured by these interfaces, some mapping techniques have been used to generate audible speech, i.e. direct signal-to-signal mappings borrowed from speaker transformation for Text-to-Speech (TTS) system (including VQ-based, LMR, DFW, speaker interpolation, ANN and GMM mapping) or chaining speech recognition and speech synthesis in order to benefit from linguistic knowledge.

In the purpose of silent telecommunication, non audible murmur and whisper, do contain some acoustic information (compared with inner speech, for instance). Therefore, a condenser NAM microphone seems to be better than other interfaces, including EMG, EMA, Ultrasound and EEG, because of its usability and its directly familiar acoustic signal (compared with electromagnetic, electric or imagery signals). This is why we focus on the NAM microphone as one of the promising devices.

In the next chapter, we concentrate on some production and perception characteristics of whisper, one type of silent speech, which is used in this thesis. Taking these characteristics into account could be useful for the realization of a silent speech-to-audible speech conversion system.

Chapter 2

Whispered speech

2.1 Introduction

In chapter 1, I have presented some capture interfaces for silent speech and an overview of the techniques used to convert silent or whispered speech to phonated speech. This chapter concentrates on the characteristics of whispered speech, a kind of silent speech that I used for my work at GIPSA-Lab. To date, many studies on the differences between whispered speech and phonated speech have been presented in the literature from different angles such as from the point of view of production theory, perception theory and more recently from the use of computerized speech to develop new communication interfaces between humans and machines. The goal of this chapter is to review some interesting studies on whispered speech in literature.

In this chapter, section 2.2 briefly summarizes basic mechanisms of the human speech production system. The specific production of whispered speech is presented in section 2.3. Section 2.4 describes the acoustic differences between whispered speech and phonated speech. Section 2.5 reviews previous studies on perceptual characteristics of whispered speech: *"pitch-like"* perception in the absence of vocal folds vibration, intelligibility and speaker identification. Finally, section 2.6 concentrates on the contribution of visual cues as complementary information for speech intelligibility, especially in the case of whispered speech where very little acoustic information is available.

2.2 The speech production system

The speech apparatus is divided into the organs of phonation (source voice production, at the laryngeal or glottal level) and articulation (settings of the supralaryngeal or supraglottal speech organs). The phonatory organs (lungs and larynx) create voice source sounds by setting the driving air pressure in the lungs and parameters for vocal fold vibration at the larynx. The two organs together

adjust the pitch, loudness, and quality of the voice, and further generate prosodic patterns of speech (Honda, 2007). The articulatory organs (lower jaw, tongue, lips, and the velum) give resonances or modulations to the voice source and generate additional sounds for some consonants. The larynx also takes a part in the articulation of voiced/voiceless distinctions. The tongue and lower lip attach to the lower jaw, while the velum is loosely combined with other articulators. The constrictor muscles of the pharynx and larynx also participate in articulation as well as in voice quality control. The phonatory and articulatory systems influence each other mutually, while changing the vocal tract shape for producing vowels and consonants. Figure 2.1 shows a schematic drawing of the speech production system.

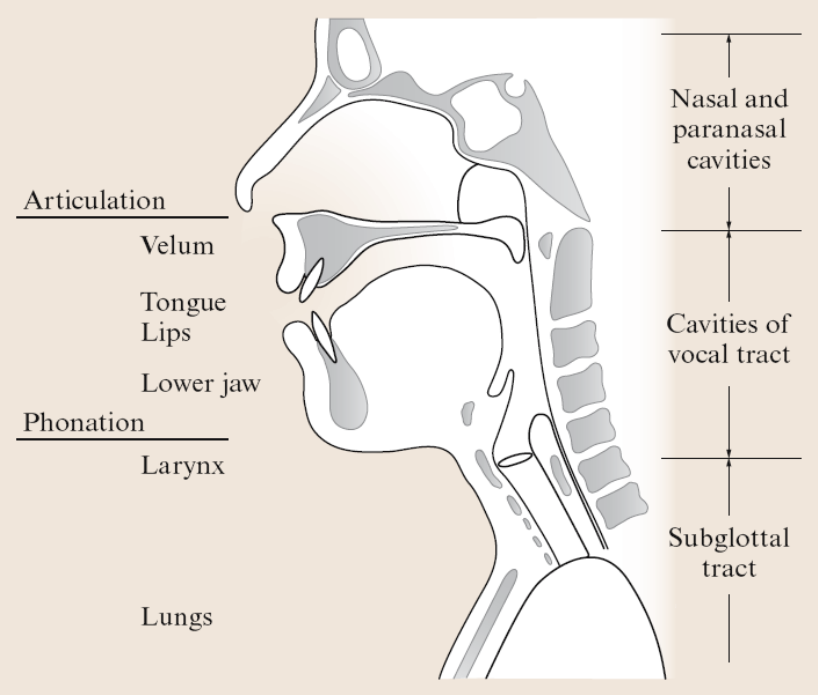


Figure 2.1. *Human speech production system (Honda, 2007)*

Speech production process is initiated by the compression of the lungs which induces a stream of air which flows through the windpipe and throat and escapes through the oral and nasal cavities. In the case of voiced sounds like */a/* and */e/*, the air flow sent from the lungs passes through an cyclically opened passage in the vocal folds, whose vibration causes that air flow to be converted into cyclic puffs of air that then become sound. In the case of unvoiced sounds like */s/* and */f/*, the air flow then passes through a narrow space formed by the tongue inside the mouth, a turbulent flow of air is produced, and this is emitted as a noise-like sound.

When we speak, we move our lower jaw, tongue, and other parts of our mouth; in fact, this changes the shape of the vocal tract and this in turn enables us to control sound resonance characteristics (Honda, 2007).

The observations on different type of sounds that the human speech production mechanism is able to produce have led to a generalized model of speech production: The speech waveform is modelled as the output of a time-varying all-pole filter driven by the source component. The source component is the glottal waveform, noise or a mixture of two. This model is known as the *source-filter model* of speech production (Stevens, 1998). This view of speech production is very powerful because it can explain the majority of speech phenomena. In the distinctions of the model, the source or excitation waveform accounts for the physiological sound sources. For example, aspiration and frication noise can be modelled as random processes, plosion as a step-function, and voicing as a pulse train. The excitation waveform can be classified into an *unvoiced* and a *voiced* signal, which can be modeled as either a random signal or an impulse-train with varying F_0 (*fundamental frequency*), respectively. Finally, the time-varying filter represents the contribution of the vocal tract shape by selectively attenuating certain frequencies of the excitation spectrum resulting in a speech spectrum with a particular spectral envelope and formant structure.

2.3 Whispered speech production

The main physiological characteristic of whisper lies in the configuration of the glottis and the epilarynx. The vocal folds are opened for whispering, thus turbulent flow produced when the air stream from the lungs is forced through this glottal constriction provides a source of sound. In normal speech, however, the vocal folds are cyclically opened and closed which produces quasi-periodic pulses of airflow. The source of whisper, therefore, is the aperiodic noise rather than quasi-periodic voice due to the lack of quasi-periodic airflow. By comparing the respiratory and laryngeal functions between whispering and normal speaking, the authors in (Hoit and Hixon, 1987) found that whispering has lower lung volumes, lower tracheal pressures and lower laryngeal airway resistances than speaking, but that whispering has higher translaryngeal flows. Tsunoda *et al.* (1991) made a physiological measurement of the laryngeal shape during whispering by using magnetic resonance imaging (MRI) and found that the supra-glottal structures were not only constricted but also shifted downward, attaching to the vocal fold to prevent vocal fold vibration. Another study by Matsuda and Kasuya (1999) used a laryngeal

endoscope inserted through the nasal tract of the subject to observe the laryngeal structure during the whispering of the five Japanese vowels */i/, /e/, /a/, /o/*, and */u/* (Figure 2.2), the author concluded that there are two major differences between the modal voice and whisper: In whisper, 1) the supra-glottal structure is constricted in the false vocal fold regions and the vocal folds are covered by the false folds; 2) the glottis is opened to a slight extent. In this study, a three-dimensional vocal tract shape measurement from a magnetic resonance image (MRI) also showed the narrowing of the tract in the false vocal fold regions and weak acoustic coupling with the subglottal system. More specifically, the laryngeal sphincter mechanism is found to be a principal contributing physiological maneuver in the production of whisper, larynx rising is more evident in whispered tense vowels than lax vowels, and the tongue root is retracted in tense vowels also (Gao, 2002).

Figure 2.2. *Laryngeal structure of modal voice (left) and whisper (right) (Matsuda, 1999)*

Recent research by Higashikawa *et al.* (2003) showed the role of lip kinematics in the production of whispered plosives (*/p/* and */b/*). The results revealed that the mean peak opening and closing velocities for */b/* were significantly greater than those for */p/* during whispered speech. No differences in peak velocity for either oral closing or opening were observed during voiced speech. Also, the maximum distance between the lips for oral opening for */b/* was significantly greater than for */p/* during whisper, whereas no difference was observed during voiced speech. These data supported the suggestion that whispered speech relies on a motor organization that is either altered or distinct from that used during normally voiced speech.

Due to these differences in the generation mechanism, the acoustic characteristics of whispered speech are different from those of phonated speech.

2.4 Acoustic characteristics of whispered speech

Due to the lack of vocal folds vibration in whispering, the most significant acoustic characteristic of whisper is the absence of fundamental frequency and consequent harmonic relationships (Tartter *et al.*, 1989). Generally, exhalation is the source of excitation in whispered speech and the shape of the pharynx is adjusted so that the vocal cords do not vibrate. "*Turbulent aperiodic airflow is therefore the only source of whisper, and it is a strong, 'rich', and hushing sound*" (Catford, 1977).

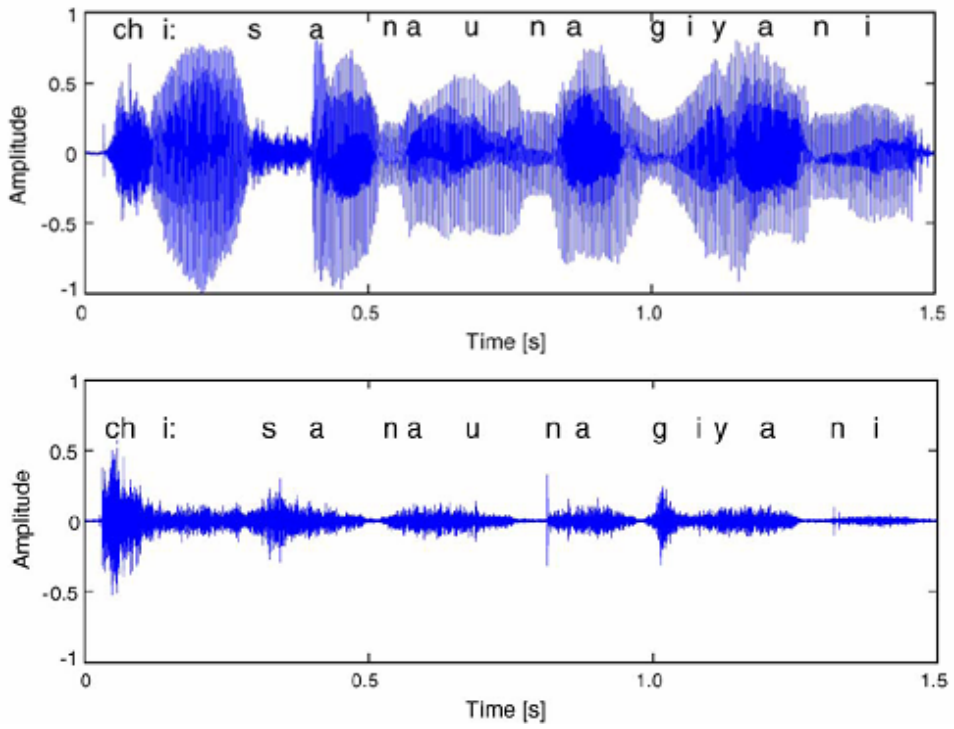


Figure 2.3. *Waveforms of the signal in normal (top) and whispered (bottom) speech modes (Ito et al., 2005)*

The acoustic energy is created by turbulence at the constriction at the vocal folds and interaction of the flow with the ventricular folds and the epiglottis. The resulting speech is completely noise excited with 20 dB lower power than its equivalent phonated speech (Jovičić *et al.*, 1996). Moreover, a study by Ito *et al.*. (2005) showed that unlike normal speech, the intensity of vowels is lower than that of consonants in the case of whispered speech. There is a significant reduction in the intensity of vowels and voiced consonants in the whispered speech compared to the phonated speech because there is no vibration of vocal folds during the production of voiced sounds in the whispered speech. However, the intensity for

unvoiced consonants is observed to be similar for normal and whispered speech. This phenomena can be observed clearly in the figure 2.3.

Most studies of whispered speech focused on vowels, especially their formantic frequencies. The upward shift of the formant frequencies for vowels in whispered speech compared to phonated speech has been found for a long time ago (Thomas, 1969, Kallail and Emanuel, 1984ab; Ito *et al.*, 2005). The studies done by Kallail and Emanuel (1984a, 1984b) or recently by Ito (2005) revealed a similar results for whispered vowel formants, which showed that the formant shift was only strongly evident for F1 and larger for vowels with low formant frequencies. The boundaries of vowel regions in the F1-F2 plane were also found to be different for phonated and whispered speech (Eklund and Traunmuller, 1996). In another experiment of Jovicic (1998), which was performed with five Serbian vowels */i/*, */e/*, */a/*, */o/* and */u/*. A total of 10 speakers, 5 males and 5 females, sustained their production of each vowel in a normal (voiced) and whispered manner. Acoustic and articulatory analysis indicated that the first and second formant frequencies of whispered vowels */i/*, */e/*, */a/* and */o/* show shifts toward higher frequencies while all formants of whispered */u/* surprisingly shifts toward lower frequencies. The authors also showed that in all cases, the formant bandwidths are expanded and are nearly constant throughout the all formants and whispered vowels.

There are also some studies on the difference between whispered and phonated consonants. Ito *et al.* (2005) found that the cepstral distances between phonated and whispered speech for vowels and voiced consonants are higher than those of unvoiced consonants. This means that the vocal tract characteristics of vowels and voiced consonants change more significantly in whisper relative to phonated speech than those of unvoiced consonants. Recent investigation of Jovičić and Šarić (2008) on Serbian consonants has a compatible result. They concluded that in intensity domain, all unvoiced consonants in whispered mode of articulation have almost unchanged intensity in comparison to phonated mode (the difference is maximum 3.5 dB). On the contrary, voiced consonants in the whispered mode were reduced in intensity by as much as 25 dB, as nasals and semivowels. Average intensity of whispered consonants is lowered by 12 dB in comparison to phonated ones, and does not depend on syllabic position inside the sentences. The authors also showed that the duration of consonants is prolonged by about 10% on average in whispered speech in comparison to phonated speech. More specifically, analysis of consonant duration in function of manner of articulation, along phonetic classes and subclasses, has showed that voiced consonants in whispering mode have prolonged duration (on average 15.3%) in comparison to unvoiced consonants (on

average 5.8%). They revealed that whispering of voiced consonants requires more articulation effort and a little more time to produce such articulation, in comparison to unvoiced consonants. The relative extension in duration of whispered consonants however depends on its position in the sentence. Jovičić and Šarić argued that the prolonged duration requires a higher precision in the motor control of vocal structures to produce an intelligible whispering. This is in agreement with the conclusion in (Rubin *et al.*, 2006) which stated that whispering causes more trauma to the larynx than normal speech. The results in this study together with the results on formant differences in phonated and whispered mode of vowel articulation in other investigations in the literature, show a high level of prosodic feature preservation (such as intensity or duration) in whispering, which is the main reason for the high intelligibility in whispered speech (Jovičić and Šarić, 2008).

2.5 Perceptual characteristics of whisper

2.5.1 Pitch-like perception

A question that has long puzzled speech scientists is how listeners identify the "*pitch*" speech sounds produced during whisper. Because of the lack of voicing, there is no fundamental frequency to guide listeners in their assignment of the whispered pitch. However, pitch is not a measure of fundamental frequency, but is defined as "*that attribute of auditory sensation in terms of which sounds may be ordered on a musical scale*" (Moore, 1997). From this definition, the study of pitch of whispered speech is valid and it is suggested that we are dealing with the pitch of something like noise bands. The purpose of this section is to clarify acoustical-perceptual relationships in identification of "*pitch*" during whispered vowel production from several previous studies.

Whispered pitch was initially explained by Von Helmholtz (1954) using the simple "listen-and-compare" method. In fact, he determined vowel resonances by listening to whispered vowels and by comparing the perceived pitches with some standard frequency source and assigned a single pitch to each of the whispered vowels. He concluded that the perceived pitches corresponded to typical values of the first formant frequency, F1, for the back vowels, and to typical values of the second formant frequency, F2, for the front vowels. Von Helmholtz's experiments with whispered vowels indicate an association between the perceived pitches of the vowels and the formant frequencies. From these results, other studies then investigated whether listeners rely more heavily on the F1 or the F2 of the

whispered speech when identifying whispered pitch. Another study of Meyer-Eppler (1957) revealed a different result by measuring the spectral characteristics of phonated and whispered German vowels on different pitch levels within a range of about a musical fifth (the subjects were asked to "sing" the first five tones of a diatonic scale: *c, d, e, f* and *g*). He found that subjective changes in pitch are apparently due either to movements of the third and higher formants (F1 and F2 remaining constant) or to changes in the intensity of the vowels. Observation of the "singing" subjects reveals also their larynx to be raised at the "higher" vowels, indicating a narrowing of the glottal fissure. The study of Thomas (1969) was with the intention of finding some compatibility in these two widely divergent views of Helmholtz and Meyer-Eppler. In his experiments, listeners were presented nine whispered vowels */i,I,ɛ,oe,ʌ,a,o,U,u/* pronounced by a male and a female speaker, then asked to adjust the frequency oscillator to match the whispered pitch. Thomas measured the formant frequencies F1 and F2 for each of these vowels and found that the frequencies associated with the perceived pitches of the whispered vowels closely approximated the frequencies of F2 for all of the vowels. These results are compatible with the findings of Von Helmholtz. In subsequent test in which the listeners were told to expect more than one pitch, two listeners identified additional pitches corresponding to F1 for the vowels */a/* and */o/* of both speakers. McGlone and Manning (1979) recorded the vowels */i,I,e,o,u/* spoken in */hV/* and */pVp/* contexts and found that listeners are more likely to rely on the resonance peak that arises from acoustical filtering in the portion of the vocal tract nearest to the lips (usually F2), even when the fundamental frequency, F0, was available in the voiced speech. Higashikawa *et al.* (1996) recorded whispered /a/ in three pitches: low, high, and ordinary. The correct order was found by a majority of the listeners (nine of the twelve speakers). The location of F1 rose significantly from low- to high-pitched whisper, and F3 was significantly higher in both the high and ordinary pitches than in the low-pitched whispers. The second formant frequency increased as well, but this was not statistically significant. In a follow-up study, male and female /a/ whispers were synthesized with the values of F1 and F2 shifted by +/- 20, 40, and 60 Hz to simulate whisper pitch. Pitch perception was stronger when F1 and F2 were moved together, and shifts in F2 created more perceptible changes in pitch than F1. The percentage of pitch matches also increased with the magnitude of the formant shifts (Higashikawa and Minifie, 1999). In a recent work on the perception of boundary tones in whispered Dutch (Heeren et Van Heuven 2009), the authors concluded that the second formant may convey pitch in whispered speech, and also that first formant and intensity differences exist between high and low boundary tones in both phonated and whispered speech. In their experiment, listeners were asked whether they were able to perceive the difference between

statements and questions as signalled by low and high boundary tones, respectively. Almost 79 % of the questions were correctly identified. This well-above-chance result shows that the rising boundary tone can be conveyed prosodically in whispered speech. Another work also shows that although whispered speech typically does not involve any vocal fold vibration, laryngeal activity may however exist during whispered speech.

From the physiological view, Coleman *et al.* (2002) have shown, using dynamic Magnetic Resonance Imaging, that larynx movements related to intonation changes (rising and falling pitch) can be observed in whispered speech. According to the authors this offers an explanation for perceived pitch in whispered speech: laryngeal movements modify the shape of the oral cavity, thus altering the vocal tract acoustics so that pitch changes can be inferred. Subvocal speech, such as the murmur that can sometimes be observed in hallucinating schizophrenic patients, has also been shown to be associated with laryngeal activity. Inouye and Shimizu (1970) reported increased EMG activity in speech-related muscles including laryngeal muscles (cricothyroid, sternohyoid, orbicularis oris, and depressor anguli oris) in 47.6% of the hallucinations of nine schizophrenic patients. These works suggest that whispered speech might carry information on laryngeal activity. This activity could be useful in the recovery of pitch because it could modify the shape of the oral cavity and hence be audible. Other sources of pitch information may exist. Zeroual *et al.* (2005) have shown with ultra-high-speed cinematography that whisper is associated with an anterior-posterior epilaryngeal compression and an abduction of the ventricular bands. These supraglottic changes may also be sources of information on pitch during whisper.

2.5.2 Intelligibility

Kallail and Emanuel (1984a,b) collected samples of phonated and whispered English vowels from male and female speakers. The vowels were */i, æ, ʌ, a, u/*. The listeners correctly categorized the phonated vowels over 80% of the time. Whispered vowels, however, were correctly identified at about a 65% rate.

Tatters' experiments had 6 speakers (three males and three females) who phonate and whisper 10 American English vowels in *[hVd]* context (Tatters *et al.*, 1991). Results exceeded 80% average identification accuracy in whisper mode, approximately a 10% falloff in identification accuracy from normally phonated speech. When they used the same vowels as in Kallail and Emanuel's experiment, the phonated and whispered accuracies were higher than the results noted by

Kallail. They concluded that the use of context from surrounding consonants makes this difference (Tatters *et al.*, 1991).

Another experiment of Tatters is about the consonants. They tested the ability of listeners to judge the voicing, place, and manner of articulation of whispered consonants. Place of a consonant is defined as the location of maximum constriction and can be categorized as labial, labio-dental, dental, alveolar, or velar. The different manners of articulation include stops, nasals, fricatives and glides. 18 consonants */b, d, g, k, m, n, p, t, r, l, w, j, f, v, s, z, sh, zh/* were whispered in the nonsense consonant-vowel /a/ syllable: */Ca/*. The overall accuracy of the identification by six listeners reached 64% while the majority of the errors occurred in the voicing decision (only 58% of accuracy for this task).

2.5.3 Speaker identification

The results of the experiments in the literature confirm the hypothesis that listeners are able to identify speaker sex from the isolated productions of whispered vowels. Schwartz and Rein tested the ability of ten listeners to determine speaker sex from whispered speech of */i/* and */a/*. In their experiment, no errors were made in the 80 identifications of speaker sex for */a/* and only 4 misidentifications were made of the 80 stimuli for */i/* (Schwartz, 1968). They concluded that the primary acoustic cue that underlies the distinction appears to be the upward frequency displacement of the resonance peaks of the female vowels. In another investigation of the relative importance of the speaker's laryngeal fundamental and vocal-tract resonance characteristics in speaker sex identification tasks from the isolated vowels */i, ɛ, æ, a, o, u/* recorded by 20 speakers (10 males and 10 females) (Lass *et al.*, 1974), the authors found that the accuracy was 95% for voiced vowels, but this drastically dropped to 75% for whispered vowels.

2.5.4 Multimodal speech perception

The contribution of the vision of the speaker's face in speech perception by humans is well known and has been well documented for a long time. Speechreading (or lipreading) is needed not only for hearing-impaired listeners (Schwartz *et al.*, 2004) but also for non-impaired hearers when auditory speech is degraded due to the ambient noise or due to the need for privacy (i.e. silent speech), in order to compensate the auditory deficit (Sumby and Pollack, 1954; Summerfield *et al.*,1989; Benoît *et al.*, 1994, 1998; Schwartz *et al.*, 2002). Vision also helps when there is no noise, but the auditory signal is degraded because we listen to speech in

a foreign language, because the speech is pronounced by a non-native speaker or because the speech is semantically complex (Reisberg *et al.*, 1987).

In the early studies of Sumby and Pollack (1954), the authors tested the speech intelligibility in noise in two conditions: audio-only and audio with visual observation of the speaker's facial and lip movements. The authors showed that when the speech signal was nearly inaudible and where only visual factors operated (S/N ratio of -30 db), the visual information contributed to about 40% of correct word perception for the 256-word vocabulary and to about 80% for the 8-word vocabulary. In contrast, under noise-free conditions, or clear speech in other words, there was little difference in the intelligibility scores associated with the two conditions. From that result, the authors concluded that under degraded acoustic conditions, visual and auditory modalities complement each other in the perception of speech and that the contribution of facial and lip movements becomes more important as the auditory perception is decreased. This conclusion was then reinforced by following studies by (Erber, 1969, 1975; Summerfield, 1979; MacLeod and Summerfield 1987; Benoît *et al.*, 1998). These studies show that what has been masked by noise in the speech spectrum can be partly recovered by the visual perception of the lips, teeth and tongue shapes that determine the place of articulation of several consonants. In another study by Benoît *et al.* (1994), the authors claimed that the complementarity between auditory and visual information provided by the vocal tract gestures is highly dependent on the phonetic context. In their experiments, stimuli of VCVCV nonsense words consisting of three French vowels (*/i/, /a/, /y/*) and six French consonants (*/b/, /v/, /z/, /Z/, /R/, /l/*), were presented under both auditory and audio-visual conditions with white noise added at various signal-to-noise ratios. The results first show that the identification scores were higher in the audiovisual condition than in the auditory-alone condition, especially in noisy acoustic situations. Secondly, the intelligibility of the 3 vowels was found to be different according to the modality. In the auditory modality, */a/* was most intelligible, followed by */i/* and then by */y/*. In the visual modality, the rounded-protruded vowel */y/* was most intelligible, followed by the open vowel */a/* and then */i/*. Thirdly, the results showed that the auditory and audio-visual intelligibility of consonants were contextually affected by the three vowels. The consonants were better identified in the */a/* context, followed by */i/* then */y/*. These results therefore support the hypothesis that vision and ausition complement eaxh other, at least in the discrimination of the 3 vowels */i/*, */a/* and */y/*.In another work by Dohen *et al.* (2004), the authors studied if the visual modality was also useful for the perception of prosody. More specifically, they tested whether the visual prosodic cues identified through an articulatory analysis were actually perceived

and could help listeners collect information about prosodic contrastive focus. In their experiment on reiterant speech, the stimuli consisted of sentences with a subject–verb–object (SVO) structure. They were recorded by a male speaker of French. Four contrastive focus conditions were studied: focus on each of the phrases (S, V or O) and broad focus. The results showed that a large jaw opening associated with a high opening velocity, a long phrase-initial lip closure and a post-focal hypo-articulation appeared to be possible visual cues to the perception of contrastive focus on reiterant speech. Moreover, when the participants were asked to identify the focus condition with only visual modality, they perceived about 86% of correct answers on average. From these results, the authors concluded that the visual modality is relevant for the perception of contrastive focus in reiterant speech in French. In further studies with non-reiterant speech, the contribution of the visual modality in the perception of contrastive prosodic focus in French was confirmed (Dohen and Lœvenbruck, 2009a; Dohen, Lœvenbruck and Hill, 2009b).

Note that, on the other hand, in the study of McGurk and MacDonald (1976), the authors showed that visual information may distort auditory perception if the acoustic and the visual cues are not coherent. It is a well known effect called the McGurk effect where an acoustic /ba/ stimulus dubbed onto a visual */ga/* stimulus is mainly perceived as */da/*. Although this effect shows that the visual articulation must be well synchronized with the auditory stream to have a benefit for speech perception, it also shows that auditory and visual information are not processed (fully) independently but are integrated in the speech perception process.

Overall, these studies suggest that it is important to develop application with virtual 3D animated talking heads that are aligned with the auditory speech (Benoît *et al.* 1998; Bailly *et al.*, 2003; Beskow, 2003; Massaro, 1998; Ouni *et al.*, 2007; Odisio *et al.*, 2004).

The studies on synthetic faces also showed that the intelligibility of the natural speech is increased when the generated facial animation are coherent and synchronized with the speech sounds (Brooke and Summerfield, 1983; Le Golf *et al.*, 1996; Benoît *et al.*, 1996). In the study by Benoît and Le Goff (1998), the authors compared the intelligibility scores obtained for visual stimuli, and audio alone stimuli under five conditions of acoustic degradation due to noise (Figure 2.4). Three types of visual stimuli were tested: 3D lip movements (controlled by five parameters and animated 25 times per second), talking head synthesis and natural video recordings. eIn the case of severely degraded acoustic information, where the subjects do not have any chance to understand the message with the audio only modality (when SNR is -18 dB in the figure), all of the three types of

visual stimuli drastically improved the scores, with the natural video recordings best identified, followed by the synthetic talking head and then by the lips alone. Specifically, with the 3D lip model, the perception score is about a third of the intelligibility observed for the natural face that reaches 60% of correct scores. With the addition of one control parameter for chin position, the lip and face models together increase the correct perception to around two thirds of the intelligibility of the natural face (Benoît and Le Goff , 1998).

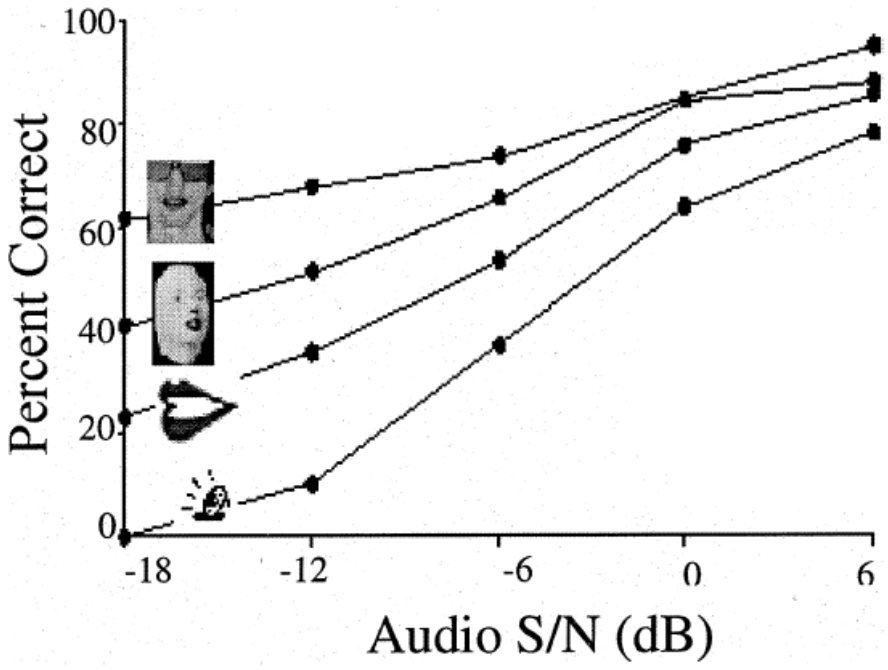


Figure 2.4. *Intelligibility scores compared across audio alone presentation of natural speech (A: bottom trace) and AV presentation of A + (from top to bottom): the front view of the original face; the face model; and the lip model. The lip and face models were controlled from measurements made from the speaker's face (Benoît and Le Goff, 1998).*

A parallel can be drawn with the subvocal visible speech that I presented in chapter 1, where there is only visual articulation without any air emission. Although synthetic visual movements are less efficient than natural one, the subjects could use synthetic visual information to compensate for the loss or degradation of auditory data.

This positive influence of synthetic visual cues will be exploited in our conversion from whispered to audible speech (chapters 4 and 5).

2.6 Summary

This chapter presents some studies from both production and perception point of views in the literature of whispered speech. The major difference between whispered speech and phonated speech is the lack of voicing and pitch cues due to the absence of vocal folds vibration. However, the investigations in the literature suggest that whispered speech might carry information on laryngeal activity that could be used in the recovery of pitch from whispered speech. Note that this recovery is a difficult task and this is one of the main reasons that limit the quality of the speech generated from whispered or non-audible acoustic signals.

This chapter also reviews the contribution of visual cues to speech perception. Facial movement information, when synchronized with speech sounds, enhances the intelligibility of speech in auditorily degraded environment, due to ambient noise or silent speech production. The promising contribution of facial movements when used as both input and output data in whisper-to-speech systems will be studied in next chapters.

Chapter 3

Audio and audiovisual corpora

3.1 Introduction

The most crucial part of any machine learning system based on a statistical framework is the data. In the case of speech, the database consists in a collection of recorded sentences, phrases, words, syllables or other units. The general purpose of a speech corpus is to acquire adequate data (i.e. acoustic-phonetic information), to provide the necessary data for training and evaluating of speech systems. Depending on the specific aim of the system, one first has to determine what kind of information must be present in the corpus to be collected. In whisper-to-speech systems, the speech corpus must satisfy particular requirements. During training, an adequate amount of data must be available for the estimation of models that represent the *mapping* between parameters of whispered speech and parameters of phonated speech. During evaluation, a sufficiently large number of additional data must be available for subjective testing.

In the design of the speech corpus for our whisper-to-speech system, four issues were at stake: database size, phonetic coverage, time alignment and multimodality.

Database size

The database size refers to the amount of data that is available in the two speech modes: whispered speech and phonated speech. In this thesis, the recording of only one speaker under studio conditions was used. In order to avoid fatigue of the speaker during the recording sessions, and thus guarantee voice quality, the number of sentences to be pronounced should not be too important. A trade-off was thus made between the minimum amount of data needed to build and test the models and the maximum number of utterances the speaker could produce.

Phonetic coverage

The phonetic coverage describes how effectively the speech utterances produced by the speaker "span" the space of possible speech sounds in the language, such as phonemes, diphones or triphones (Kain, 2001).

Time-alignment

As described in (Kain, 2001), before training the mapping functions to be used in the whisper-to-speech conversion system (based on GMM), source (whispered speech) and target (phonated speech) features must be aligned in time to display similar phonetic features. A natural way to do this task is to have the speaker utter the same sentences in both whispered and phonated mode. The use of identical sentences maximizes the probability of a consistent transcription across modes of pronunciation (whisper and speech). Note however that such speech conversion systems do not always require such constraints (Mouchtaris *et al.*, 2006; Toda *et al.*, 2006; Ohtani *et al.*, 2007). This approach ensures an accurate time-alignment of the training data with only a minimum of additional signal processing or with additional manual transcription of both whispered and phonated speech. In our case, both orthographic and audio prompts were used for recording whispered speech (cf. 3.2.2).

Multimodality

Speech is naturally multimodal. Visual cues, i.e. lip shape, tongue position and teeth visibility can compensate for lost of audio information or augment the perception of speech. In (Summerfield, 1992), the author has shown that noise up to 4-6 dB can be tolerated in speech understanding if one can see the speaker's lips. Our corpus was thus recorded with both audio and video modalities to evaluate the contribution of facial movement information in both input and output modalities of our whisper-to-speech conversion system.

To sum up, the recorded speech corpus should maximally represent all the facial movements and corresponding sounds which can be found in the language. The chosen sentences must cover maximally allophonic variations of each phoneme in context. We choose to maximize diphone coverage for our corpus. Moreover, as described above, the size of the corpus should be limited, especially in the case of a audio-visual corpus, because the speaker should not feel fatigue and also because the recording should be carried out in one day. The accurate repositioning of the NAM microphone and the markers on the speaker's face from one recording session to another is quite difficult.

In this thesis, we used two corpora, for two languages (French and Japanese), with two different speakers. The construction and the analysis of these two corpora will be presented in sections 3.2 and 3.3.

3.2 French corpus

3.2.1 Text material

As explained above, the main objective in the constitution of a speech corpus is to gather the maximum of phonetic coverage for a minimum of recording time to guarantee the constancy of the voice quality during the recording. It is thus necessary to design "dense" sentences, that is a set of homogeneous sentences, neither too short nor too long, with a simple typography and capturing the maximum of phonetic variability with a limited number of sentences (several hundreds). The *Greedy search* (Van Santen and Buchsbaum, 1997; Franois and Boffard, 2002; Bozkurt *et al.*, 2003) algorithm is often used for this task. The iterative principle of this algorithm is to eliminate the sentences whose elements are already covered by the others.

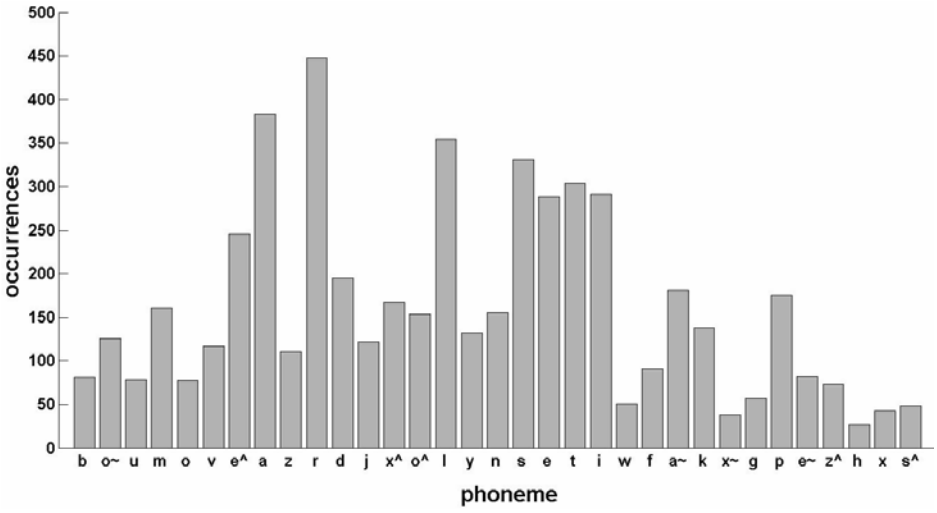


Figure 3.1. *Histogram of phoneme content of the final 288 sentences selected by the greedy search algorithm.*

For our corpus, we ran the *greedy* algorithm on a list of phonetic transcriptions of 4289 sentences, extracted from the journal *Le Monde 2003*. These sentences have from 5 to 7 words in length with no abbreviations nor acronyms. Our selection criterion was to constitute a dense list of sentences for the recording while maximizing the number occurrences of diphones. We ended up with a list of selected sentences to 288 sentences in order to limit the duration of each recording session to approximately 30 minutes. In the final selection, each phoneme was represented at least 27 times (see figure 3.1 for a phoneme histogram). The number of covered diphones is 935, with only 157 possible diphones missing.

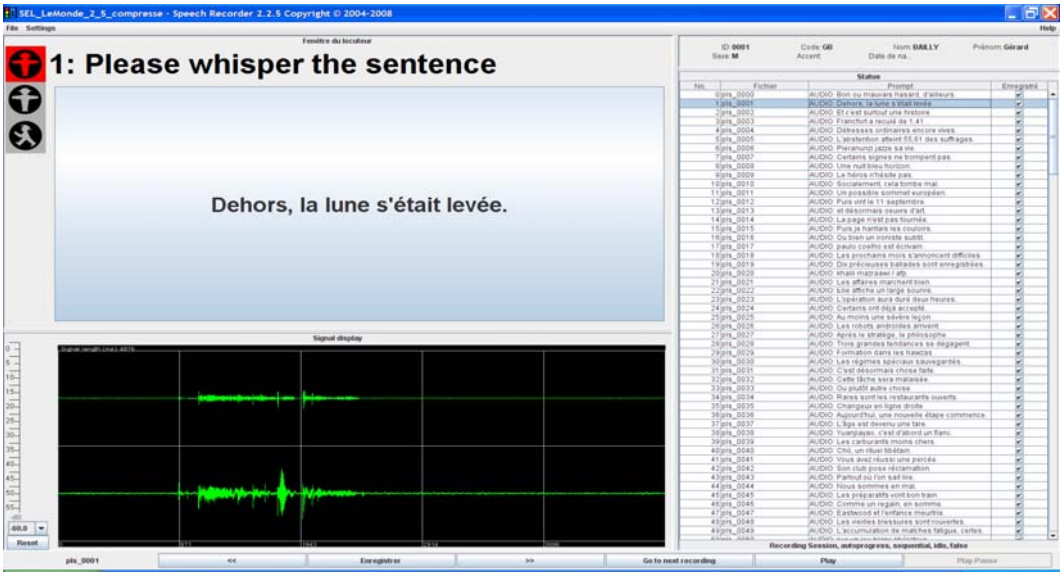


Figure 3.2. *Instructions presented on a screen guide the recording of each sentence in the corpus.*

3.2.2 Recording protocol

A trained male native speaker of French participated in the recording. The 288 sentences were first read by the speaker in a normal (voiced) speaking mode, with a natural speaking rate. Then to make sure that the whispered speech was uttered as close as possible to the normal voiced speech, that is with similar prosody (i.e. similar syllable durations, intensity and intended intonation, the speaker listened to its own voiced production of a sentence before whispering it. The sentences to be produced by the speaker in a whispered mode, were presented on a computer screen placed in front him. After a beep, the sentence was presented on the screen and the corresponding sound for the sentence produced in the normal speaking mode was played. The *SpeechRecorder Toolkit* developed by Draxler and Jänsch (2008) was used. Figure 3.2 shows an example of the software interface that guided the speaker to record each sentence. The speaker was asked to try to "mimic" the sentence in a whispered mode, with a similar intended intonation as the voiced utterance. After each sentence, the following one was immediately presented. When the speaker made a mistake, the sentence was re-recorded. Each sentence was presented for 5 seconds, which allowed enough time for the sentence to be pronounced while ensuring that all the sentences were pronounced in a limited time. All of the 288 selected sentences were recorded in one afternoon, in two different sessions: the first session for the phonated speech and the second session for the whispered speech, without any change in position for the NAM and headset microphone.

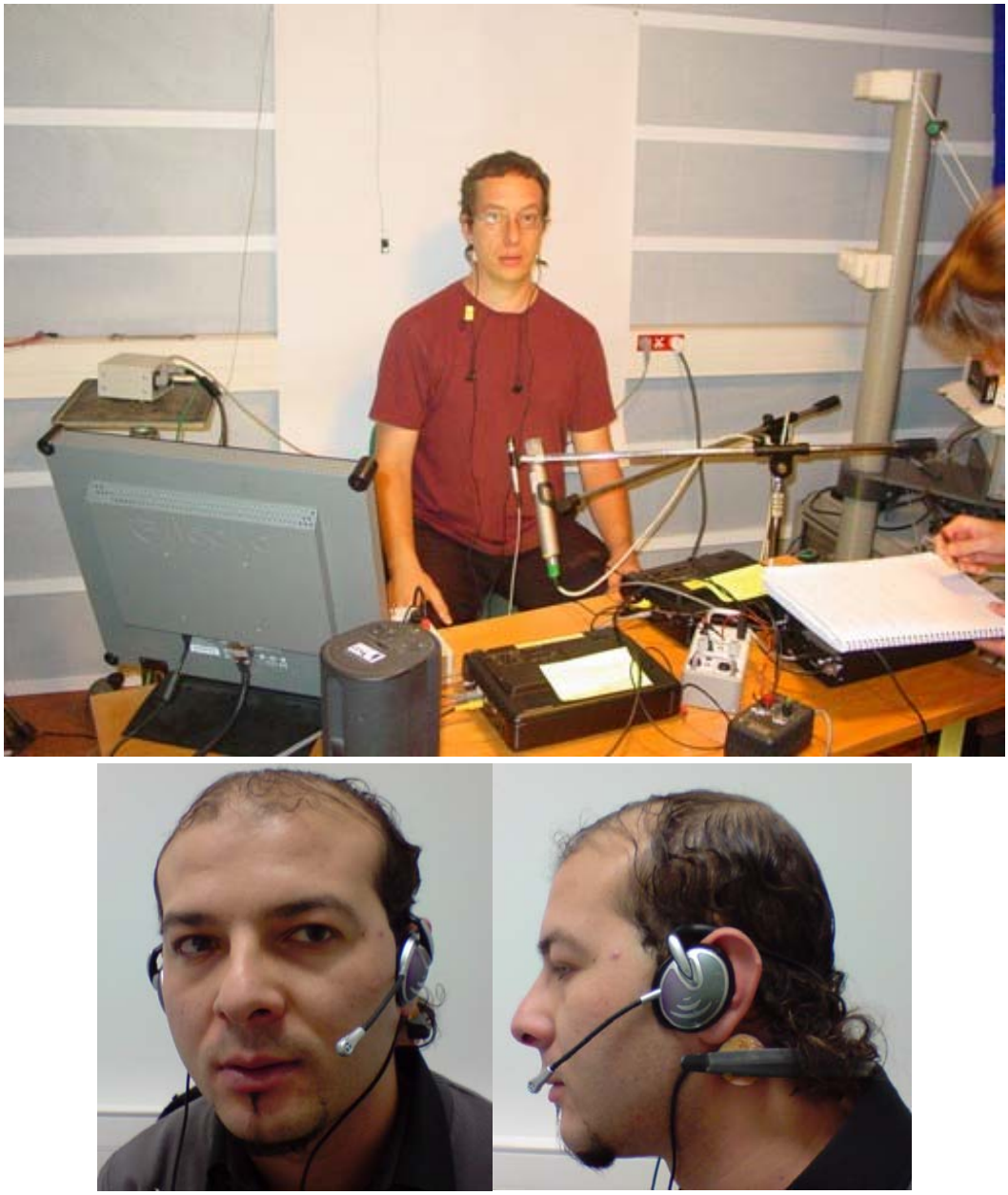


Figure 3.3. *Setup for the French corpus recording*

We recorded speech at a sampling rate of 44 kHz, using a 16 bit encoding. The speaker was seated in a professional soundproof room and wore a high quality headset microphone and a NAM microphone. The NAM microphone was fixed on the neck, under the ear of the speaker, with a plastic arch . The headset microphone was used for easing the phonetic segmentation in the GMM training phase. In addition, a calibrated microphone was placed 40 cm away from the speaker to assess the intensity of the whisper. This calibrated microphone ensured that the

whispered speech could not be heard from a distance of 40 cm, a common definition of non-audible murmur in the NAM literature. Figure 3.3 shows the setup for the recording.

3.3 Audiovisual Japanese corpus

An audiovisual corpus was recorded in order to build an audiovisual synthesis system which generates speech-related facial movements (Revéret, Bailly *et al.*, 2000; Elisei *et al.*, 2001; Badin, Bailly *et al.*, 2002; Bailly *et al.*, 2003, 2006; Gibert, 2006).

3.3.1 Text material

The text used for this corpus was composed of two parts:

- The first part was obtained from a set of 503 isolated sentences from the ATR Japanese speech database B-set. This database was collected from newspapers and magazines, by maximizing phonemic variance. The description of the text material can be found in more detail in Sagisaka *et al.* (1990).

- The second part is a list of all the Japanese consonants in symmetrical context VCV where V is one of the Japanese vowels /a/, /e/, /i/, /o/, /u/ and C is one of the Japanese consonants: /p/, /pj/, /b/, /bj/, /m/, /mj/, /d/, /t/, /s/, /ts/, /z/, /j/, /n/, /nj/, /k/, /kj/, /g/, /gj/, /f/, /Σ/, /tΣ/, /Z/, /h/, /hj/, /r/, /rj/, /w/. This second part was designed to be able to carry out a subjective evaluation of the contribution of facial movement parameters in the whispered-to-phonated speech conversion.

3.3.2 Materials and recording

The apparatus used to record this corpus is the setup traditionally used in the building of talking heads, developed at DPC (ex. ICP) (Revéret, Bailly *et al.*, 2000; Badin, Bailly *et al.*, 2002; Bailly *et al.*, 2003, 2006). The setup is composed of three analogical cameras (where uncompressed images are grabbed by three DPS stations, one for the speaker's face and two others for the profile) (figure 3.4), a helmet to maintain the speaker's head. In addition to the traditional setup, two NAM microphones, a headset microphone for the phonetic segmentation (see explanation in Chapter 4), a calibration microphone were used. The calibration

microphone was positioned at a distance of 40 cm away for the speaker and was used to check that the produced utterances were not audible at this distance. Finally, a synchronization device (see below) was used to synchronize the audio and video data and a screen computer was positioned in front of the speaker to display the sentences to be pronounced.

Figure 3.4. *Three views of the apparatus used to record the audiovisual data.Colored beads are glued on the speaker's face, to capture facial movements. The NAM microphones are strapped to the speaker's neck. The speaker wears a helmet to restrain head movements.*

The speaker was a 27-year old Japanese student at NAIST. 142 markers (small colored beads) were glued on his face and his neck.

The same protocol as the one used to record the French corpus was applied. During data acquisition, the sentences to be pronounced were displayed on a computer screen in front of the speaker.

The audio signals from a NAM microphone and a head-set microphone were recorded at the sampling rate of 44 kHz. The video are recorded at frequency of 25 Hz with interleaving, 50 Hz after de-interleaving. The visual parameters extracted from the video signals were then upsampled to 200 Hz to have a better alignment with audio parameters.

The raw audio-visual was then processed and visual parameters related to the movements of the lips, jaw and larynx were extracted by using the talking head cloning methodology that will be presented in more detail in chapter 4.

3.4 Summary

In this chapter, the data used to evaluate the contribution of the solutions explored in this thesis to improve whisper-to-clear speech conversion are presented.

The first corpus on French only consisted of acoustic signals captured by both a NAM microphone and a head-set microphone. A set of 288 phonetically-balanced sentences was recorded. The sentences were extracted from *Le Monde* journal, 2003. This French corpus was designed to test new methods to improve the existing GMM-based whisper-to-speech conversion procedure.

The second corpus on Japanese, on the other hand, is multimodal data. This corpus contains audio data (captured by two NAM microphones and a head-set microphone) and video data for face movements (captured by three cameras and using markers glued on the speaker's face). A set of 278 sentences, extracted from the 503 sentences in the ATR Japanese speech database B-set, was recorded. The audio signal was recorded in synchrony with the video signal. This Japanese corpus is designed to examine the contribution of visual movements in whisper-to-speech conversion.

The next chapters in this thesis will concentrate on our improvements to the naturalness and the intelligibility of the synthesized speech generated from whispered speech.

Chapter 4

Improvement to the GMM-based whisper-to-speech system

4.1 Introduction

As explained in chapter 1, voice transformation has been recently promoted in order to obtain quickly and inexpensively new voices for Text-to-Speech (TTS) systems. This paradigm creates a "speaker model" from a small number of speech utterances produced by the desired target speaker. Among with other different techniques already presented in chapter 1, the mapping technique based on Gaussian Mixture Model (GMM) proposed by Stylianou et *al.* (1998) is a mature technology.

As mentioned earlier, another application of voice conversion is whisper-to-speech transformation. Based on the GMM technique, the NAM-to-speech conversion system proposed by Toda and Shikano (2005) at NAIST which converts Non-Audible Murmur (NAM) to phonated speech is very promising. The quality of the converted speech is however still insufficient for computer-mediated communication, notably because of the poor estimation of F_0 from unvoiced speech and because of impoverished phonetic contrasts.

In this chapter, section 4.2 briefly presents the original NAIST system which uses a Gaussian Mixture Model (GMM) to match whispered-speech to voiced-speech. Then I explain why the original system needs improvement. The naturalness and intelligibility of the obtained converted speech is still poor, and I suggest that one of the reasons for this low quality is the weak F_0 and voicing rendering in the converted speech. This chapter therefore deals with possible improvements to the original NAIST system, concerning F_0 estimation and voicing decision. One way to do this is to use a better alignment between whispered- and voiced-speech. Section

4.3.1 describes an alignment procedure for whisper- and voiced-speech samples. This procedure has to overcome the difficulties in getting accurate phonetic segmentation in whispered-speech. The feature extraction procedure is presented in section 4.3.2. Then a method to improve voicing decision and F_0 estimation is proposed in section 4.3.3 and 4.3.4. Instead of combining voicing decision and F_0 estimation in a single GMM, a simple feed-forward neural network is used to detect voiced segments in whispered speech while a GMM estimates a continuous melodic contour based on voiced segments only. Section 4.3.5 describes how to further improve the performance of the system by using different input time window sizes and by using a different vector dimension reduction technique. Finally, I present an experiment in which I tested adding visual parameters both as input and output parameters (section 4.3.6). The contribution of these suggestions is then evaluated by subjective tests (section 4.4). Finally, the conclusion for this chapter is presented in section 4.5.

4.2 Original NAIST NAM-to-Speech system

The acoustic spectrum of NAM or whispered speech is considerably different from that of natural speech. Furthermore, NAM (or whispered speech) does not offer input F_0 characteristics since they are uttered without the vibration of the vocal folds. NAM-to-Speech or whisper-to-speech conversions thus need to estimate acoustic features of speech typically used for synthesis namely, spectral envelope, power and F_0, from the acoustic features of NAM (or whisper) namely spectral envelope and power). Figure 4.1 shows a schematic diagram of the NAM-to-Speech system proposed by Toda *et al.* (Toda and Shikano, 2005; Nakagiri *et al.*, 2006; Nakamura *et al.*, 2006; Sekimoto *et al.*, 2006) at NAIST.

The NAM-to-speech system can be decomposed into two conversion processes:

- Deriving spectral parameters from a spectral sequence of NAM or whispered speech to characterize the time-varying filter part of the converted speech signal. Mel-Frequency Cepstral Coefficients (MFCC) are used to represent the spectral envelope.
- Deriving F_0 and an aperiodic component from a spectral sequence of NAM or whispered speech to characterize the source excitation of the converted speech signal. The F_0 includes voiced/unvoiced information in this case (0-label presents unvoiced frame) and the aperiodic component includes the non-periodic part of voiced sounds (e.g. fricative noise in */v/*) or sound

emitted without any vocal fold vibration (e.g. unvoiced fricatives, or plosives).

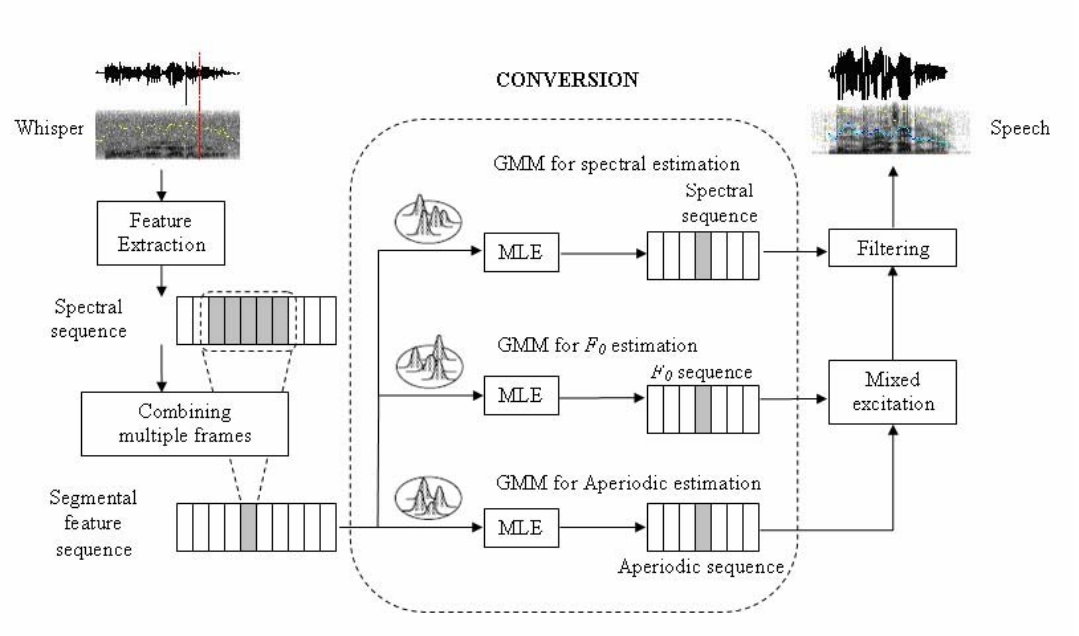


Figure 4.1. *Conversion Process of NAM-to-Speech (Toda and Shikano, 2005; Nakagiri et al., 2006)*

Finally, the mixed excitation based on estimated F_0 and aperiodic component using is then filtered with estimated spectra to generate a speech signal by a vocoder STRAIGHT (Kawahara *et al.*, 1999, 2001; Ohtani *et al.*, 2006).

4.2.1 Spectral estimation

The spectral estimation is described in Toda *et al.* (2009). The input static feature x_t and output static feature y_t at frame t are presented followed:

$$x_t = \left[x_t(1),...,x_t(D_x)\right]^{\mathrm{T}} \tag{4.1}$$

$$y_t = \left[y_t(1),...,y_t(D_x)\right]^{\mathrm{T}} \tag{4.2}$$

In order to compensate for the lost characteristics of some phonemes due to body transmission, a segment feature

$$X_t = W_x\left[x_{t-L}^{\mathrm{T}},...x_t^{\mathrm{T}},x_{t+L}^{\mathrm{T}}\right]^{\mathrm{T}} + b_x \tag{4.3}$$

extracted over several frames ($t \pm L$) is used as an input speech parameter vector where W_x and b_x are determined by Principle Component Analysis (PCA). As an output speech parameter vector, $Y_t = \left[y_t^{\mathrm{T}}, \Delta y_t^{\mathrm{T}} \right]$ is used, consisting of both static and dynamic feature vectors of the target spectrum. In order to alleviate the spectral discontinuities, these dynamic features are used with a parameter generation algorithm to take into account the correlation between frames (Toda et *al.*, 2005). Moreover, Maximum Likelihood Estimation (MLE)-based mapping is used instead of Minimum Mean Square Error (MMSE)-based mapping proposed by Stylianou *et al.* (1998) to improve the mapping performance. As stated in (Toda, Black and Tokuda, 2008), MMSE-based algorithm determines the target parameter from the given source parameter frame by frame using the minimum mean-square error (MMSE) criterion. Although it works reasonably well, it is not appropriate for multiple probability density distributions because it ignores the covariances of the individual distributions even when they are different from each other and inappropriate parameter trajectories having unnatural movements are caused by the frame-by-frame mapping process. On the other hand, in the MLE-based mapping, the determination of a target parameter trajectory having appropriate static and dynamic properties is obtained by imposing an explicit relationship between static and dynamic features. (Toda, Black and Tokuda, 2004).

As explained in Toda *et al.* (2007b, 2009), using parallel training data set consisting of time-aligned input and output parameter vectors $Z_1 = \left[X_1^{\mathrm{T}}, Y_1^{\mathrm{T}} \right]^{\mathrm{T}}$, $Z_2 = \left[X_2^{\mathrm{T}}, Y_2^{\mathrm{T}} \right]^{\mathrm{T}}, \ldots, Z_T = \left[X_T^{\mathrm{T}}, Y_T^{\mathrm{T}} \right]^{\mathrm{T}}$ as proposed by Kain *et al.* (1998abc), the joint probability density of the input and output parameter vectors is modelled by a GMM as follows:

$$P(Z_t \left| \lambda^{(Z)}) = \sum_{m=1}^{M} w_m \mathrm{N}(Z_t; \mu_m^{(Z)}, \Sigma_m^{(Z)}) \tag{4.4}$$

with

$$\mu_m^{(Z)} = \begin{bmatrix} \mu_m^{(X)} \\ \mu_m^{(Y)} \end{bmatrix} \tag{4.5}$$

and

$$\Sigma_m^{(Z)} = \begin{bmatrix} \Sigma_m^{(XX)} & \Sigma_m^{(XY)} \\ \Sigma_m^{(YX)} & \Sigma_m^{(YY)} \end{bmatrix} \tag{4.6}$$

A parameter set of the GMM is $\lambda^{(Z)}$, which consists of weights w_m, mean vectors $\mu_m^{(Z)}$ and full covariance matrices $\Sigma_m^{(Z)}$ for each individual mixture component m. The parameters GMM $\lambda^{(Z)}$ is trained in advance using aligned training data.

The spectral conversion is performed by maximizing the following likelihood function:

$$P(Y \mid X, \lambda^{(Z)}) = \sum_{all\ m} P(m \mid X, \lambda^{(Z)}) P(Y \mid X, m, \lambda^{(Z)})$$

$$= \prod_{t=1}^{T} \sum_{m=1}^{M} P(m \mid X_t, \lambda^{(Z)}) P(Y_t \mid X_t, m, \lambda^{(Z)}) \tag{4.7}$$

where m = {m_1, m_2,…, m_T} is a mixture component sequence. The conditional probability density at each frame is modeled as a GMM. At frame t, the m-th mixture component weight $P(m \mid X_t, \lambda^{(Z)})$ and the m-th conditional probability distribution $P(Y_t \mid X_t, m, \lambda^{(Z)})$ are given by

$$P(m \mid X_t, \lambda^{(Z)}) = \frac{w_m N(X_t; \mu_m^{(X)}, \Sigma_m^{(XX)})}{\sum_{n=1}^{M} w_n N(X_t; \mu_n^{(X)}, \Sigma_n^{(XX)})} \tag{4.8}$$

$$P(Y_t \mid X_t, m, \lambda^{(Z)}) = N(Y_t; E_{m,t}^{(Y)}, D_m^{(Y)}) \tag{4.9}$$

where
$$E_{m,t}^{(Y)} = \mu_m^{(Y)} + \Sigma_m^{(YX)} \Sigma_m^{(XX)^{-1}} (X_t - \mu_m^{(X)}) \tag{4.10}$$

$$D_m^{(Y)} = \Sigma_m^{(YY)} + \Sigma_m^{(YX)} \Sigma_m^{(XX)^{-1}} \Sigma_m^{(XY)} \tag{4.11}$$

In the MLE-based framework, a time sequence of the converted feature vectors is determined as follows:

$$\hat{y} = \arg\max P(Y \mid X, \lambda^{(Z)}) \tag{4.12}$$

$$\text{subject to } Y = W_y y \tag{4.13}$$

where W_y is a window matrix to extend the static feature vector sequence to the parameter vector sequence consisting of static and dynamic features (Tokuda *et al.*, 1995ab). This matrix is the 2DT-by-DT matrix written as

$$W = \left[W_1, W_2, ..., W_T\right]^{\mathrm{T}} \otimes I_{D \times D}, \tag{4.14}$$

where
$$W_t = \left[w_t^{(0)}, w_t^{(1)} \right] \tag{4.15}$$

$$w_t^{(n)} = \left[\overset{1st}{0},...,0, \overset{(t-L_-^{(n)})-th}{w^{(n)}(-L_-^{(n)})},..., \overset{(t)-th}{w^{(n)}(0)}, \overset{(t+L_+^{(n)})-th}{w^{(n)}(L_+^{(n)})}, 0,..., \overset{T-th}{0} \right]^{\mathrm{T}}, \quad n = 0,1 \tag{4.16}$$

with $L_-^{(0)} = L_+^{(0)} = 0, w^{(0)}(0) = 1.$

The likelihood function in (4.7) can be approximated with a single mixture component sequence to effectively reduce the computational cost. The likelihood is now represented as follows

$$P(Y \mid X, \lambda^{(Z)}) \approx P(m \mid X, \lambda^{(Z)}) P(Y \mid X, m, \lambda^{(Z)}) \tag{4.17}$$

In the first step, the suboptimum mixture component sequence $\hat{m}$ is estimated by

$$\hat{m} = \arg\max P(m \mid X, \lambda^{(Z)}) \tag{4.18}$$

Then the approximated log-scaled likelihood function is written as

$$L = \log P(\hat{\mathrm{m}} \mid \mathrm{X}, \lambda^{(Z)}) P(\mathrm{Y} \mid \mathrm{X}, \hat{\mathrm{m}}, \lambda^{(Z)}) \tag{4.19}$$

The converted static feature vector sequence $\hat{y}$ that maximize L under the constraint 4.13 is given by

$$\hat{y} = (W^T D_{\hat{m}}^{(Y)^{-1}} W)^{-1} W^T D_{\hat{m}}^{(Y)^{-1}} E_{\hat{m}}^{(Y)} \tag{4.20}$$

with
$$E_{\hat{m}}^{(Y)} = \left[E_{\hat{m}_1,1}^{(Y)}, E_{\hat{m}_2,2}^{(Y)},..., E_{\hat{m}_t,t}^{(Y)},..., E_{\hat{m}_T,T}^{(Y)} \right], \tag{4.21}$$

$$D_{\hat{m}}^{(Y)^{-1}} = diag\left[D_{\hat{m}_1}^{(Y)^{-1}}, D_{\hat{m}_2}^{(Y)^{-1}},..., D_{\hat{m}_t}^{(Y)^{-1}},..., D_{\hat{m}_T}^{(Y)^{-1}} \right] \tag{4.22}$$

Further, the degradation of the converted speech quality caused due to an over-smoothing of the converted spectra can be alleviated by using "*global variance*" (GV) as proposed in (Toda *et al.*, 2005). The GV of the target static feature vectors over a time sequence is written as

$$v(y) = \left[v(1),..., v(D_y) \right]^{\mathrm{T}} \tag{4.23}$$

where
$$v(d) = \frac{1}{T} \sum_{t=1}^{T} \left(y_t(d) - \frac{1}{T} \sum_{\tau=1}^{T} y_\tau(d) \right)^2 \tag{4.24}$$

A new likelihood function consisting of two probability density functions for a sequence of the target static and dynamic feature vectors and for the GV of the target static feature vectors is defined as follows:

$$P(Y\left|X,\lambda^{(Z)},\lambda^{(v)}\right.)=P(Y\mid X,\lambda^{(Z)})^{w}P(v(y)\left|\lambda^{(v)}\right.) \tag{4.25}$$

The probability density of the GV of the output static feature vectors over an utterance is also modeled by a Gaussian distribution,

$$P(v(y)\left|\lambda^{(v)}\right.)=N(v(y);\mu^{(v)},\Sigma^{(vv)}) \tag{4.26}$$

A parameter set $\lambda^{(v)}$ consists of a mean vector $\mu^{(v)}$ and a diagonal covariance matrix $\Sigma^{(vv)}$.

In the conversion procedure, the converted static feature sequence $y=\left[y_1^{\mathrm{T}},...,y_2^{\mathrm{T}},...,y_T^{\mathrm{T}}\right]^{\mathrm{T}}$ is determined by maximizing a product of the conditional probability density of *Y* given *X* and the GV probability density as follows:

$$\hat{y}=\arg\max_{y}P(Y\left|X,\lambda^{(Z)}\right.)^{w}P(v(y)\left|\lambda^{(v)}\right.) \tag{4.27}$$

Again, the approximated log-scaled likelihood function is used to effectively perform the conversion process without significant quality degradation compared with the EM algorithm (Toda *et al.*, 2007b):

$$L=\log\{P(Y\left|X,\lambda^{(Z)}\right.)^{w}P(v(y)\left|\lambda^{(v)}\right.)\}^{w}P(v(y)\mid\lambda^{(v)}) \tag{4.28}$$

The converted parameter trajectory is iteratively updated by using the first derivative given by

$$\begin{aligned}\frac{\partial L}{\partial y}&=w\left(-W^{T}D_{\hat{m}}^{(Y)^{-1}}Wy+W^{T}D_{\hat{m}}^{(Y)^{-1}}E_{\hat{m}}^{(Y)}\right)\\&\quad+\left[v'^{T}_{1},v'^{T}_{2},...v'^{T}_{t},...,v'^{T}_{T}\right]\end{aligned} \tag{4.29}$$

where

$$v'_{t}=\left[v'_{t}(1),v'_{t}(2),...,v'_{t}(d),...,v'_{t}(D)\right]^{\mathrm{T}} \tag{4.30}$$

$$v'_{t}(d)=-\frac{2}{T}p_{v}^{(d)^{T}}(v(\hat{y})-\mu_{v})(\hat{y}_{t}(d)-\overline{\hat{y}}(d)) \tag{4.31}$$

4.2.2 Source excitation estimation

Various vocoders – called Harmonic plus Noise Models (HNM) (Laroche *et al.*, 1993; Stylianou *et al.*, 1995) - used in speech synthesis or coding decompose the speech excitation into two components (Richard and Alessandro, 1996):

- A periodic or quasi-periodic component which takes into account the quasi-periodic segments of speech produced by the regular vibrations of the vocal folds. This component is characterized by the *fundamental frequency* F_0.
- an aperiodic component which corresponds to the noise source (frication, aspiration, bursts) in the speech production.

Some results (Laroche *et al.*, 1993; Dutoit and Leich, 1993) indicate that a separate processing of the periodic and aperiodic components of speech excitation may improve the quality of synthetic speech for time-scale/pitch-scale modifications.

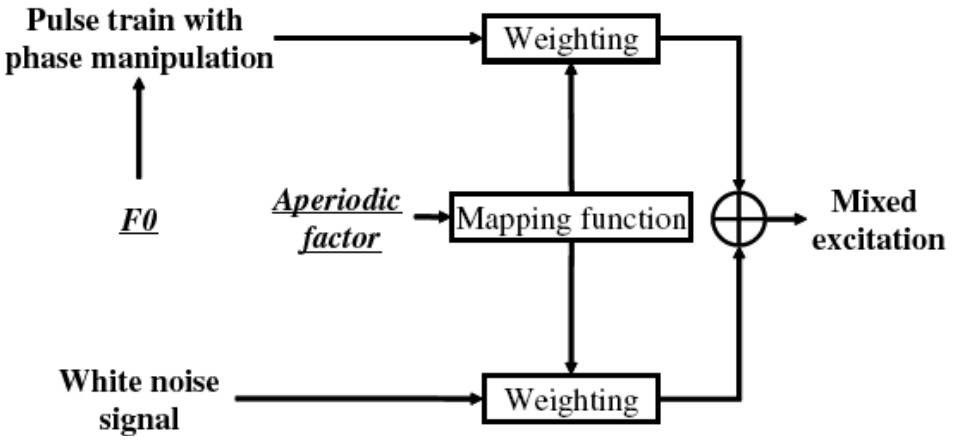


Figure 4.2. *STRAIGHT mixed excitation (Ohtani et al., 2006)*

In the STRAIGHT system (Kawahara *et al.*, 1999; Kawahara *et al.* 2001) that we use for the speech analysis and synthesis, the mixed excitation is defined as the frequency-dependent weighted sum of periodic and aperiodic components (figure 4.2).

The weight is determined based on an aperiodic component in each frequency band which is calculated from the smoothed power spectrum by a subtraction of the lower spectral envelope from the upper spectral envelope. The upper envelope is calculated by connecting spectral peaks and the lower envelope is calculated by connecting spectral valleys as presented in the figure 4.3 (Kawahara *et al.* 2001; Ohtani *et al.*, 2006). Note that the noise envelope is characterized by the energy in 5 frequency bands.

Two GMM models for F_0 estimation and aperiodic estimation (c.f. figure 4.1) are constructed in the same way as the spectral estimation except that the *global variance* (GV) is not used because GV does not cause large differences to the converted speech in the aperiodic conversion (Ohtani *et al.*, 2006). Static and dynamic features Y_t of F_0 and aperiodic components are used while keeping the same feature vector of whisper X_t as that used for the spectral estimation.

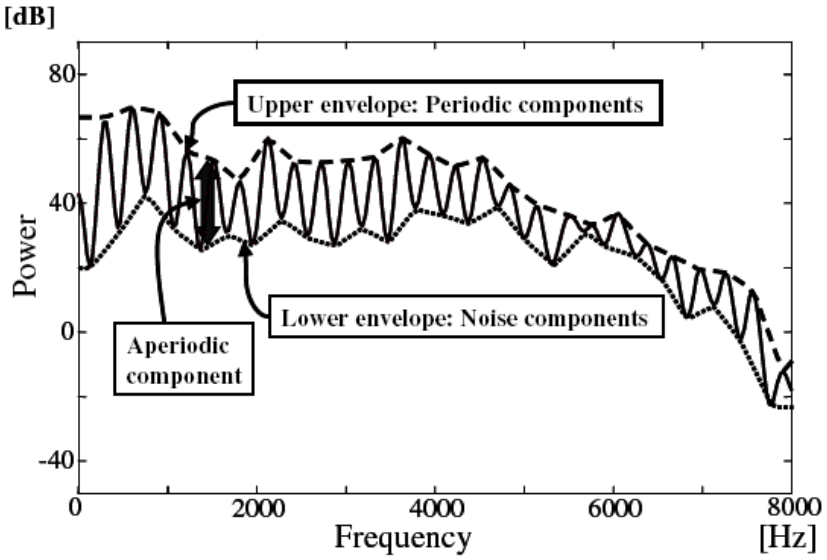


Figure 4.3. *Aperiodic component extraction from liftered power spectrum keeping periodicity. (Ohtani Y. et al., 2006).*

4.3 Improvement to the GMM-based whisper-to-speech conversion

The NAM-to-speech conversion system proposed by Toda and Shikano (2005) opens the promise for efficient silent-speech communication. Although this system successfully generates intelligible speech using only a small number of training sentences, the quality of the synthesized speech is insufficient and is far from reaching a commercial level, especially concerning the naturalness of the converted speech. This is due to the difficulties in F_0 estimation from unvoiced speech. The authors claimed that it is inevitable to improve the performance of their NAM-to-Speech system. Nakagiri and Toda *et al.* (2006) proposed another system which converts NAM-to-whisper in order to avoid this F_0 estimation. But this solution seems to be inappropriate because whisper is only another type of unvoiced speech, just like NAM, just a little more intelligible.

In this section, I will present our different investigations to improve both the intelligibility and the naturalness of the generated audible speech, especially for the F_0 contour prediction of the synthesized speech.

4.3.1 Aligning data

Due to the differences in speaking rate between recording sessions, the feature streams of both whispered speech and phonated speech must be aligned in time with each other before training the transformation function. The most straightforward way is to use manual transcription information. However, it could be more time-saving to use automatic methods such as Dynamic Time Warping (DTW) (Itakura, 1975) or a Hidden Markov Model (Kim *et al.*, 1997; Kain and Macon, 1998abc; Arslan, 1999).

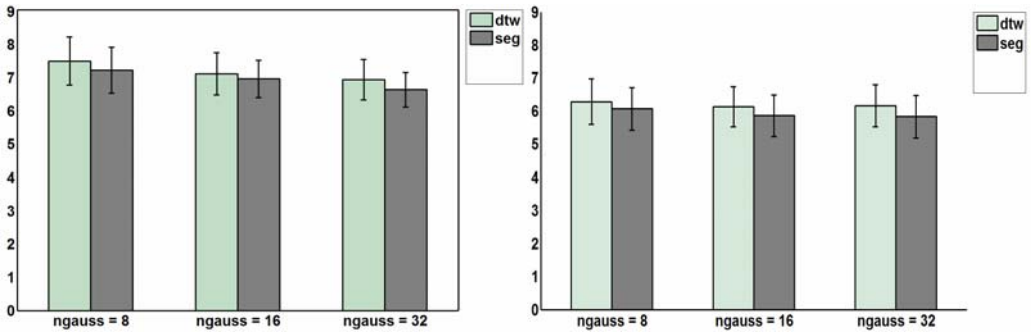
In our case, the acoustic data of both whispered speech and phonated speech was first semi-automatically segmented into phonemes. A careful manual correction was carried out after a forced alignment procedure based on a general-purpose Viterbi phone recogniser using a HMM. The HTK[3] toolkit (Young *et al.*, 2000) and the labelling software Praat[4] were used here. In this alignment paradigm, the whispered speech signal captured by a headset microphone was used instead of that captured by the NAM microphone, because the NAM microphone severely degraded the quality due to the lack of lip-radiation characteristics and the low-pass characteristics of soft tissues as well as the presence of several impulsive pop noises caused by the contact between the NAM microphone and the skin (Shimizu *et al.*, 2009).

Table 4.1. *Cepstral distortion (dB) between converted speech and target audible speech by using phonetic segmentation (seg) and by using DTW.*

Number of Gaussians	French corpus		Japanese	
	Seg	DTW	Seg	DTW
8	7.23 (±0.69)	7.50 (±0.73)	6.06 (±0.64)	6.29 (±0.648)
16	6.96 (±0.56)	7.12 (±0.64)	5.99 (±0.63)	6.13 (±0.61)
32	6.64 (±0.53)	6.94 (±0.61)	5.83 (±0.64)	6.16 (±0.64)

[3] http://htk.eng.cam.ac.uk/
[4] http://www.fon.hum.uva.nl/praat/

This segmentation information was then used to obtain a better alignment between the two different modes of pronunciation than the one obtained with the blind dynamic time warping (DTW) used in the NAIST system.



(a). French data *(b). Japanese data*

Figure 4.4 *Mean cepstral distortion and standard deviation between converted speech and target audible speech by using phonetic segmentation and by using DTW.*

Both the French and Japanese corpora presented in chapter 3 were used to evaluate the alignment. For the French corpus, the training data consisted of 200 utterances and 70 utterances were used for the test. For the Japanese corpus, the training data consisted of 150 sentences and 40 sentences were used for the test corpus. The cepstral distortion between the spectral parameters of converted speech from whispered speech captured by NAM microphone and phonated speech was used as a quantitative measurement. 25 mel-cepstral coefficients were used for the French data and 20 mel-cepstral coefficients were used for the Japanese data. A better alignment does provide a smaller distortion (cf. table 4.1, figure 4.4). The results of an ANOVA with the alignment method as a 2-level factor show that we have an improvement by using the segmentation information compared to DTW (($F(1,70) = 17.26$, $p < 0.001$) for the French corpus and ($F(1,40) = 11.32$, $p < 0.001$) for the Japanese corpus).

4.3.2 Feature extraction

The French corpus presented in chapter 3 was used to evaluate our system and the original system proposed by NAIST.

The training corpus consists of 200 utterance pairs of whisper and speech uttered by a native male speaker of French. Respective speech durations are 4.9 minutes for whisper (9.7 minutes with silences) and 4.8 minutes for speech (7.2 minutes with silences).

The 0th through 24th mel-cepstral coefficients (MFCC) are used as spectral features at each frame (with 25 ms Hanning window, shifted at the rate of 5 ms) for both whisper and phonated speech. The spectral segment features of whisper are then constructed by concatenating feature vectors at each current whispered frame ± 8 frames (i.e. representing 105 ms of speech) to take into account the context variation. Then the vector dimension is reduced to 50 using a Principal Component Analysis (PCA) technique. Log-scaled F_0 extracted with fixed-point analysis (Kawahara *et al.*, 1999) and 5 average dB values of the aperiodic components on five frequency bands (0-1, 1-2, 2-4, 4-6, 6-8 kHz) are used to characterize the excitation feature (Ohtani *et al.*, 2006). All these parameters were extracted using the STRAIGHT toolkit (Kawahara *et al.*, 1999).

The test corpus consists of 70 additional utterance pairs of whisper and audible speech not included in the training data. This corpus is used to evaluate the performance of the two systems.

4.3.3 Voiced/unvoiced detection

In the original system, all whispered feature segments (including voiced and unvoiced segments) are used to train the GMM model estimating F_0 values for converted speech. In this training paradigm, the undefined F_0 value on each unvoiced segment is assigned to 0. Therefore, in the conversion phase, F_0 and voicing are then jointly estimated by a unique GMM: voicing decision is determined using a very simple threshold. If the predicted F_0 value at the current frame is greater than the threshold, this segment is considered as a voiced segment, otherwise this segment is considered as an unvoiced segment and the value of F_0 is then set to 0. The main drawback of this approach is that some Gaussian components are wasted for representing null F_0 values for unvoiced segments in the training phase. The risk is also to predict unstable F_0 values for unvoiced segments incorrectly labelled as voiced in the conversion phase.

In order to improve the naturalness of this whisper-to-speech system by a better voicing decision and a better F_0 estimation, we estimate each one by two separate models instead of combining them in a single GMM. Our system predicts a continuous melodic contour that is sampled by voiced/unvoiced decision (Tran *et al.*, 2008b).

A non-linear feed-forward neural network is used to predict the voiced segments from the whispered speech. Since the optimization of a neural network is a difficult

task, I do not deal with this problem in this thesis. The topology of the network is empirically defined as follows:

- 50 input nodes corresponding to the dimension of the whispered features used to train the GMM.
- One hidden layer with 17 nodes. The best number of hidden nodes depends on several variables: the numbers of input and output units, the number of training cases, the amount of noise in the targets, the complexity of the function or classification to be learned, the architecture, the type of hidden unit activation function, the training algorithm, etc... In most situations, there is no way to determine the best number of hidden units without training several networks and estimating the generalization error of each. If we have too few hidden units, high training error and high generalization error occur due to *underfitting* and high statistical bias. If we have too many hidden units, we have a low training error but still have a high generalization error is occurred due to *overfitting* and high variance (Xu and Chen, 2008). For our network, we use a very simple "*rule of thumb*" proposed in (Blum, 1992, p. 60): "*A rule of thumb is for the size of this [hidden] layer to be somewhere between the input layer size ... and the output layer size ...*".
- One output node corresponding to the voiced/unvoiced decision: for tracking, the network assign 0 to unvoiced segments and 1 to voiced segments. For generating, the predicted output is larger than .5, the segment is labeled as voiced, otherwise, it is labeled as unvoiced.

The features at each frame of the whispered utterances extracted from the training corpus for the GMM estimation (including spectral, F_0 and aperiodic component estimation) are used as input vector for this network. The voiced/unvoiced label for each segment in the training whispered data is obtained from the voiced/unvoiced label of the corresponding speech utterance by aligning the two utterances as described above.

The neural network is trained by the *BackpropMomentum* learning algorithm (Bishop, 1996), using the Stuttgart Neural Network Simulator (SNNS) (Zell *et al.*, 1996). The configuration parameters for this algorithm are as followed:

- η = 0.01: learning parameter, specifies the step width of the gradient descent.

- μ = 0.01: momentum term, specifies the amount of the old weight change (relative to 1) which is added to the current change.
- c = 0.05: flat spot elimination value, a constant value which is added to the derivative of the activation function to enable the network to pass at spots of the error surface.
- d_{max} = 0.2: the maximum difference $d_j = t_j - o_j$ between a teaching value t_j and an output o_j of an output unit which is tolerated.

The general formula for back-propagation used here is

$$\Delta w_{ij}(t+1) = \eta \delta_j o_i + \mu \Delta w_{ij}(t)$$

$$\delta_j = \begin{cases} (f'_j(net_j) + c)(t_j - o_j) & if\ unit\ j\ is\ a\ output-unit \\ (f'_j(net_j) + c)\sum_k \delta_k w_{jk} & if\ unit\ j\ is\ a\ hidden-unit \end{cases}$$

with f is the sigmoid activation function in each unit : $f(x) = \frac{1}{1+e^{-x}}$

Table 4.2 and figure 4.5 show the voiced/unvoiced detection by this network compared with the GMM method in the original system which uses a single GMM with F_0 estimation. The number of Gaussian mixtures in this model is fixed at 16. Compared with the error in the original system, the statistical analysis shows that we have a significant improvement of the voiced/unvoiced detection (6.8% compare with 9.3% ($F(1,70)=24.33$, $p<0.001$)). In this table, we additionnally describe two types of error: the voiced error is the proportion of unvoiced frames that the system detects as voiced one and the unvoiced error is the proportion of voiced frames that the system detects as unvoiced. The ANOVA analysis also showed that we have a significantly better performance in both cases: 2.3% compared with 3.2% for the voiced error ($F(1,70)=10.54$, $p=0.0014$) and 4.5% compared with 6.1% for the unvoiced error ($F(1,70)=8.45$, $p = 0.042$).

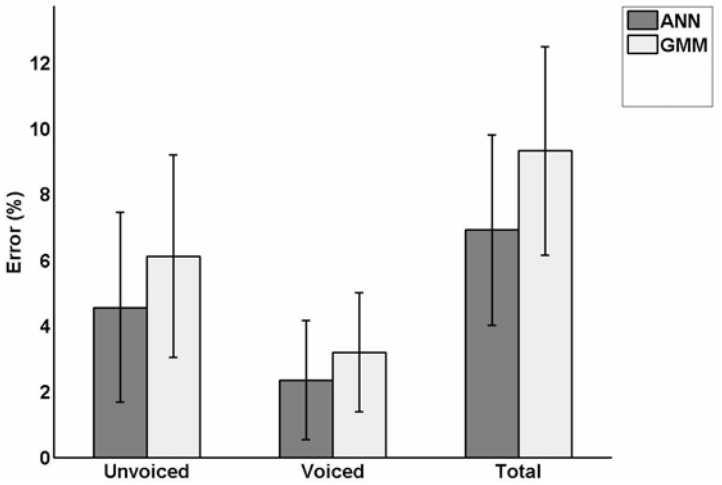


Figure 4.5 *Mean and standard deviation error of voicing decision by a neural network (ANN) and by a GMM.*

Table 4.2. *Voicing error using neural network and GMM on the French data*

Type of error	Feed-fwd NN (%)	GMM (%)
Voiced	2.3 (± 1.81)	3.2 (± 1.82)
Unvoiced	4.5 (± 2.88)	6.1 (± 3.07)
Total	**~ 6.8** (± 2.89)	~ 9.3 (± 3.16)

4.3.4 F0 estimation

Only voiced segments in whispered speech were used to train the GMM for the F_0 estimation in order to avoid wasting Gaussian components for unvoiced frames. Continuous F_0 values are then predicted. They are then sampled by voicing decisions made by the Artificial feed-forward Neural Network (ANN) presented above. This study was presented in (Tran *et al.*, 2008b). Figure 4.6 presents the synopsis of the system.

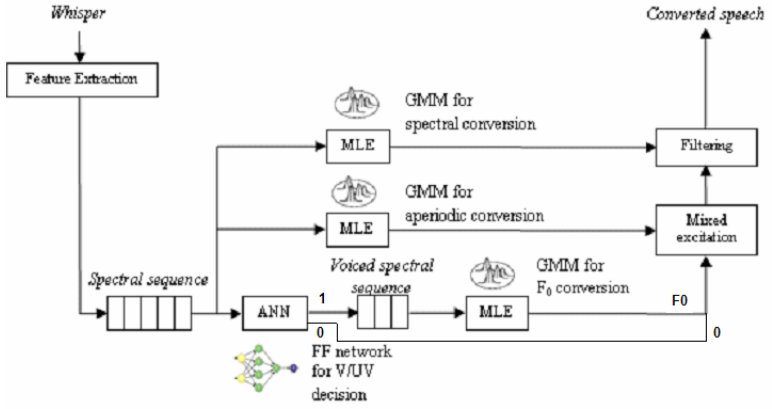


Figure 4.6. *Synopsis of the whisper-to-speech conversion system*

We compared the F_0 rendering of our system with the original one, with different numbers of mixtures (8, 16, 32 and 64) for both training and test data. Full covariance matrices are used for both GMMs. The performance of the system is measured as the normalized difference between synthetic F_0 and natural F_0 in the voiced segments that were well detected by the two systems. This difference, which we will refer to as the "deviation from natural F_0", is given by the following formula:

$$Diff = \frac{1}{N}\sum_{i=1}^{N}\frac{\left|synthetic\_F_{0i} - natural\_F_{0i}\right|}{natural\_F_{0i}} \times 100\%$$

where N is the number of frames that well detected by the two systems.

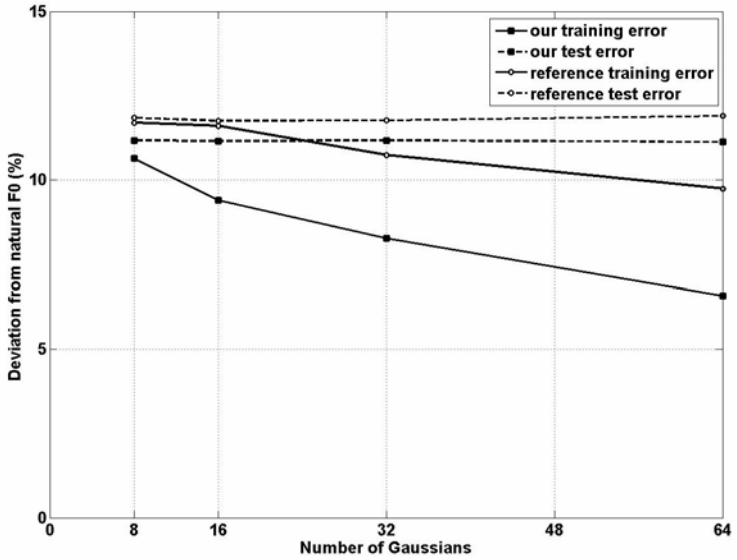


Figure 4.7. *Parameter evaluation on training (solid line) and test corpus (dashed line).*

Figure 4.7 shows that the proposed framework outperforms the original system, since the deviation from natural F_0 is lower. In addition, when the number of Gaussian mixtures increases, the deviations of both systems on the training data decrease, but the deviations on the test data are almost insensitive to the number of mixtures.

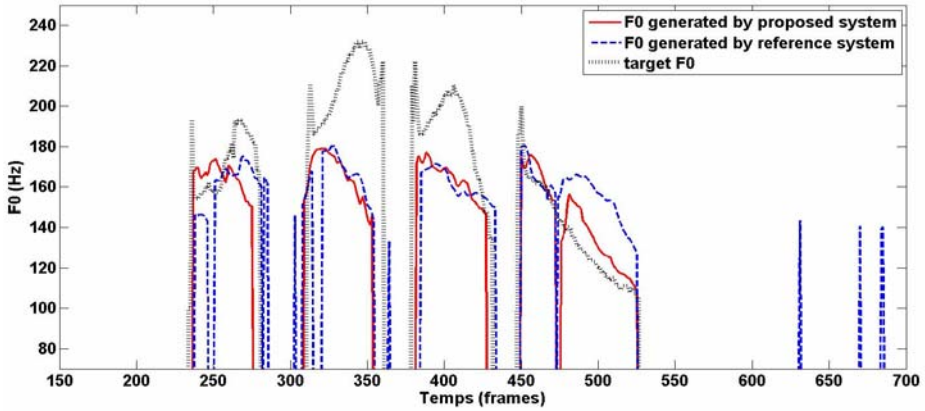


Figure 4.8. *Natural and synthetic F_0 curve for the test utterance: "Armstrong tombe et s'envole"*

Figure 4.8 shows an example of a natural (target) F_0 curve and the synthetic F_0 curves generated by the two systems. It shows that our proposed system is closer to the natural F_0 curve than the original system, especially for the final contour. An ANOVA analysis showed that this improvement is statistically significant ($F(1,70) = 20.72$, $p < 0.001$).

Figure 4.9 shows an example of whispered-, converted- (by our system) and natural-speech. As can be seen the formant patterns of the converted speech are flatter than those of natural speech. Global variance was used to partly attenuate this difference (Toda *et al.*, 2004).

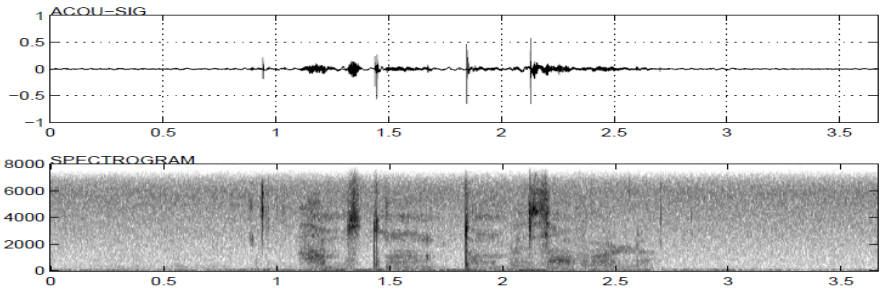
4.3.5 Influence of spectral variation on the predicted accuracy

4.3.5.1 Context window

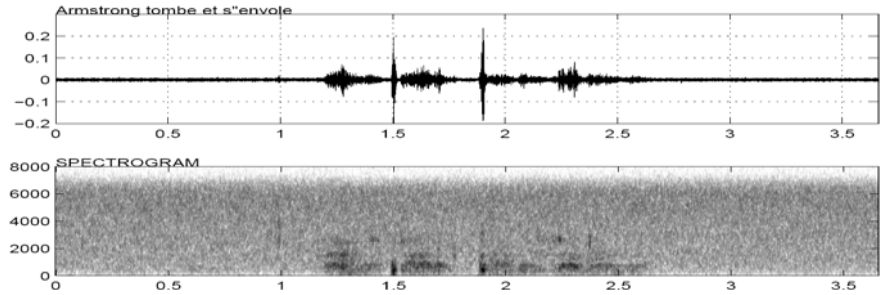
In this section, we compare two different data reduction methods: Linear Discriminant Analysis (LDA) versus Principal Component Analysis (PCA) to compute the input spectral features for whispered speech. Classes of F_0 ranges or phonemes were determined to label whispered frames for LDA training. Furthermore, we used different sizes of context window to study the influence of

spectral variation on the pitch estimation performance as well as the spectral estimation performance of GMM-based system. This was presented in (Tran *et al.*, 2008a).

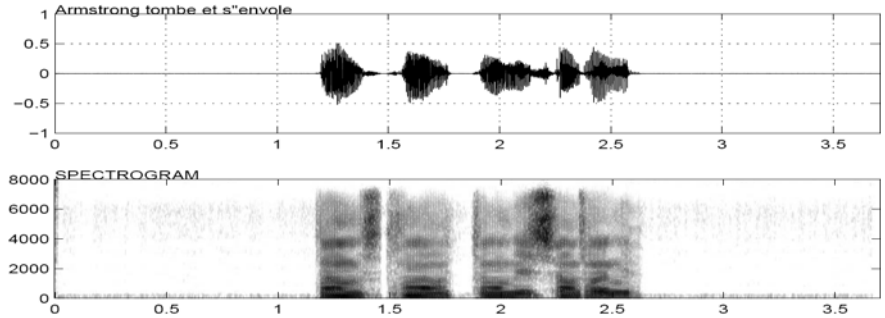
a. / Whisper captured by a headset microphone /



b. / Whisper captured by NAM microphone /



c. / Converted speech /



d. / Natural speech /

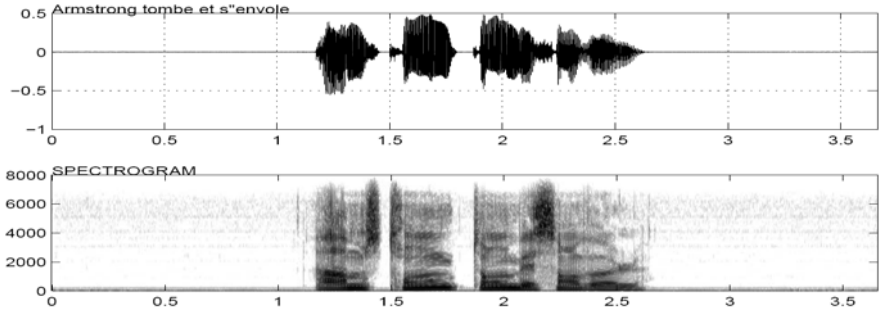


Figure 4.9. *Whispered speech captured by headset microphone (a), NAM sensor (b), converted speech (c) and ordinary speech (d) for the same utterance: "Armstrong tombe et s'envole".*

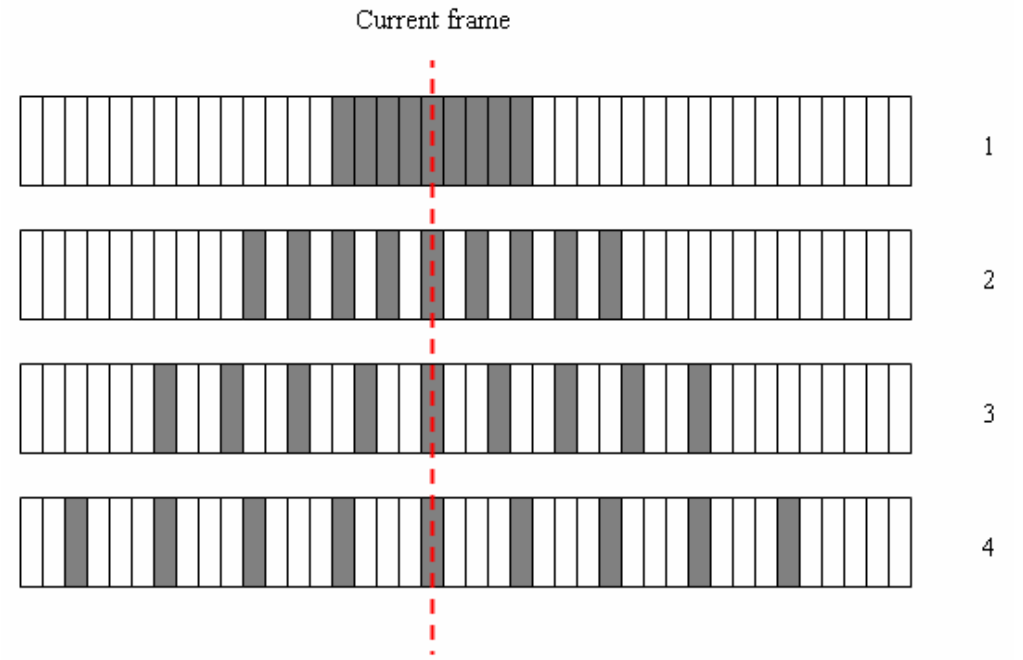


Figure 4.10. *Combining multiple frames for input features*

In (Toda, Black and Tokuda 2008), the authors showed that the use of feature vectors constructed by concatenating multiple acoustic frames, employed as an input feature to take into account the dynamic constraints on acoustic parameters, is effective for improving the mapping accuracy. In this article, the size of the context window is increased by concatenating more adjacent frames. In our case, this context window is widened by picking one frame every *m* frames (as presented in the figure 4.10 for the cases of *m* = 1 to 4) keeping thus the number of frames for the concatenation remains constant. PCA or LDA is then applied to this multi-frames vector for dimensionally reduction.

4.3.5.2 F_0 estimation

For this evaluation, the dimension of the whispered characteristics is reduced by using a PCA as proposed in the original system or by using a LDA. For the LDA, the target speech frames are classified into 13 classes: unvoiced frames are labelled with a '0' label and voiced frames fall into 12 other labels, depending on which interval the F_0 value in this frame belongs to (bark scale between 70Hz and 300 Hz). The class of a whispered frame is deduced from the class of the corresponding speech frame using the warping path generated by the transcription boundaries for each utterance pair. This information is then used to guide the LDA in the dimension reduction of the whispered sequence features. While the PCA reduces the dimension of the whispered sequence without regarding the corresponding

speech features, we hope that the seek for traces of the F_0 control performed by LDA will improve the performance of the system.

For the voiced/unvoiced estimation, we used the same neural network as the one proposed in 4.3.3 to ensure that the results were not influenced by this step.

The number of Gaussian mixtures for F_0 estimation varied from 8 to 64 (8, 16, 32, 64). The size of the context window was also varied from the phoneme size (90 ms) to the syllable (or foot) size (350 ms) (by picking one frame every 1-5 frames).

Table 4.3. *Mean and standard deviation of F_0 difference (%) between converted and target speech on the French data.*

method	window size		Number of Gaussian mixtures			
	frame interval	Context size (ms)	8	16	32	64
PCA	1	105	10.96 (± 2.04)	**10.90** (± 2.02)	10.92 (± 1.99)	10.90 (± 1.94)
	2	185	10.77 (± 2.02)	10.41 (± 2.03)	10.29 (± 2.13)	10.44 (± 2.2)
	3	265	10.33 (± 2.03)	9.98 (± 2.19)	10.08 (± 2.17)	10.28 (± 2.09)
	4	345	9.90 (± 2.29)	9.58 (± 2.24)	9.47 (± 1.98)	9.82 (± 2.03)
	5	425	9.44 (± 2.14)	**9.17** (± 2.16)	9.32 (± 2.21)	9.31 (± 2.14)
LDA	1	105	10.85 (± 2.02)	10.58 (± 1.99)	10.56 (± 2.08)	10.64 (± 1.82)
	2	185	10.36 (± 2.16)	10.23 (± 2.07)	10.11 (± 2.09)	10.36 (± 1.88)
	3	265	9.98 (± 2.3)	9.94 (± 2.23)	9.93 (± 2.16)	10.29 (± 2.2)
	4	345	9.45 (± 2.06)	9.43 (± 2.19)	9.62 (± 2.2)	9.67 (± 2.21)
	5	425	**9.15** (± 2.08)	9.22 (± 1.99)	9.25 (± 2.19)	9.37 (± 2.18)

Table 4.3 shows that using LDA and a large window size improves the precision of pitch estimation with respect to PCA with a small window. When using LDA with a window about 400 ms, the F_0 error decreases by 16% compared to our previous system with PCA and a smaller window size (10.90% with 16 mixtures → 9.15% with 8 mixtures). However, using LDA instead of PCA with the same context window size does not significantly improve the precision (9.17% with 16 mixtures

→ 9.15% with 8 mixtures). But with a reduced number of mixtures (8), LDA outperforms PCA. Furthermore, The ANOVA on three factors (number of Gaussians, context size and dimensional reduction method) also showed that using a larger context window (including at least one syllable) is more significant than using a smaller window (including only a phoneme) for the F_0 estimation ($F(4,70)=52.81$, $p<0.001$). The LDA is also statistically outperforms PCA ($F(1,70)=6.25$, $p=0.013$) while the number of Gaussians does not have significant effect ($F(3,70)=1.7$, $p = 0.17$).

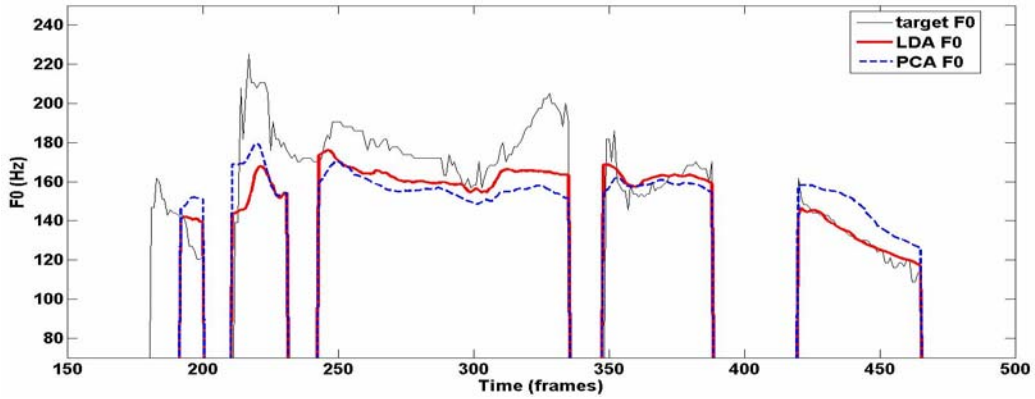


Figure 4.11. *Comparing natural and synthetic F0 curves (French data).*

Figure 4.11 shows an example of a natural (target) F_0 curve and the synthetic F_0 curves generated by the two systems (LDA + large context window *vs.* PCA + small context window). The predicted F_0 contour is closer to the natural F_0 curve than the one generated by the reference system and also smoother. This suggests that increasing the context window size probably helps generating smooth F_0 contours. Largest effect takes place at the end of the sentences.

4.3.5.3 Spectral estimation

We also evaluated the influence of LDA and long-term spectral variation to the spectral estimation. The phonetic segmentation was used to guide the LDA: each whispered frame was classified into one of 34 classes, depending on which phonetic segment it belonged to.

Table 4.4 provides the cepstral distortion between the converted and the target speech (the higher the distortion, the worse the performance). It shows that LDA is better than PCA ($F(1,70)=22.59$, $p<0.001$). But contrary to F_0 estimation, the spectral distortion increases when the size of the time window increases ($F(3,70)=18.55$, $p<0.001$). The most plausible interpretation is that a phoneme-sized window optimally contains necessary contextual cues for spectral conversion.

Table 4.4. *Mean cepstral distortion (dB) and standard deviation between converted and target speech (French data).*

Method	window size		Number of Gaussian mixtures	
	frame interval	**Context size** (ms)	8	16
PCA	1	105	7.23 (± 0.69)	**6.96** (± 0.56)
	2	185	7.20 (± 0.59)	7.01 (± 0.57)
	3	265	7.42 (± 0.77)	7.26 (± 0.66)
	4	345	7.25 (± 0.71)	7.55 (± 0.66)
LDA	1	105	6.96 (± 0.6)	**6.83** (± 0.57)
	2	185	6.98 (± 0.57)	7.01 (± 0.59)
	3	265	7.03 (± 0.61)	7.17 (± 0.59)
	4	345	7.19 (± 0.65)	7.34 (± 0.54)

4.3.6 Influence of visual information for the conversion

To convey a message, humans produce sounds by controlling the configuration of oral cavities. The speech articulators determine the resonance characteristics of the vocal tract during speech production. Movements of visible articulators such the jaw and lips are known to significantly contribute to the intelligibility of speech during face-to-face communication (Summerfield 1979; Summerfield, MacLeod *et al.* 1989). In the field of person-machine communication, visual information can be helpful both as input and output modalities (Bailly, Bérar *et al.* 2003; Potamianos, Neti *et al.* 2003).

To evaluate the contribution of the facial movements to the performance of the GMM-based system, the Japanese audiovisual corpus presented in chapter 3 was used.

4.3.6.1 Visual parameters extraction

The facial movement parameters were obtained using the face cloning methodology developed at DPC (ex. ICP) of GIPSA-Lab (Revéret *et al.*, 2000), (Elisei *et al.*, 2001), (Badin *et al.*, 2002), (Bailly *et al.*, 2003). In our case, we identified the contribution of the jaw rotation, lip rounding vertical movement of upper and lower lips, lip corner movements and throat movement.

The face and profile views of the subject have been video-monitored under good lighting conditions. From the raw motion data of the 142 colored beads glued on

the speaker's face (particularly those on the lips and the jaw since their movements are supposedly dominating for the speech), a linear model composed of 5 degrees of freedom was computed. A so-called guided Principal Component Analysis (PCA) was used, which iteratively estimates and subtracts the elementary movements of segments (lips, jaw, etc.) known to drive facial motion. The resulting articulatory model also includes other components for head movements and facial expressions but only components related to speech articulation were considered in this thesis, namely:

1. Jaw Raising/lowering (Jaw1 parameter). The PCA was applied to the vertical coordinates (*y)* of the beads on the jaw and the lower teeth;
2. Lip Protrusion (Lip1 parameter). The PCA was applied to the coordinates (residue) *xyz* of the beads on the lips;
3. Lower lip raising/lowering (Lip2 parameter). The PCA was applied to the vertical coordinates (residue) (*y)* of the beads on the lower lip;
4. Upper lip raising/lowering (Lip3 parameter). The PCA was applied to the vertical coordinates (residue) (*y)* of the beads on the upper lip;
5. Larynx raising/lowering (Lar1 parameter). The PCA was applied to the vertical coordinates (residue) (*y)* of all the beads on the neck.

These five visual parameters are used as input and output modalities to obtain a multimodal whisper-to-speech system.

Figure 4.12. *Video-realistic rendering of computed movements by statistical shape and appearance models driven by the same articulatory parameters*

4.3.6.2 Facial animation

The audiovisual rendering of the estimated visual parameters is performed by texturing the mesh piloted by the shape model introduced above. An appearance model that computes a texture at each time frame is built using a technique similar to AAM (Active Appearance Models) (Cootes, Edwards *et al.*, 2001) in three steps: (1) shape-free (SF) images are obtained by morphing key images to a reference mesh; (2) contrary to AAM, these pooled SF images are then directly linked to articulatory parameters via a multilinear regression; (3) the resulting appearance model is then used to compute a SF synthetic image given the set of articulatory parameters of each frame and used to texture the corresponding shape. The main difference with AAM is the direct multilinear regression used instead of a joint PCA and the number of configurations used: while AAM typically uses a few dozen images on which a generic mesh is adapted by hand or semi-automatically, we use here more than a thousand configurations on which the mesh is positioned automatically thanks to marked fleshpoints (Bailly, Bégault *et al.* 2008). The videorealistic audiovisual rendering of computed facial movements is illustrated in Figure 4.12. This figure presents from left to right: original frame; its shape-free transform (note the texture distortion in the mouth region because the reference mesh is chosen with an open mouth) and three different synthesized frames sampling a bilabial closure (note the nice rendering of the inner mouth despite the linear modelling of the nonlinear appearance/disappearance of the inner parts, notably the teeth).

4.3.6.3 Contribution of the visual parameters to the conversion system

For this evaluation, the database consists of 150 sentences for training and 40 sentences for testing, pronounced by a native Japanese speaker. The audiovisual feature vector combines whispered spectral and visual feature vectors in an identical way to the AAM where two distinct dimensionality reductions by PCA are performed in shape and appearance and further combined into a third PCA to get a joint shape and appearance model. Acoustic and visual features are fused, using an identical process. The 0^{th} through 19^{th} mel-cepstral coefficients are used as spectral features at each frame. The input feature vector for computing the speech spectrum is constructed by concatenating feature vectors at current ±8 frames and further reduced to a 40-dimension vector by a PCA. Each visual frame is interpolated at 200 Hz – so as to be synchronous with the audio signal – and characterized by a feature vector obtained by concatenating and projecting ±8 frames centred

around the current frame on the first *n* principal components. The dimension of the visual vector *n* is set to 10, 20 or 40. Prior to joint PCA, the visual vector is weighted. The weight *w* varied from 0.25 to 2. The conversion system uses the first 40 principal components of this joint audiovisual vector. In the evaluation, the number of Gaussian is fixed at 16 for the spectral estimation, 16 for the F_0 estimation and 16 for the aperiodic components estimation.

Table 4.5. *Influence of visual information on voicing decision (Japanese data)*

Type of error	**Feed-fwd NN (%)**					**GMM (%)**
	AU	**AUVI**				
		w = 1	w = 0.75	w = 0.5	w = 0.25	
Voiced error	3.71	3.58	3.74	3.34	**2.75**	4.29
Unvoiced error	4.34	**4.19**	4.71	4.37	4.87	5.47
Total	8.05 (± 3.99)	7.77 (± 3.45)	8.45 (± 3.46)	7.71 (± 3.48)	**7.62** (± 3.86)	9.76 (± 4.43)

4.3.6.3.1 Voicing decision

In the same way as for the evaluation of voicing decision described in section 4.2, the audiovisual vectors of whisper in the training corpus of the conversion system are used to train the network. This network has 40 input neurons, 17 hidden neurons and 1 output neuron. Table 4.5 shows that visual information improves the accuracy of voicing decision. With a visual weight empirically set at 0.25, the voicing error is decreased by 5.4 % (8.05 % → 7.62%) compared to audio-only 21.9 % compared with the baseline system (9.76 % → 7.62%) ($F(1,40) = 2.8$, $p = 0.065$).

4.3.6.3.2 Spectral and excitation estimations

The visual information also enhances the performance of the conversion system. As shown in Table 4.6, the best results are obtained with weighting equally acoustic and visual parameters (w = 1) and with the dimension of the visual vector equal to 40. The spectral distortion between the converted speech and the target speech is decreased by 3.7% (5.99 dB → 5.77 dB) while the difference between the converted speech and that of target speech is decreased by 6.9 % (11.56 % → 10.76 %) for F_0 estimation and 3.6 % for aperiodic components estimation.

Figure 4.13 shows an example of F_0 curves converted from audio and audiovisual input whisper. With visual information as an additional input, we have a better converted F_0.

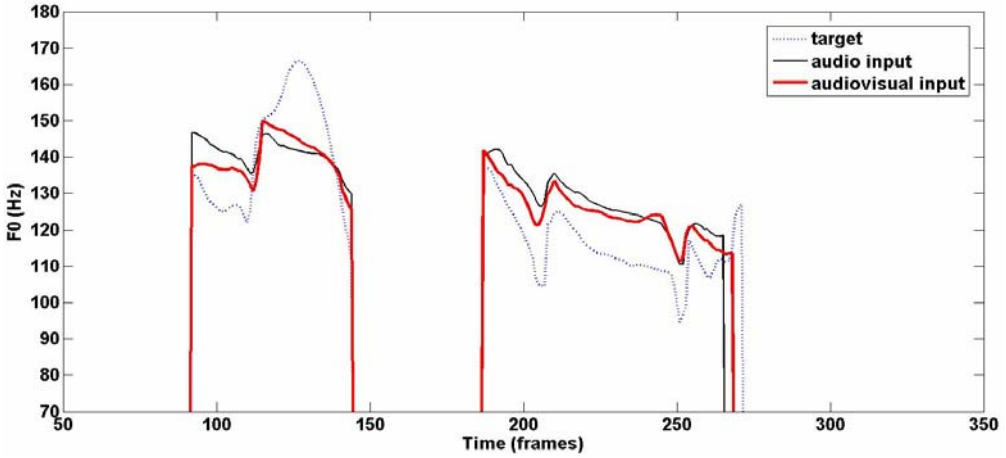


Figure 4.13. *Natural and synthetic F0 curve converted from audio and audiovisual whisper.*

Table 4.6 *Influence of visual information on the estimation of spectral and excitation features*

		Visual weight									
Distorsions	Visual dimension	**Audio-only**	0.25	0.5	0.75	**1**	1.25	1.5	1.75	2	**Video-only**
Cepstral distortion (dB)	10	**5.99**	7.01	5.97	5.80	5.78	5.88	5.87	5.86	5.84	**9.28**
	20		7.01	5.97	5.79	5.77	5.86	5.83	5.95	5.89	
	40		7.02	5.96	5.79	**5.77**	5.86	5.82	5.94	5.88	
F_0 estimation (%).	10	**11.56**	14.55	12.44	11.42	10.99	11.71	12.75	14.24	15.77	**12.95**
	20		14.79	12.38	11.40	10.78	11.66	12.62	14.31	15.43	
	40		14.57	12.39	11.04	**10.76**	11.66	12.62	14.33	15.46	
AP distortion (dB)	10	**38.72**	41.46	38.26	37.73	37.57	37.33	37.95	37.57	37.95	**51.45**
	20		41.42	38.25	37.79	37.55	37.33	38	37.90	37.95	
	40		41.42	38.21	37.78	37.53	**37.33**	37.99	37.53	37.93	

4.4 Perceptual evaluations

In this section, the objective improvement of our system over the original one proposed by Toda *et al.* is further evaluated by subjective tests. We recall that two corpora were recorded. The French corpus is used for the acoustic evaluation. The Japanese corpus is used for the audiovisual evaluation.

4.4.1 Acoustic evaluation

For this evaluation, sixteen French listeners, who had never listened to NAM, participated in our perceptual tests on the comparison of the intelligibility and naturalness of the converted speech in the two systems (Tran *et al.*, 2008c). We used 20 test utterances not included in the training set.

Each listener heard an utterance pronounced in normal (original voiced) speech and the converted utterances obtained from the whispered speech with both systems. For intelligibility testing, the spectral, F_0 and aperiodic component parameters were obtained by the two voice conversion procedures and the synthesis was performed using STRAIGHT (Kawahara, Masuda-Katsuse *et al.*, 1999). For naturalness, the spectral and aperiodic component parameters were original and the stimuli were obtained by substituting only predicted voicing and F_0 values to the original warped target frames.

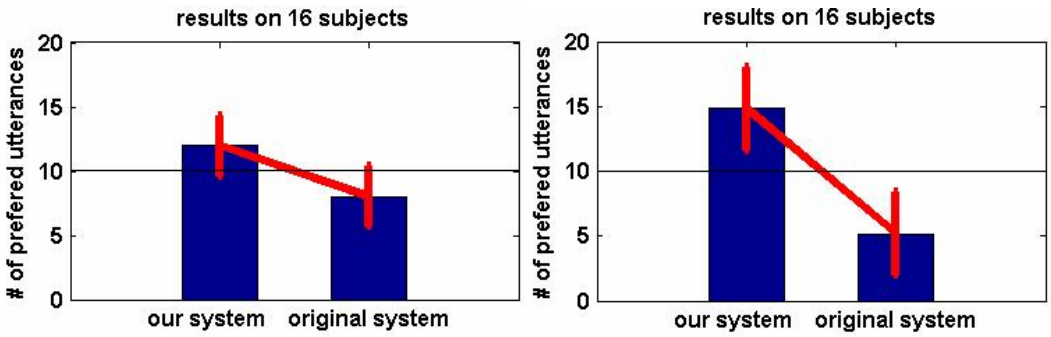


Figure 4.14. *ABX results. (left) intelligibility ratings (right) naturalness ratings*

This procedure was chosen because the quality of the converted spectrum could also influence the perception in the naturalness test. We want to make sure that we were testing the prediction of F_0 and voicing only.

For each utterance, listeners were asked which one was closer to the original one, in terms of intelligibility and in terms of naturalness. An ABX test (or matching-to-sample test) was used. It is a discrimination procedure involving presentation of two test items and a target item. The listeners are asked to tell which test item (A or

B) is closest to the target (X). Figure 4.14 (left) displays the mean intelligibility scores for all the listeners for the converted sentences using the original system and our new system. An ANOVA showed that the intelligibility score is higher for the sentences obtained with our new system ($F = 23.41$, $p < .001$). Figure 4.14 (right) shows the mean naturalness. Again the proposed system is strongly preferred to the original one, as shown by an ANOVA ($F = 74.89$, $p < .001$).

4.4.2 Audiovisual evaluation

To confirm the positive contribution of visual information, eight Japanese listeners participated in our perceptual tests on audiovisual converted speech (see also Tran *et al.*, in press). The stimuli consisted of Japanese VCV (with V chosen amongst five vowels and C amongst twenty-seven consonants) sequences with four conditions:

1. Speech generated from whispered audio (named 'condition A from A')
2. Speech generated from whispered audio and video ('condition A from AV')
3. Speech and facial animation generated from whispered audio ('condition AV from A')
4. Speech and facial animation generated from whispered audio and video ('condition AV from AV')

The five vowels were /a/, /i/, /e/, /o/, /u/. The 27 consonants were the following: /p/, /pj/, /b/, /bj/, /m/, /mj/, /d/, /t/, /s/, /ts/, /z/, /j/, /n/, /nj/, /k/, /kj/, /g/, /gj/, /f/, /Σ/, /tΣ/, /Z/, /h/, /hj/, /r/, /rj/, /w/. Only 115 combinations of vowels and consonants were tested.

Each participant heard and viewed a list of randomized synthetic audio and audiovisual VCV. For each VCV, she/he was asked to select what consonant she/he heard among a list of 27 possible Japanese consonants.

Figure 4.15 provides the mean recognition scores for all the participants. Visual parameters significantly improve consonant recognition: the identification ratios for 'AV from A' (28.37 %), 'A from AV' (30.56 %) and 'AV from AV' (36.21 %) are all significantly higher than that of the 'A from A' condition (23.84 %) ($F = 1.23$, $p < 0.303$). The figure also shows that providing visual information to the speakers is more beneficial when it is synthesized from audiovisual data ('AV from AV') than when it is derived from audio data alone ('AV from A'). Furthermore the addition of visual information in the input data ('A from AV') increases

identification scores compared with audio data alone ('A from A'), but is not significantly different from providing audiovisual information derived from the audio alone ('AV from A').

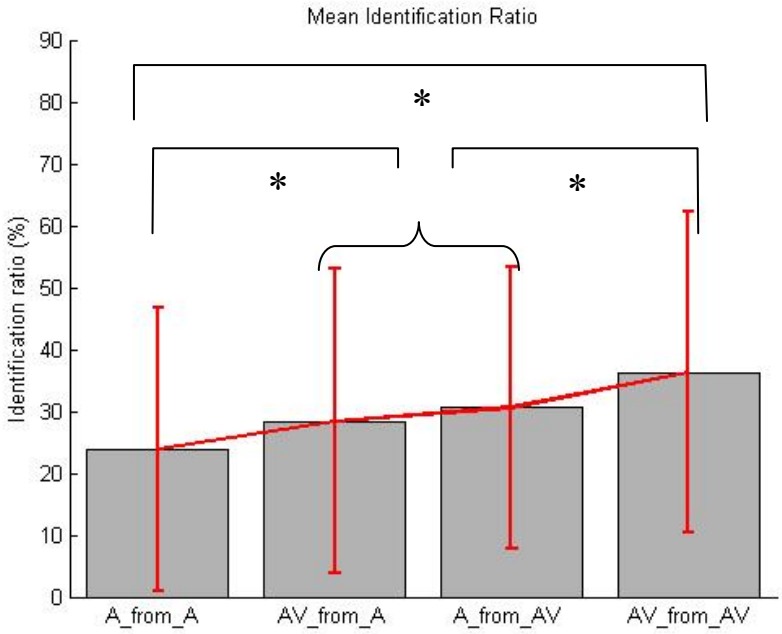
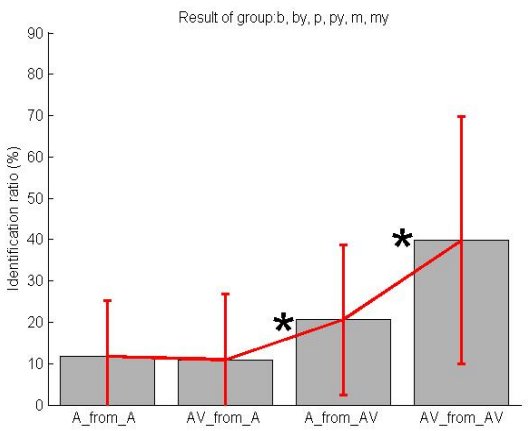
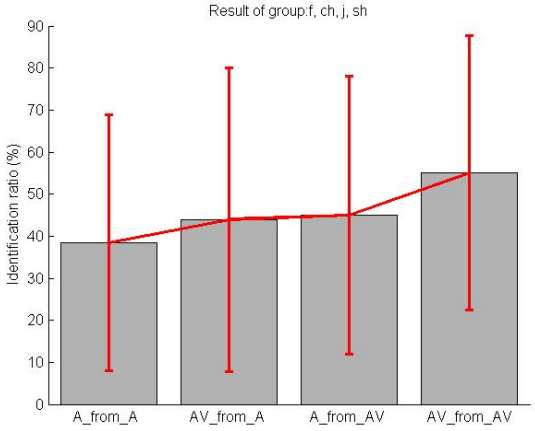


Figure 4.15. *Mean of consonant identification ratio.*

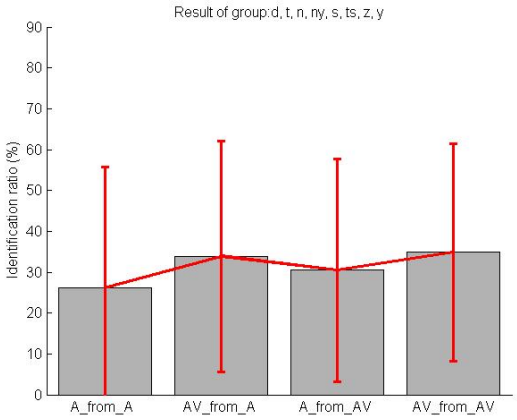
Although adding visual information greatly increases the identification scores on all the tested VCVs, the scores are still quite low (less than 40%). It should be noted however that nonsense VCV recognition is a difficult task, especially in a language with a lot of consonants. It could therefore be argued that identification scores on words or sentences could be much higher, with the help of lexical, syntactic and contextual information. With this in mind, it could be interesting to check whether the addition of visual information provided in fact some cues to place of articulation, even if accurate phoneme detection was too difficult. Therefore we also grouped the consonants into different place of articulation categories to examine which consonantal groups benefited most from the addition of visual information. The 27 consonants were thus grouped into bilabials (/p/, /pj/, /b/, /bj/, /m/, /mj/), alveolars (/d/, /t/, /s/, /ts/, /z/, /j/, /n/, /nj/), palatals (/k/, /kj/, /g/, /gj/), non-alveolar fricatives (/f/, /Σ/, /tΣ/, /Z/) and others (/h/, /hj/, /r/, /rj/, /w/). The first aim was to check whether consonants belonging to the bilabial category,



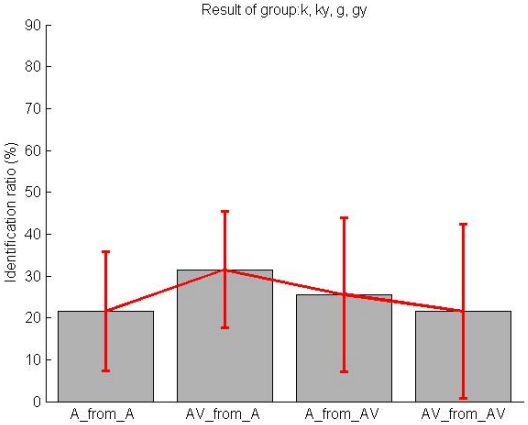
(a) bilabials (F = 2.62 , p < 0.079)



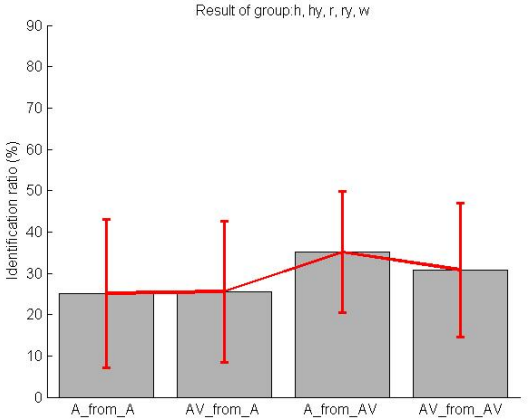
(b) non-alveolar fricatives (F = 0.17, p < 0.912)



(c) alveolars (F = 0.16, p < 0.922)



(d) palatals (F = 0.3 , p < 0.822)



(e) others (F = 0.42, p < 0.744

Figure 4.16. *Recognition ratios for different groups of consonants, classified according to their places of articulation*

which are intrinsically more salient visually, were better identified as bilabials, when visual information was taken into account in the speech conversion procedure and when audiovisual stimuli were presented. The second aim was to examine what kind of perceptual confusions were observed.

Table 4.7. *Confusion matrices for the 5 places of articulation in the 4 conditions*

		bilabial	alveolar	palatal	fricative	other
A from A	bilabial	18.29	10.86	22.86	34.86	13.14
	alveolar	2.38	59.52	11.90	8.73	17.46
	palatal	3.86	15.44	54.83	12.74	13.13
	fricative	0.95	25.71	21.90	47.62	3.81
	other	5.00	20.71	25.71	9.29	39.29
AV from A	bilabial	15.00	10.50	28.00	36.50	10.00
	alveolar	3.47	65.97	9.03	9.72	11.81
	palatal	2.36	16.89	65.20	9.12	6.42
	fricative	0.83	21.67	27.50	49.17	0.83
	other	7.50	19.38	27.50	10.63	35.00
A from AV	bilabial	32.57	5.14	20.57	30.29	11.43
	alveolar	0.79	63.49	15.87	6.35	13.49
	palatal	6.56	13.51	57.14	12.74	10.04
	fricative	3.81	21.90	24.76	47.62	1.90
	other	5.00	20.00	21.43	6.43	47.14
AV from AV	bilabial	75.00	1.00	9.50	10.50	4.00
	alveolar	0.69	70.83	11.81	2.78	13.89
	palatal	1.01	17.91	62.84	10.14	8.11
	fricative	0.83	23.33	26.67	47.50	1.67
	other	4.38	16.88	25.00	6.88	46.88

Figure 4.16 displays the evolution of the identification scores along the four conditions (‘A from A’, ‘AV from A’, ‘A from AV’, ‘AV from AV’) for the five articulatory groups. The average of the correct identification rates for each consonant in an articulatory group was calculated to provide an overall correct identification rate for each group. This articulatory grouping shows that visual information effectively helps the participants to identify bilabials. The identification ratio for the bilabial consonants rises from 11.67 % in the ‘A from A’

condition to 39.83 % in the ‘AV from AV’ condition ($F = 2.62$, $p < 0.079$). The identification ratios for all the other groups are quite stable and are not significantly increased by the addition of visual information.

Table 4.7 provides the confusion matrices with places of articulation chosen among 5 categories (bilabial, alveolar, palatal, fricative, or other), rather than among 27 individual consonants. The focus here is on the broad category the consonant belongs to, not on the score of individual consonants. Interestingly, the bilabial place of articulation, which is poorly identified in the ‘A from A’ and ‘AV from A’ conditions (less than 20 %), becomes quite well identified in the ‘AV from AV’ condition (75 %). When the bilabial place of articulation is wrongly identified, it is most often mistaken for what we named “non-alveolar fricatives”, a group which contains labiodentals or rounded consonants. This suggests that the labial place of articulation remains quite perceptible. Table 4.6 also shows that the alveolar place of articulation is quite well recovered from the audio alone (approximately 60 %). Although its recovery does benefit from the addition of visual information, the ‘AV from AV’ score does not reach as high a score as the one for the bilabial place of articulation (70.83 % vs. 75 %). We must recall here that the visual synthesis used in this study does not provide information on tongue movements, which could in fact be visible for alveolar consonants. This probably explains why the addition of visual information does not provide such a drastic improvement on the scores. The palatal place of articulation reaches a good score (above 50 %) but does not benefit much from visual information, as could be expected. The other two groups do not benefit from visual information and are not well identified (less than 50 %).

4.5 Summary

This chapter proposes several solutions to improve the intelligibility and the naturalness of the speech generated by a whisper-to-speech conversion system. The original system proposed by Toda and Shikano (2005) is based on a GMM predicting the three parametric streams of the STRAIGHT speech vocoder (F_0, harmonic and noise spectrum) from spectral characterization of the non-audible-murmur input.

The first improvement concerns characteristics of the voiced source. We have shown that the estimation of voicing and F_0 by separate predictors improves both predictions. We have also shown that F_0 prediction is improved by the use of a large input context window (about 425 ms) to get more spectral variation compared

to the original smaller window (about 105 ms). Predictions of all parametric streams are further improved by a data reduction technique that makes use of the phonetic structure of the speech stream (LDA).

The second part of the chapter compared objective and subjective benefits offered by multimodal data. Improvements obtained by adding visual information in the input stream as well as including articulatory parameters in the output facial animation are very significant.

Although the performance of the system is improved, the estimated pitch as well as the spectral structures of converted speech is still flat due to the impoverished phonetic contrasts of the GMM-based method. Listeners therefore have sometimes difficulty in chunking the speech continuum into meaningful words due to incomplete phonetic cues provided by output signals. The next chapter consequently studies another approach consisting in combining HMM-based statistical speech recognition and synthesis techniques to convert silent speech to audible voice. By introducing phonological constraints, such systems are expected to improve the phonetic consistency of output signals.

Chapter 5

Multi-streams HMM-based whisper-to-speech system

5.1 Introduction

As presented in chapter 1, two main approaches have been proposed to generate audible – and visible – speech from signatures of non-audible articulation:

- Mapping techniques based on a GMM can be used to directly convert whispered signals into sound using aligned training corpora (of paired non-audible and audible sentences, words or other units) without the need for any phonetic information: joint multi-frame representations of non-audible and speech signals are either stored or modelled and then used to perform direct estimation – or inversion – of audible speech given the sole representation of non-audible signals. A conversion system based on this approach was presented in chapter 1 and improvements to this system were proposed in chapter 4.

- The second approach consists in plugging a speech synthesis system to a speech recognizer. The approach is quite straightforward: the recognizer segments the speech flow into phonemic units using both signal-dependent information and a language model. A standard speech synthesis system then converts the phonetic string into a synthetic voice either using the pre-recorded modal voice of the speaker or built-in available resources. The performance of such a system is mainly dependent on the recognition performance: correct recognition will result in a perfect reconstructed speech while recognition failures or inadequate language models result in drastic degradations.

In this chapter, we examine the feasibility of the second approach in a whisper-to-speech conversion system, including audiovisual data. A summary of this study was published in (Tran *et al.*, 2009). In this section, we only focus on the segmental intelligibility of the converted speech. The source excitation (F_0) generation is beyond the scope of this study. We first study the impact of visual information for

a recognition – synthesis system. This system is based on statistical Hidden Markov Models (HMM), which are traditionally used in speech recognition. Then, this system is compared with the original GMM-based system proposed by Toda and Shikano (2005).

The chapter is organized as follows. Section 5.2 provides an overview of the multi-stream HMM-based whisper-to-speech conversion system. The promising contribution of visual information to this system and the comparison between the two approaches (GMM-based mapping and HMM-based recognition and synthesis) are presented in the section 5.3. Finally, conclusions are drawn in section 5.4.

5.2 Multi-stream HMM-based Whisper-to-Speech system

In order to compare the performance of the GMM-based voice conversion technique proposed by Toda and Shikano (2005) with the approach of combining NAM recognition and speech synthesis, HMM-based whisper-to-speech conversion system was developed which combines 2 modules, namely HMM recognition and HMM synthesis.

Advantages of an HMM Speech synthesis method

Instead of using the diphone-based concatenative synthesis proposed in (Hueber *et al.*, 2008ab), we used an HMM-based synthesis, as described in (Tokuda *et al.*, 2000) because integrating a speaker specific parametric model offers some benefits. First of all, the whole synthesis system is parameterised rather than consisting of stored pre-recorded waveforms. The system is therefore *trainable*. Secondly, HMMs provide a much more compact representation compared to the concatenated approach. Moreover, the synthesized speech from an HMM-based system has an overall lower variance, more constrained articulatory quality than that obtained from a concatenative approach and the voice characteristics, speaking styles, emotions can be controlled parametrically. Finally, an HMM-based parametric model could bring a more intimate coupling between speech recognition and synthesis components, by tackling both problems in a unified coherent statistical framework (Tokuda *et al.,* 2004; Zen *et al.*, 2004, 2009; Zhang, 2009)

Training methods

As stated in Zhang *et al.* (2009), two training frameworks could be used to estimate the models.

The first uses separate training: while the whispered speech HMMs are trained on the whispered data only, and the phonated speech HMMs are built from the speech data alone using the standard HMM training procedure. The idea behind the separate training is clearly that training the two types of HMMs individually is likely to bring out the best performance from each channel. This framework does not need a prior alignment of whisper and speech data.

The second scheme, on the other hand, tries to find a way to jointly optimise a single model for both whispered speech and phonated speech information. The model therefore has two components, both are modelled as multi-state phone-level HMMs: a speech synthesis model which generates speech parameters given a phone sequence and a whisper model which derives the phone sequence for synthesis from the whispered features extracted from a whispered utterance. Both the whispered and speech model have the same topology: i.e. they have exactly the same set of HMM states and allophonic variations. This structure enables to bridge the whisper and speech domains. The Hidden Markov Model Toolkit[5] (HTK)'s multi-stream functionality makes this type of training possible by combining the two sets of HMMs into a single (for our system, we have a two-stream model when only acoustic information is available, and a four-stream model when facial data are added as a complementary information). This framework, on the other hand, needs a prior alignment of whisper and speech data represented the same phonetic information.

The overview of the HMM-based conversion system is presented in figure 5.1. The system consists of two processes: training and conversion. The training process is performed by the second scheme described above. The conversion process is carried out in 2 steps:

- The first step performed phoneme recognition, based on the acoustic HMMs. The result is a sequence of recognised allophones together with their durations.
- The second step of the conversion aims at reconstructing the speech parameters (cepstral coefficients in this work) from the chain of phoneme labels and boundaries delivered by the recognition procedure. As described in (Govokhina *et al.*, 2006), the synthesis is performed as follows, using the software developed by the HTS[6] group (Tamura *et al.*, 1999; Zen *et al.*, 2004). A linear sequence of HMM states is built by concatenating the corresponding segmental HMMs. The proper state durations are estimated

[5] http://htk.eng.cam.ac.uk/
[6] http://hts.sp.nitech.ac.jp/

by z-scoring. A sequence of observation parameters is generated using a specific ML-based parameter generation algorithm (Zen *et al.*, 2004).

Table 5.1. *Phone occurrences in the Japanese training corpus*

Phone	No. of occ	Phone	No. of occ	Phone	No. of occ	Phone	No. of occ
a	472	gj	6	n	171	s	83
a:	5	h	63	nj	8	Σ	68
b	61	hj	5	N	134	t	154
bj	1	i	370	o	347	ts	38
tΣ	37	i:	3	o:	60	u	290
d	78	Z	59	p	24	u:	72
e	255	k	179	pj	1	j	73
e:	17	kj	20	r	164	z	40
f	23	m	109	rj	7	w	47
g	101	mj	3	q	56	sil	362

Data

The data used to train and test this HMM-based conversion system are the Japanese corpus used for the evaluation of audiovisual GMM system presented in chapter 4 (c.f. 4.3.6) which includes 150 sentences for the training and 40 sentences for the test. These data are used in order to compare the spectral conversion performance of the two systems. Table 5.1 and 5.2 show the number of occurrences of each phone in the training and the test corpus respectively.

As can be seen in Table 5.1, some phones have a very low number of occurrences in the training corpus (e.g. */bj/* and */pj/* occur only once) but these phonemes have a very low frequency in the language (Amano & Kondo, 1999).

Table 5.2 shows that some phones are not represented in the test corpus. These phones correspond to phonemes that are rare in the language, however.

The 0^{th} through 19^{th} mel-cepstral coefficients extracted by STRAIGHT[7] and their first deltas were used as spectral features for the acoustic information while 5 visual parameters and their first deltas, extracted by the "talking head" cloning system developed at DPC (Bailly *et al.*, 2006), were used to characterize the movements of the jaw, throat and lips.

[7] http://www.wakayama-u.ac.jp/~kawahara/index-e.html

Table 5.2. *Phone occurrences in the Japanese test corpus*

Phone	No. of occ	Phone	No. of occ	Phone	No. of occ	Phone	No. of occ
a	141	gj	2	n	47	s	24
a:		h	18	nj		Σ	24
b	13	hj	2	N	35	t	50
bj		i	90	o	87	ts	4
tΣ	13	i:	1	o:	22	u	61
d	18	Z	12	p	6	u:	11
e	55	k	51	pj		j	13
e:	8	kj	4	r	38	z	10
f	6	m	20	rj	2	w	10
g	31	mj		q	22	sil	89

5.2.1 Concatenative feature fusion

To build the multi-stream HMM-based system by HTK Toolkit (Young *et al.*, 2006), a simple audio-visual fusion approach is concatenative feature fusion. Whispered and phonated speech audio and visual data need first to be aligned due to the different speaking rates and syllable durations in the two speaking modes (whispered and phonated). This alignment is carried out by the same semi-automatic procedure as presented in chapter 4.

Acoustic and visual parameters of whispered speech and phonated speech are then stored in 4 streams (whispered audio spectral stream, whispered visual stream, phonated audio spectral stream and phonated visual stream). Static and dynamic features are used for each stream to generate a realistic parameter trajectory in which the variations in parameters are much smoother. The joint bimodal feature vector is therefore presented as:

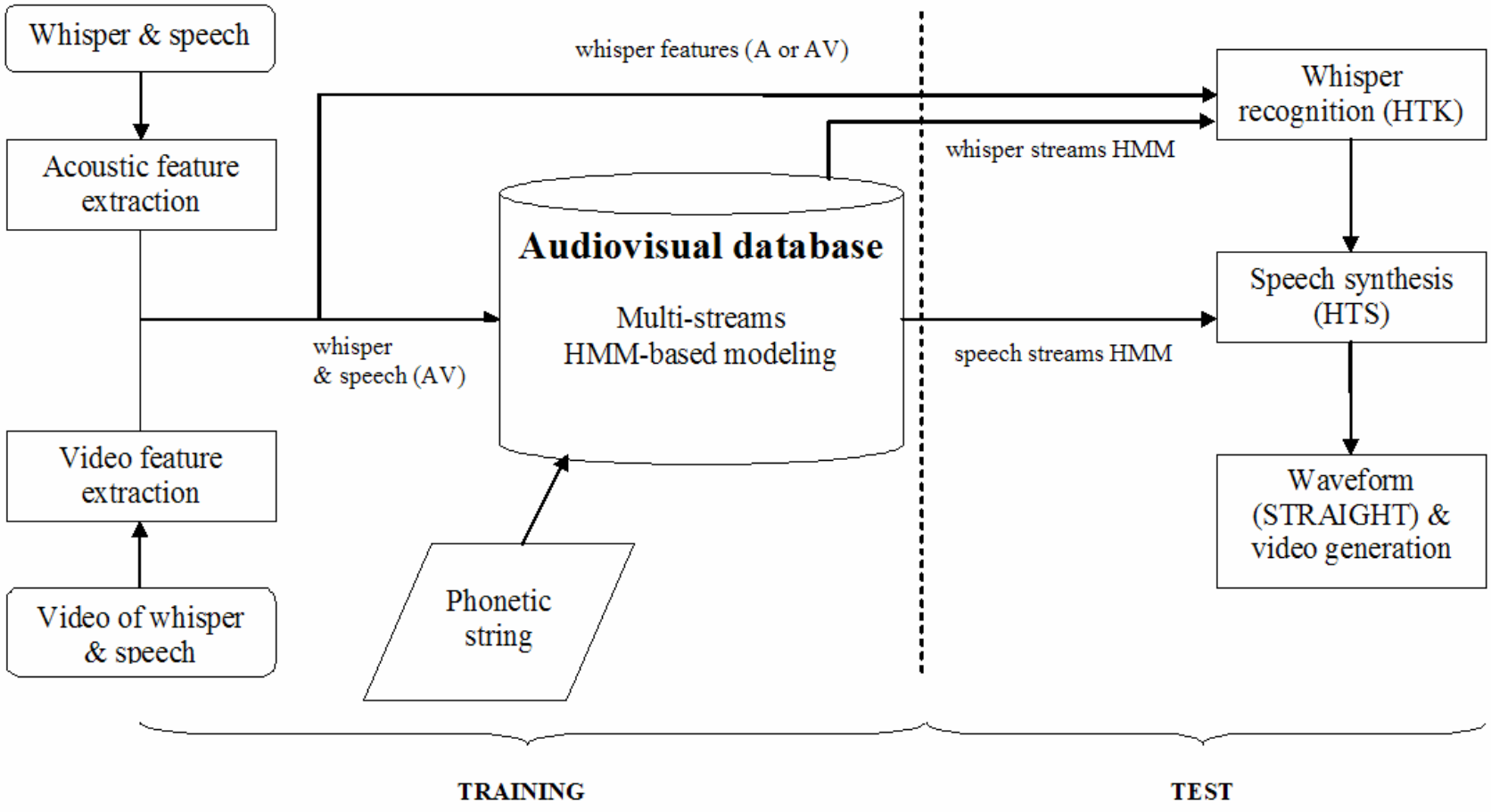


Figure 5.1. *HMM-based voice conversion system combining whisper recognition with speech synthesis*

$$o_t^{AV} = \left[o_t^{(A\_W)^T}, o_t^{(V\_W)^T}, o_t^{(A\_S)^T}, o_t^{(V\_S)^T} \right]^T \in R^{D^{AV}} \tag{5.1}$$

where o_t^{AV} is the joint audio-visual feature vector of aligned whispered speech and phonated speech, $o_t^{(A\_W)}$, $o_t^{(V\_W)}$ are the audio spectral feature and the facial feature of whisper, $o_t^{(A\_S)}$, $o_t^{(V\_S)}$ are the audio spectral feature and facial feature of phonated speech respectively and D^{AV} is the dimensionality of the *augmented* joint feature vector o_t^{AV}. Each state-output vector o_t of each stream, consists of the static feature c_t, and its first-order dynamic feature Δc_t:

$$o_t = \left[c_t^T, \Delta c_t^T \right]^T \tag{5.2}$$

where the dynamic feature is calculated as proposed in (Zen, Tokuda *et al.*, 2009):

$$\Delta c_t = c_t - c_{t-1} \tag{5.3}$$

The relationship between *augmented* output vector o_t and static vector c_t can be presented in matrix form as

$$\begin{bmatrix} \vdots \\ \boldsymbol{c}_{t-1} \\ \Delta \boldsymbol{c}_{t-1} \\ \boldsymbol{c}_t \\ \Delta \boldsymbol{c}_t \\ \boldsymbol{c}_{t+1} \\ \Delta \boldsymbol{c}_{t+1} \\ \vdots \end{bmatrix} = \begin{bmatrix} \cdots & \vdots & \vdots & \vdots & \vdots & \cdots \\ \cdots & \boldsymbol{0} & \boldsymbol{I} & \boldsymbol{0} & \boldsymbol{0} & \cdots \\ \cdots & -\boldsymbol{I} & \boldsymbol{I} & \boldsymbol{0} & \boldsymbol{0} & \cdots \\ \cdots & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I} & \boldsymbol{0} & \cdots \\ \cdots & \boldsymbol{0} & -\boldsymbol{I} & \boldsymbol{I} & \boldsymbol{0} & \cdots \\ \cdots & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I} & \cdots \\ \cdots & \boldsymbol{0} & \boldsymbol{0} & -\boldsymbol{I} & \boldsymbol{I} & \cdots \\ \cdots & \vdots & \vdots & \vdots & \vdots & \cdots \end{bmatrix} \begin{bmatrix} \vdots \\ \boldsymbol{c}_{t-2} \\ \boldsymbol{c}_{t-1} \\ \boldsymbol{c}_t \\ \boldsymbol{c}_{t+1} \\ \vdots \end{bmatrix} \tag{5.4}$$

where $c = \left[c_1^\mathsf{T}, ..., c_T^\mathsf{T} \right]^\mathsf{T}$ is a static feature-vector sequence and $\boldsymbol{W}$ is a matrix that appends dynamic features to c.

As stated above, we use 20 mel-cepstral coefficents and their first delta, together with 5 PCA-coefficients and their first delta to present static and dynamic acoustic parameters and visual parameters respectively in our experiments. The dimension D^{AV} of o_t^{AV} is therefore 100 (40+10+40+10).

5.2.2 Multi-stream HMM

The fusion data captures the reliability of each stream by combining the likelihoods of single-modality HMM classifiers. Such an approach has been used in multi-band audio-only ASR (Bourlard and Dupont, 1996) and in audio-visual speech recognition (Potamianos *et al.*, 2003). The emission likelihood of multi-stream HMM is the product of emission likelihoods of single-modality components weighted appropriately by stream weights. Given the *O* multi-streams observation vector, i.e., whispered and speech acoustic and facial modalities, the emission probability of multi-streams HMM is given by

$$b_j(O_t) = \prod_{s=1}^{N_s} \left[\sum_{m=1}^{M_s} c_{jsm} N(O_{st}; \mu_{jsm}, \Sigma_{jsm}) \right]^{\lambda_{sjt}} \tag{5.5}$$

where $N(O_{st}; \mu_{js}, \Sigma_{js})$ denotes the multivariate Gaussian distribution with mean vector μ_{js} and covariance matrix Σ_{js} for each state *j* in stream *s*. In contrast to the GMM-based system presented in chapter 4, the covariance matrix in each state of the HMM is modelled by a diagonal matrix due to the limitation of the training data. For each stream *s*, M_s Gaussians in a mixture are used, each weighted with c_{jsm}. The contribution of each stream is controlled by the weight λ_{sjt}. Although finding the optimized weights for the streams is interesting, it is a time consuming task for this system with a 4-stream topology. This optimization is therefore not studied in this thesis. The stream weights are set by default to 1.0 for all the streams in our system.

5.2.3 HMM Training

The joint probability densities of whispered speech and phonated speech parameters are modelled by "left-right" phone-sized HMM. Each HMM topology consists of five states. Three of these are emitting states and have output probability distributions associated with them. The transition matrix for this model will have five rows and five columns. The HMMs are trained by the following basic procedure (Young *et al.*, 2005).

Context-independent monophone training

Before starting the training process, the HMM parameters must be properly initialised with training data in order to allow a fast and precise convergence of the training algorithm. This initialization is considered as an isolated estimation:

- Extract all the corresponding segments (audio or audio+visual of aligned whisper-speech) for a particular phone model.
- The segments are then used ? an iterative Viterbi-training scheme until the parameters converge (*HInit* command in HTK).
- The Viterbi-estimated parameters are then refined by a further *Baum-Welch* training cycle (*HRest* command in HTK).

Whereas the isolated training above is sufficient for initialization using hand-labelled data, the HMM parameters are estimated by an "*embedded training*" procedure (*HERest* command in HTK).

For each utterance:

- Join together all of the HMMs corresponding to the phone list presented in the transcription to make a single composite HMM.
- Apply a *Forward-Backward* algorithm to collect the necessary statistics.

When all of the training utterances have been processed, the total set of accumulated statistics is used to update the parameters of all HMMs.

At the end of this stage, we have 40 HMMs for the phone presented in the training corpus (cf. table 5.1).

Context-dependent training

Starting from single Gaussian state output distribution, context-independent monophone, the model was enriched by two ways during the training:

- Due to coarticulatory effects, it is unlikely that a single context-independent HMM could optimally represent a given allophone. Therefore context-dependent HMM is used as another way to enrich the model. Because of limited training data (about 12 minutes), we only used bi-phone context for the acoustic models. In the recent study of Ben Youssef *et al.* (2009), the authors showed that the "next" context is better than the "precedent" context. The "next" context bi-phone is also chosen for our system by first grouping phonemes in context classes. In addition to using *a priori* phonetic knowledge to define such classes, confusion trees have been built based on the matrix of Manhattan distances of visual parameters between each pair of phone. Two different confusion trees were computed: one from the whispered visual (facial) parameters and one from the whispered acoustic features.

The confusion tree obtained from visual data (figure 5.2) provides groupings between consonants that can be interpreted articulatory. In particular, most of the bilabials are grouped together (*/p/, /b/, /m/, /mj/*). Another group gathers dentals *(/d/, /t/, /n/, /nj/*). Another group gathers dental fricatives and affricates (*/s/, /ts/ /z/, /Σ/, /tΣ/, /ts/*). Some results may seem difficult to interpret, such as the non-inclusion of */pj/* and */bj/* in the bilabial group. They can be explained by the fact that these consonants were very unfrequent in the corpus. Table 5.1 provides the number of occurrences of each phoneme in the corpus.

The confusion tree obtained from audio data (figure 5.3) provides groupings between consonants that can be interpreted acoustically, considering that in the whispered mode many phonemes are confused. Importantly, most of the stops are gathered in one group (*/t/, /k/, /b/, /ky/, /g/, /gy, /p/*). Voiced and unvoiced consonants are not separated, which is explained by the fact that they are undistinguishable acoustically in whisper.

Taking these two confusion trees as well as general phonetic knowledge, phonemes were classified coarsely into 3 groups for vowels ({/a/}, {/i/,/e/}, {/u/,/o/} without distinguishing between long and short vowels) and 7 groups for consonants: bilabials ({*/p/, /pj/*}, {*/b/, /bj/*}, {*/m/, /mj/*}), alveolars (*/d/, /t/, /n/, /nj/, /s/, /ts/, /z/, /j/*), palatals (*/Σ/, /tΣ/, /Z/*), velars ({*/k/, /kj/*},{*/g/, /gj/*}), */f/, /w/* and others ({*/h/, /hj/*}, {*/r/, /rj/*}). We added */f/* and */w/* to the list of contextual groups because they are visually distinguished from other consonants (see figure 5.2). Silences were also classified into 2 groups for utterance-final and internal silences. Only phone with high number of occurrences are extended to "next-context" biphone (>20). The number of HMM models in the training corpus is 149.

The biphones are initialized by making a copy of the corresponding monophones estimated in the previous step. These new biphone sets are then re-estimated twice by "embedded training" (*HERest*).

– The monogaussian state output distribution is then replaced by Gaussian mixture models. The number of Gaussians is increased from 1-2-3 and 4. Due to the limitation of training data, the number of Gaussians in each state of the HMM differs for each biphone, depending on its number of occurrences in the corpus. For biphones with a high number of occurrences in the corpus (>40), the number of Gaussians is set to 4. For biphones with lower number of occurrences, the number of Gaussians is set to 3, 2 and 1.

Finally, the number of parameters of HMM models is 114989 for audio system and 143545 for audiovisual system, calculated by the formula:

Number of parameters = models × states × Gaussians × (means + variances) + models × transitions.

5.2.4 Recognition

The recognition phase consists in identifying the most probable phoneme sequence from the sequence of observations (acoustic or audiovisual observations of whisper). This phase allows us to introduce additional linguistic constraints to enhance the recognition performance by using a language model. This model is especially essential for large vocabularies recognition task.

The most common kind of language model in use today is based on estimates of word string probabilities from large collections of text or transcribed speech. In order to make these estimates tractable, the probability of a word given the preceding sequence is approximated to the probability given the preceding one (bigram) or two (trigram) words (in general, these are called n-gram models). In our experiments, a bigram language model was used.

$$P(w_n \mid w_1, w_2, ..., w_{n-1}) = P(w_n \mid w_{n-1}) \tag{5.6}$$

This model was simply built from the labelled files in the training corpus via the following steps.

HLStats command reads all of the context-dependent transcriptions of training corpus, builds a table of bigrams in memory and then outputs a *back-off* bigram. The probability values of this table is calculated by the following formula

$$p(i,j) = \begin{cases} (N(i,j) - D) / N(i) & \text{if } N(i,j) > t \\ b(i)\,p(j) & \text{otherwise} \end{cases} \tag{5.7}$$

where *N(i,j)* is the number of times word *j* follows word *i* and *N(i)* is the number of times that word *i* appears. D is a constant for the discounting process in which a small part of the available probability mass is deducted from the higher bigram counts and distributed amongst the infrequent bigrams. When a bigram count falls below the threshold *t*, the bigram is backed-off to the unigram probability suitably scaled by a back-off weight in order to ensure that all bigram probabilities for a given history sum to one.

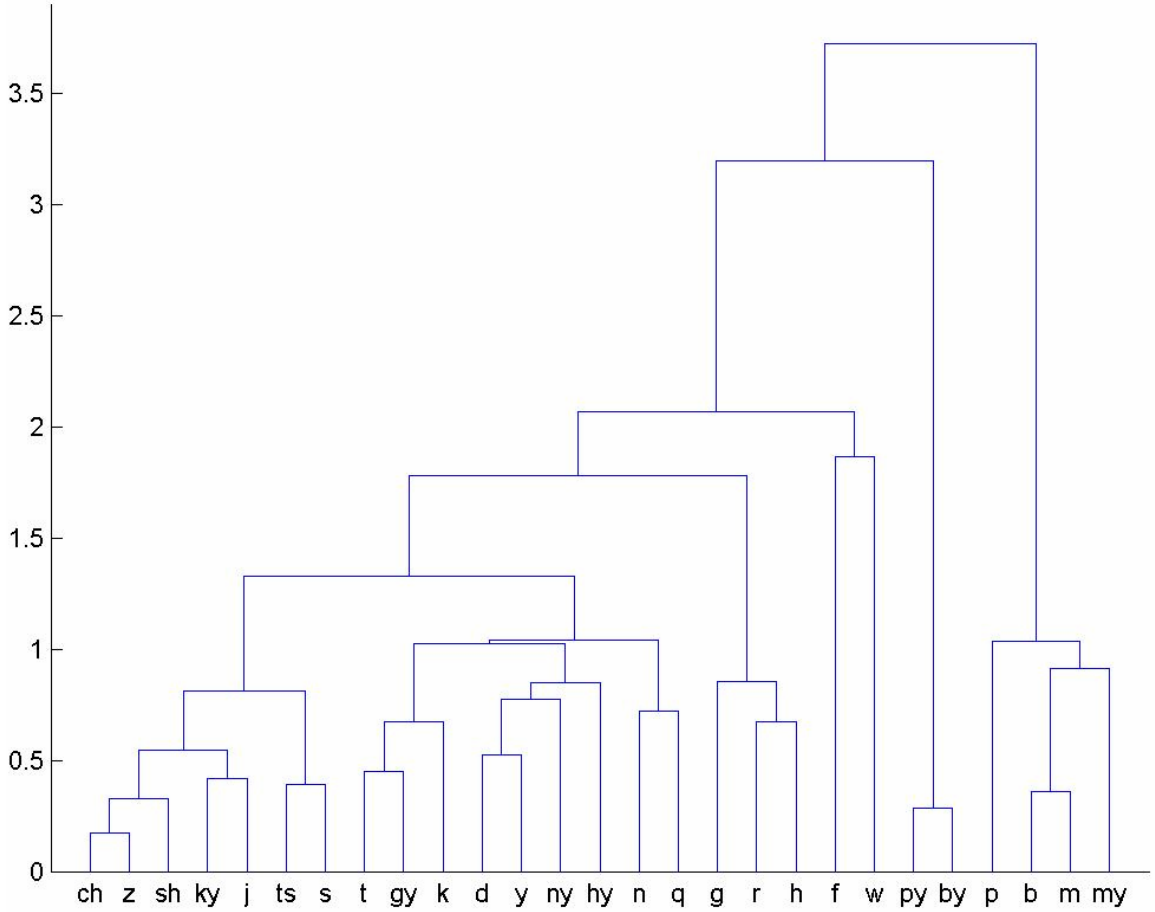


Figure 5.2. *Confusion tree of whispered visual movements of consonants (the smaller the ordinate, the more confused the two categories are)*

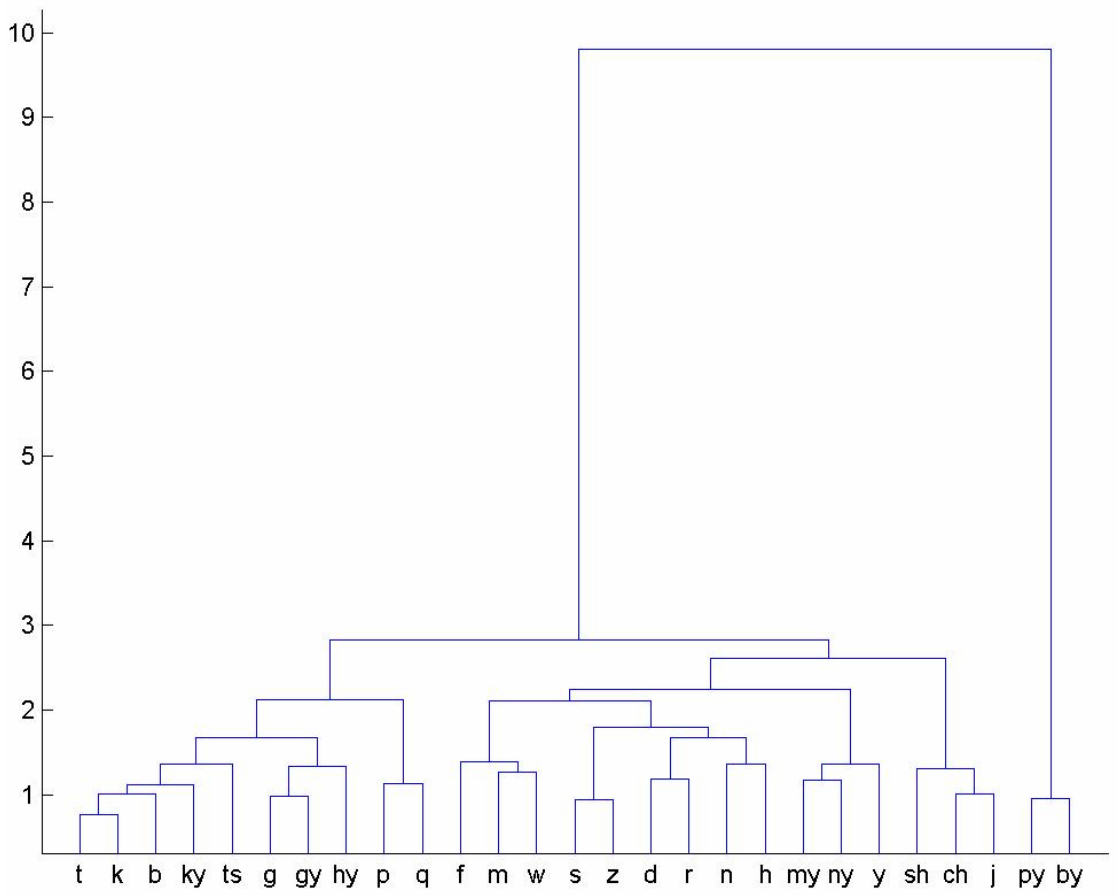


Figure 5.3. *Confusion tree of whispered acoustic parameters of consonants (the smaller the ordinate, the more confused the two categories are)*

The next step is to use *HBuild* command to construct the recognition network. One of the commonest forms of this network is the word-loop where all vocabulary items are placed in parallel with a loop-back to allow any word sequence to be recognized. In our experiments, *HBuild* reads the back-off bigram generated by *HLStats* and attaches a bigram probability to each word transition.

Finally, for an unknown input utterance with T frames, every path from the start node to the exit node of the network passes through T emitting HMM states. Each of these paths has a log probability which is computed by summing the log probability of each transition in the path and the log probability of each emitting state generating the corresponding observation. Within-HMM transitions are obtained ? from the HMM parameters while word-end transitions are determined by the language model likelihoods attached to the recognition networks. The "*Token Passing*" algorithm is used to find those paths which have the highest log probability (Young *et al.*, 2005).

5.3 Experiments and results

This section presents the objective performance of our multi-streams HMM-based whispered-to-speech by different criteria: recognition rate and cepstral distance between the converted speech and target audible speech. The impact of facial information for recognition and synthesis, as well as the comparison of this system with the GMM-based system described in chapter 4 will be presented here.

5.3.1 Impact of visual information for whisper recognition

Table 5.3 provides the recognition scores for all phones as well as separately for all vowels and consonants presented in the test corpus. These results show the positive contribution of visual information for the recognition task. On average, all phones considered, the input facial movements improve recognition rate by 13.2% (61.35% to 74.52%) for the biphone system and 5.9% (65.25% to 71.10%) for the monophone system. In the case of vowel recognition, the accuracy obtained by using the visual information is 78.28% for the biphone system and 77.36% for the monophone system, showing an improvement of 7% and 3.3% respectively compared with using acoustic information only. In the case of consonant recognition, this improvement is of 15.7% (56.62% to 72.35%) for the biphone system and 6.4% for the monophone one (61.71% to 68.1%). Note that in Japanese phonology, the number of vowels is only 5 while the number of consonants is 27.

The vowel recognition may be easier than that of consonants. The lesser improvement of vowels in the present of facial movements compare to that of consonants can be attributed to this fact.

Table 5.3. *Recognition ratio for all vowels, consonants and all the phones represented in the test corpus*

a. monophone system

Number of Gaussians	AU (%)			AUVI (%)		
	Vowels	Cons	All	Vowels	Cons	All
1	70.76 (±29.70)	58.90 (±30.27)	62.83 (±30.28)	75.70 (±23.59)	66.91 (±29.01)	69.95 (±27.79)
2	74.10 (±30.99)	59.44 (±30.83)	64.08 (±31.29)	77.36 (±23.93)	67.94 (±29.11)	**71.10** (±27.93)
3	61.83 (±30.56)	58.63 (±26.67)	**65.25** (±30.52)	75.28 (±30.80)	68.10 (±30.25)	70.73 (±30.04)
4	71.08 (±31.18)	61.71 (±30.58)	65.04 (±30.67)	76.91 (±31.33)	67.22 (±30.86)	70.47 (±30.75)

b. right biphone context system

Number of Gaussians	AU (%)			AUVI (%)		
	Vowels	Cons	All	Vowels	Cons	All
1	66.70 (±25.73)	57.71 (±28.35)	61.07 (±28.06)	77.94 (±21.05)	71.68 (±28.27)	73.98 (±26.52)
2	71.30 (±25.10)	56.62 (±28.96)	**61.35** (±28.80)	78.28 (±21.80)	72.35 (±28.43)	**74.52** (±26.71)
3	64.41 (±29.78)	58.63 (±26.67)	61.21 (±27.55)	75.20 (±25.63)	71.82 (±27.71)	73.41 (±26.83)
4	62.46 (±31.81)	58.20 (±28.45)	60.44 (±29.25)	76.79 (±23.19)	66.33 (±31.14)	69.75 (±29.43)

Table 5.4 shows the contribution of facial movements to the recognition of consonants clustered according to their place of articulation. The consonants considered here are classified into 4 groups: bilabials, alveolars, palatals and velars. The bilabials benefit from a very significant improvement (32.4%, from 63.13% to 95.57% for the biphone system and 26.2%, from 61.97% to 88.17% for the monophone system) while alveolars display an improvement (11.7%, from 65.37%

to 77.07% and 6.5%, from 66.59% to 73.09% in biphone and monophone case respectively). Note that facial movements impact the recognition of the other consonants. For palatals, the improvement due to facial movements is of 11.3% for the biphone system and of 6% for the monophone system. For velars, the facial contribution is null in the biphone system: the recognition ratio is 80.82% with audio only and 80.65% with an audiovisual input; in the monophone system, adding facial movements for velars decreases the score by 8.9%. The small number of occurrences of velars (*/kj/, /gj/*) in the test corpus probably cause the significant improvement of velar recognition when the number of Gaussians increases (improvement from 56.45% to 80.82% in the biphone system). A possible interpretation, is that since articulatory movements associated with velar phones are not visible, adding visual information does not help recognition and may even degrade it, if the visual data add confusing information.

Table 5.4. *Recognition ratio with different places of articulation*

a. monophone system

Num of Gauss	AU (%)				AUVI (%)			
	Bil	Pal	Alv	Vlr	Bil	Pal	Alv	Vlr
1	54.03 ±4.15	62.93 ±19.20	60.06 ±26.58	60.93 ±44.53	80.83 ±21.05	68.27 ±9.53	69.31 ±15.70	54.63 ±37.44
2	56.27 ±7.31	58.17 ±32.54	59.74 ±23.82	63.18 ±44.79	82.60 ±18.25	68.93 ±11.91	70.83 ±14.46	55.20 ±37.66
3	60.50 ±0.87	57.13 ±25.07	65.27 ±25.16	66.05 ±45.49	**88.17** ±12.55	63.40 ±20.92	73.09 ±17.23	57.33 ±39.46
4	**61.97** ±10.44	57.60 ±25.85	66.59 ±23.72	66.25 ±45.84	88.17 ±12.55	65.37 ±20.81	71.60 ±20.46	52.15 ±38.31

b. right biphone system

Num of Gauss	AU (%)				AUVI (%)			
	Bil	Pal	Alv	Vlr	Bil	Pal	Alv	Vlr
1	**63.13** ±11.76	56.30 ±11.11	60.08 ±24.90	54.77 ±40.21	87.23 ±17.96	67.90 ±18.53	72.64 ±23.51	80.42 ±22.34
2	54.43 ±29.84	50.40 ±19.33	62.99 ±23.56	56.45 ±41.68	90.0 ±13.23	62.33 ±22.92	77.07 ±22.85	80.65 ±21.90
3	47.16 ±23.70	48.50 ±25.96	65.37 ±21.38	80.82 ±21.90	**95.57** ±4.18	58.10 ±21.62	74.21 ±14.07	78.78 ±22.25
4	51.06 ±27.12	45.10 ±21.80	59.48 ±28.29	80.70 ±21.92	95.30 ±4.56	56.80 ±26.15	71.41 ±19.87	51.78 ±39.16

5.3.2 Impact of visual information for speech synthesis

The GMM-based system that we used as a reference for this comparison is described in chapter 4. A GMM with 16 Gaussians, full covariance matrix is used for the spectral estimation. The number of estimated parameters is of 103680. Global variance is also used to reduce the over-smoothing, which is inevitable in the conventional ML-based parameter estimation.

Table 5.5. *Cepstral distortion between converted speech and target speech (dB) with ideal recognition*

a. monophone system

Modality	**HMM**			
	ngauss=1	**ngauss=2**	**ngauss=3**	**ngauss=4**
AU (dB)	**5.93** (± 0.57)	6.98 (± 0.45)	6.28 (± 0.54)	6.51 (± 0.63)
AUVI (dB)	**5.92** (± 0.57)	7.60 (± 0.45)	8.45 (± 0.64)	6.54 (± 0.66)

b. Right context-biphone system

Modality	**HMM**				
	ngauss=1	**ngauss=2**	**ngauss=3**	**ngauss=4**	**ngauss=5**
AU (dB)	6.48 (± 0.51)	6.47 (± 0.6)	6.42 (± 0.59)	**6.37** (± 0.59)	6.52 (± 0.6)
AUVI (dB)	**5.88** (± 0.56)	6.01 (± 0.58)	6.49 (± 0.6)	6.76 (± 0.64)	

For the HMM system, in order to evaluate the contribution of the stochastic parameter generation to the intelligibility of the converted speech in terms of cepstral distortion between target speech and synthesized speech, we first synthesized spectral parameters from the original phonetic transcription, i.e. simulating a perfect recognition step. The results presented in table 5.5 show that the facial movements also have a positive contribution to the performance of the synthesis task in the biphone system, since the cepstral distortion is improved. In the monophone system, however, facial information does not provide any improvement. In the context-dependent system, the cepstral distortion decreases by 7.7% relatively, from 6.37 dB to 5.88 dB ($F(1,40) = 4.13$, $p = 0.034$). This table allows us to select the best synthesis model: in the monophone system, it is the monogaussian model, for both audio only and audiovisual input; in the biphone

system, the best synthesis models are the monogaussian one with audiovisual input and the quadruple-gaussian one with audio only input. Table 5.6 presents the cepstral distortion obtained with synthesized parameters which were generated by the best synthesis models from the phonetic sequences decoded by the recognition module. Again, adding visual information improves the cepstral distortion of the converted spectral parameters by 8.3% (from 6.98 dB to 6.4 dB) with $F(1,40) = 35.1$, $p < 0.001$, for the biphone system and by 1.5% (from 6.76 dB to 6.66 dB) with $F(1,40) = 0.5$, $p = 0.48$ for the monophone one.

Table 5.6 *Cepstral distortion between converted speech and target speech (dB)*

a. monophone system

Modality	**HMM**			
	ngauss=1	**ngauss=2**	**ngauss=3**	**ngauss=4**
AU (dB)	6.79 (± 0.74)	**6.76** (± 0.74)	6.77 (± 0.72)	6.78 (± 0.73)
AUVI (dB)	**6.66** (± 0.74)	6.68 (± 0.72)	6.76 (± 0.68)	6.77 (± 0.73)

b. Right context-biphone system

Modality	**HMM**				**GMM**
	ngauss=1	**ngauss=2**	**ngauss=3**	**ngauss=4**	
AU (dB)	7.10 (± 0.65)	7.02 (± 0.72)	**6.98** (± 0.70)	7.03 (± 0.75)	5.99 (± 0.66)
AUVI (dB)	**6.40** (± 0.67)	6.40 (± 0.67)	6.57 (± 0.79)	6.74 (± 0.97)	5.77 (± 0.61)

Although facial movements have a positive contribution in both HMM-based and GMM-based systems, the performance of the HMM-based system is currently inferior compared with the direct signal-to-signal system based on GMM model. First, the diagonal covariance currently used for each state of the models in the HMM-based system does not take into account the covariance between whispered speech parameters and speech parameters, but the GMM-based system does, by using a full covariance matrix. Second, synthesis and recognition are used separately: the trained HMM models tend to minimize the recognition error, but not the final reconstruction error.

5.4 Summary

This chapter describes an audio-visual whisper to speech conversion procedure that couples a speech synthesis system with a speech recognizer instead of using direct mapping function. The facial movements seem to act as a compensation for lip radiation loss in the signal captured by the NAM microphone. Integrating them noticeably improves the performance of such a system, especially for the recognition task. The experimental results also show that the contribution of visual data depends on place of articulation.

The performance of the HMM-based system is however currently inferior compared with the direct signal-to-signal system based on GMM model. This should be confirmed by a subjective test but objective performances are still too different to motivate such an additional benchmark. As mentioned above, this inferior score could be explained by two reasons.

- First, the diagonal covariance currently used for each state of the models in the HMM-based system does not take into account the covariance between whispered speech parameters and speech parameters, but the GMM-based system does, by using a full covariance matrix. We hope that by modelling the covariance between whispered speech parameters and speech parameters, using more data, extending the acoustic models as well as the linguistic model, and by exploiting state-dependent variance, the performance of this system will further improve.
- Second, synthesis and recognition are used separately, therefore the trained HMM models tend to minimize the recognition error, but not the final reconstruction error. We think that a more intimate coupling of recognition and synthesis – obtained for example by considering trajectory formation accuracy in HMM training based on Minimum Generation Error (MGE) criterion (Wu *et al.*, 2006, 2008) or by considering N-best solutions in the synthesis process – should overcome the limitation of the proposed approach.

Conclusion and perspectives

A brief review

The focus of this thesis is the conversion of whispered speech to phonated speech by using a statistical framework. If this conversion could successfully be carried out, silent cell-phones, speech communication in adverse conditions, speaking-aids for laryngeal handicaps or new human-computer interfaces would become feasible.

By using a condenser microphone attached on the neck, under one ear of the speaker to capture whispered speech, we propose several solutions to improve the intelligibility and the naturalness of the speech generated by whisper-to-speech conversion systems.

The first system that we used is based on a GMM model. It is inspired by the original system proposed by Toda and Shikano (2005) at NAIST. Although the original system successfully predicts the three parametric streams of the STRAIGHT speech vocoder (F_0, harmonic and noise spectrum) from a spectral characterization of the non-audible-murmur input, the authors claimed that the quality of the converted speech is however still insufficient for computer-mediated communication, notably because of the poor estimation of F_0 from unvoiced speech and because of impoverished phonetic contrasts. The substantial improvements that we brought to the original system are presented in Chapter 4. Our first improvement unsurprisingly concerns characteristics of the voiced source. We have shown that the estimation of voicing and F_0 by separate predictors improves both predictions. F_0 prediction is then further improved by the use of a large input context window (>400 ms) to get more spectral variation compared to the original smaller window (105 ms). Predictions of source and spectral streams are both improved by a data reduction technique using LDA instead of PCA. LDA slightly helps with subtle traces of that PCA blindly discards. Another issue concerned with this system is to compare objective and subjective benefits offered by multimodal data. We have shown that improvements obtained by adding visual information in the input stream as well as predicting articulatory parameters for output facial animation are very significant, especially in the case of bilabial consonants where the visual information has an important impact for both production and perception.

We also studied in this thesis another whisper-to-speech conversion system that couples a speech synthesis system with a speech recognizer by jointly optimising a single model for both whispered speech and phonated speech information using the multi-stream functionality supported by the HTK toolkit. The joint probability densities of whispered audiovisual and phonated audiovisual parameters and common state duration densities are modelled by context-dependent phone-sized HMM. Our experiments show that including facial movement noticeably improves the performance of such a system, especially for the recognition task. Furthermore, this positive contribution depends on the place of articulation. Another results from our experiments also proved the inconsistency between the training and the generation of such system. Gaussian mixture models (GMMs) were chosen instead of single Gaussian densities for the output distribution in each state to provide a richer modelling capacity. The best performance of recognition doesn't bring about the best performance for synthesis. Finally, our experiments also show that the objective performance of our HMM-based system is currently inferior compared with the direct signal-to-signal system based on GMM model.

Future research

Based on the research in this work, I would like to highlight some possible directions for future works.

HMM-based system

The major difficulty for the HMM-based whisper-to-speech conversion system is how to tackle both speech recognition and synthesis in a unified framework due to the lack of a coherent statistical model that can be used for both problems. HMMs that have been successfully used in ASR as recognition components, were found to be inadequate for use as the generative output component in a synthesis system (Wu *et al.*, 2006). In the current system, we use Maximum Likelihood Estimation (MLE) criterion for the training of HMM models used by recognition and synthesis modules. However, the MLE criterion is the origin of this inconsistency between training and generation. In fact, the MLE criterion only evaluates the model's pertinence to the data in the likelihood sense which does not reflect the final distance between the generated parameters and the target vectors. MLE criterion is clearly not suitable for the aim of HMM-based speech synthesis which is to generate a synthetic speech (acoustic parameters) as close to the natural speech as possible. Some studies in speech recognition suggested that a training criterion that

directly minimises the error measure on the training set, such as the Minimum Classification Error (MCE) criteria (Juang *et al.*, 1997) or Maximum Mutual Information (MMI) (Valtchev *et al.*, 1997), tends to produce a better model for recognition. In the same context, we could use the Root Mean Square (RMS) criterion proposed by Zhang (2009) or Minimum Generation Error (MGE) based HMM training (Wu and Wang, 2006) to eliminate this inconsistency between training and generation. Furthermore, the over-smoothing problem of generated speech features could be alleviated by incorporating the Global Variance (GV) which represent a penalty for a reduction of the variance of the generated trajectory (Wu *et al.*, 2008).

The second critical issue is the lack of mutual constraints between static and dynamic features in HMM (Wu and Wang, 2006). The conditional independence assumption between state outputs of the hidden Markov model, which constraints the distribution at each time depends only on the state at that time, obstructs the modelling of temporal correlation characteristics in human speech, such as the coarticulation phenomenon and is regarded as the major drawback of HMM (Zhang, 2009). Some solutions are proposed to fix this problem. A simple method to enhance HMM's modelling capacity proposed by Furui (1986) is to use the static acoustic feature vector with dynamic features computed as a linear transformation of several adjacent acoustic feature vectors to introduce the intra-frame dependence. Although the use of these dynamic features (delta and delta-delta features) also improves the performance of HMM-based speech recognizers, it has been thought of as an ad hoc rather than an essential solution (Tokuda *et al.*, 2003). Trajectory-HMM proposed by Tokuda *et al.* (2004) offers an interesting direction which replaces conventional HMM with a new model that can explicitly model the inter-frame dependencies and hidden dynamics in speech. This model also allows to share coherent knowledge between speech recognition and synthesis components.

Noise reduction

The signal captured by NAM microphone contains noise that influences the performance of the conversion system (cf. section 1.3.4). Noise reduction is obviously an important task. In order to enhance the Signal to Noise Ratio (SNR) of the NAM microphone, we could use the correlations between signals captured by two NAM microphones fixed on the speaker's neck or we could use other multimodal correlations, i.e. the correlation of NAM and visual information.

Multimodal data

Speech is multimodal in nature. The research on exploiting visual cues from speaker's face focuses on the robustness in noise for speech recognition systems. The introduction of speaking faces or avatars in speech synthesis systems also improves their naturalness and intelligibility. In general, accounting for the visual aspect of speech in ways inspired by the human speech production and perception mechanisms can substantially benefit to automatic speech processing and human-computer interfaces.

In this context, we used visual cues extracted from the speaker's face as a complementary information to the acoustic signal captured by the NAM microphone, whose high frequency components are attenuated due to the absence of lip radiation. Although the results show the positive contribution of this information for our conversion systems, the visible face movements only present a partial articulatory apparatus: the lips and the jaw movements, and this information is insufficient to adequately convey all spoken information. Including other data related to the movements of inner speech organs (i.e. tongue and velum displays such as EMA, ultrasound image of the tongue), or to laryngeal activity (EGG, EMG), etc., would obviously be very promising.

Real-time issue

Real-time issues also need to be considered in order to use a NAM microphone for silent communication in daily life. Mapping algorithms work on large speech chunks often equal to isolated sentences and combine context-dependent estimations of speech frames using sliding windows. The amount of contextual information has a strong impact on performance. Realistic applications will require this context to be truncated to a maximum duration to allow conversation. Telecom companies estimate that a maximal delay of 200ms is tolerated to allow for full duplex (Guéguin, 2006; Guéguin *et al.*, 2008). Above this threshold speech overlap is prohibited and the conversation enters a mode similar to the push-to-talk mechanism. We will thus study the impact of the limitation of contextual information with regards to performance and subjective quality.

References

Abe, M., Nakamura, S., Shikano, K. and Kuwabara, H., 1988. Voice conversion through vector quantization. In Proc. of ICASSP '88, pp. 655-658.

Abe, M., Nakamura, S., Shikano, K. and Kuwabara, H., 1990. Voice conversion through vector quantization. J. Acoust. Soc. Jpn. (E), E-11(2), pp.71-76.

Amano, S. and Kondo, T., 1999. Lexical Properties of Japanese. Sanseido.

Arslan, L. M., 1999. Speaker transformation algorithm using segmental codebooks (STASC). Speech Communication vol. 28 (3), pp. 211-266.

Badin, P., Bailly, G., Revéret, L., Baciu, M., Segebarth, C., & Savariaux, Christophe. 2002. Three-dimensional linear articulatory modeling of tongue, lips and face based on MRI and video images. Journal of Phonetics, 30(3), pp. 533-553.

Bailly, G., Bérar, M., Elisei, F. and Odisio, M., 2003. Audiovisual speech synthesis. International Journal of Speech Technology, vol. 6, no. 4, pp. 331–346.

Bailly, G., Elisei, F., Badin, P. and Savariaux, C. 2006. Degrees of freedom of facial movements in face-to-face conversational speech. International Workshop on Multimodal Corpora, Genoa – Italy, pp. 33-36.

Bailly, G., Bégault, A., Elisei, F. and Badin, P., 2008. Speaking with smile or disgust: data and models. In Proc. of AVSP, Tangalooma - Australia, pp. 111-116.

Baudoin, G. and Stylianou, Y., 1996. On the transformation of the speech spectrum for voice conversion, In Proc. of ICSLP, vol. 3, pp. 1405-1408.

Benoît, C., Mohamadi, T. and Kandel, S., 1994. Effects of phonetic context on audio-visual intelligibility of French, Journal of Speech and Hearing Research, vol. 37, no. 5, pp. 1195–1203.

Benoît, C. and Le Goff, B., 1998. Audio-visual speech synthesis from French text: Eight years of models, designs and evaluation at the ICP. Speech Communication, vol. 26, pp. 117-129.

Ben Youssef, A., Badin, P., Bailly, G. and Heracleous, P., 2009. Acoustic-to-articulatory inversion using speech recognition and trajectory formation based on phoneme hidden Markov models. In Proc. of Interspeech, Brighton, UK, pp. 2255-2258.

Beskow, J., 2003. Talking heads - models and applications for multimodal speech synthesis, Ph.D. thesis, Department of Speech, Music and Hearing, KTH, Stockholm, Sweden.

Bishop, C. M., 1996. Neural Networks for Pattern Recognition.

Black, A., 2003. Unit selection and emotional speech. In: Proc. Eurospeech, pp. 1649–1652.

Blum, A., 1992, Neural Networks in C++, NY: Wiley.

Bourlard, H. and Dupont, S., 1996. A new ASR approach based on independent processing and recombination of partial frequency bands. In Proc. of ICSLP, pp 426-429.

Bozkurt, B., Ozturk, O. and Dutoit, T. 2003. Text design for tts speech corpus building using a modified greedy selection. In Proc. of Eurospeech, Genova, Swizerland, pp. 277-280.

Callies, J.,2006. Further Investigations on Unspoken Speech. Institut for Theoretische Informatique, University Karlsruhe.

Catford, J. C., 1977. Fundamental Problems in Phonetics. Edinburgh, Edinburgh University Press.

Chan, A.D.C., Englehart, K., Hudgins, B. and Lovely, D. F., 2002. Hidden markov model classification of myoelectronic signals in speech, IEEE/EMBS. vol.2, pp. 1727- 1730.

Coleman, J., Grabe, E. and Braun, B., 2002. Larynx movements and intonation in whispered speech, Summary of research supported by British Academy.

Cootes, T.F., Edwards, G.J. and Taylor, C.J., 2001. Active Appearance Models. IEEE Trans. on Pattern Analysis and Machine Intelligence. 23(6): pp. 681-685.

Dasalla, C.S., Kambara, H., Sato, M. and Koike, Y., 2009. Single-trial classification of vowel speech imagery using common spatial patterns, Neural Networks Journal, Special Issue, Volume 22(9), pp. 1334-1339.

Dempster A. P., N. M., Laird and D. B. Rubin, 1977. Maximum likelihood from incomplete data via the EM algorithm. J. Royal Statist. Soc. Ser. B (methodological), 39(1), pp. 1-22 and 22-38 (discussion).

Denby, B. and Stone, M., 2004. Speech synthesis from real time ultrasound images of the tongue. In: Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing, (ICASSP'04), Montreal, Canada, 17–21, Vol. 1, pp. I685–I688.

Denby, B., Oussar, Y., Dreyfus, G. and Stone, M., 2006. Prospects for a Silent Speech Interface Using Ultrasound Imaging. IEEE ICASSP, Toulouse, France, pp. I365–I368.

Denby, B., Schultz, T., Honda, K., Hueber, T., Gilbert, J.M. and Brumberg, J.S., in press. Silent speech interfaces. Speech Communication, Special Session.

Desai, S., Raghavendra E.V., Yegnanarayana, B., Black, A.W. and Prahallad, K., 2009. Voice conversion using artifical neural networks. In Proc. of ICASSP, pp. 3893-3896.

Dohen, M., Loevenbruck, H., Cathiard, M.-A. and Schwartz, J.-L., 2004. Visual perception of contrastive focus in reiterant French speech, Speech Communication, vol. 44, pp. 155-172.

Dohen, M. and Lœvenbruck, H., 2009a. Interaction of audition and vision for the perception of prosodic contrastive focus. Language and Speech, 52 (3), pp. 177-206.

Dohen, M., Lœvenbruck, H. and Hill, H., 2009b. Recognizing prosody from the lips: is it possible to extract prosodic focus from lip features? In Alan Wee-Chung Liew & Shilin Wang (Eds.), Visual Speech Recognition: Lip Segmentation and Mapping, Medical Information Science Reference, Hershey, New York, ISBN 978-1-60566-186-5, pp. 416-438.

Duda R. O., Hart, P. E. and Stork, D. G., 2000. Pattern classification (2nd edition). John Wiley & Son, Inc., New York.

Dutoit, T. and Leich, H., 1993. MBR-PSOLA: Text-To-Speech synthesis based on an MBE re-synthesis of the segments database. Speech Communication, Vol. 13, Nos. 3-4, pp. 435-440.

Eklund, I. and Traunmuller, H., 1996. Comparative study of male and female whispered and phonated versions of the long vowels of Swedish. Phonetica 54, pp. 1–21.

Elisei, F., Odisio, M., Bailly, G. and Badin, P., 2001. Creating and controlling video-realistic talking heads. Auditory-Visual Speech Processing Workshop, pp. 90-97.

Erber, N.P., 1969. Interaction of audition and vision in the recognition of oral speech stimuli. Journal of Speech and Hearing Research 12, pp. 423-425.

Erber, N.P., 1975. Auditory-visual perception of speech. Journal of Speech and Hearing Research 40, pp. 481-492.

Fadiga, L., Craighero, L., Buccino, G. and Rizzolatti, G., 2002. Speech listening specifically modulates the excitability of tongue muscles: a TMS study. Eur. J. Neurosci., 15 (2), pp. 399-402.

Fagan, M.J., Ell, S.R., Gilbert, J.M., Sarrazin, E. and Chapman, P.M., 2008. Development of a (silent) speech recognition system for patients following laryngectomy. Med. Eng. Phys. 30 (4), pp. 419-425.

Franois, H. and Boffard, O., 2002. The Greedy Algorithm and its Application to the Construction of a Continuous Speech Database, in 3rd International Conference on Language Resources and Evaluation (LREC 2002), vol. 5, pp. 1420-1426.

Gao, M., 2002. Tones in whispered Chinese: articulatory features and perceptual cues. MA thesis, British Columbia Canada: Department of Linguistics, University of Victoria.

Gibert, G., 2006. Conception et évaluation d'un système de synthèse 3D de Langue française Parlée Complétée (LPC) à partir du texte. PhD thesis, Institut National Polytechnique, Grenoble.

Govokhina, O., Bailly, G., Breton, G. and Bagshaw, P., 2006. TDA: A new trainable trajectory formation system for facial animation. In Proc. of InterSpeech, Pittsburgh, pp. 2474-2477.

Govokhina, O., 2008. Etude et modélisation de la coarticulation pour l'animation d'un humanoïde de synthèse. PhD thesis, Institut National Polytechnique, Grenoble.

Grant, K. W. and Seitz, P. F., 2000. The use of visible speech cues for improving auditory detection of spoken sentences. JASA 108, pp. 1197-1208.

Guéguin, M., 2006. Evaluation objective de la qualité vocale en contexte de conversation. PhD Thesis. LTSI. l'Université de Rennes 1.

Guéguin, M., Le Bouquin-Jeannès, R., Gautier-Turbin, V., Faucon, G. and Barriac, V. 2008. On the evaluation of the conversational speech quality in telecommunications. EURASIP Journal on Advances in Signal Processing.

Heeren, W., Van Heuven, V.J., 2009. Perception and Production of Boundary Tones in Whispered Dutch, In Proc. of Interspeech, pp. 2411-2414.

Heracleous, P., Nakajima, Y., Saruwatari, H. and Shikano, K., 2005. A tissue-conductive acoustic sensor applied in speech recognition for privacy. International Conference on Smart Objects & Ambient Intelligence, Grenoble – France, vol. 121, pp. 93 - 97.

Heracleous, P., Beautemps, D., Tran, V.-A., Loevenbruck, H. and Bailly, G., 2009. Exploiting visual information for NAM recognition, IEICE Electronics Express, vol. 6(2), pp. 77-82.

Higashikawa, M., Nakai, K., Sakakura, A. and Takahashi, H., 1996. Perceived Pitch of Whispered Vowels – Relationship with formant frequencies: A preliminary study. Journal of Voice, pp. 155-158.

Higashikawa, M. and Minifie, F. D., 1999. Acoustical-perceptual correlates of whispered pitch in synthetically generated vowels. In Journal of Speech, Language, and Hearing Research. vol 42, pp. 583-591.

Higashikawa, M., Green, J. R., Moore, C. A. and Minifie F. D., 2003. Lip kinematics for /p/ and /b/ production during whispered and voiced speech. Folia Phoniatrica Logopaedica, vol. 55, pp. 1–9.

Hoit, J., and Hixon, T., 1987. Age and speech breathing. In Journal of Speech and Hearing Research, vol. 32, pp. 351-366.

Honda, K., 2007. Physiological Processes of Speech Production. In the Springer Handbook of Speech Processing. Jacob Benesty, M. Mohan Sondhi, Yiteng Huang.

Huang X. D., Y. Ariki and Mervyn A. Jack, 1991. Hidden Markov Models for Speech Recognition. (Edinburgh Information Technology Series, 7).

Hueber, T., Chollet, G., Denby, B., Dreyfus, G. and Stone, M., 2007. Continuous-speech phone recognition from ultrasound and optical images of the tongue and lips. In Proc. of Interspeech, Antwerp, Belgium, pp. 658-661.

Hueber, T., Chollet, G., Denby, B., Dreyfus, G. and Stone, M., 2008a. Towards a segmental vocoder driven by ultrasound and optical images of the tongue and lips. In Proc. of Interspeech, Brisbane, Australia, pp. 2028-2031.

Hueber, T., Chollet, G., Denby, B., Dreyfus, G. and Stone, M., 2008b. Phone Recognition from Ultrasound and Optical Video Sequences for a Silent Speech Interface. In Proc. of Interspeech, Brisbane, Australia, pp. 2032-2035.

Hueber, T., Benaroya, E.-L., Chollet, G., Denby, B., Dreyfus, G., Stone, M., 2009. Visuo-Phonetic Decoding using Multi-Stream and Context-Dependent Models for an Ultrasound-based Silent Speech Interface. In Proc. of Interspeech, pp. 640-643.

Inouye, T. and Shimizu, A., 1970. The electromyographic study of verbal hallucinations. Journal Nerv. Mental Disease 151: 415-422.

Itakura, F., 1975, "Minimum Prediction Residual Principle Applied to Speech Recognition", IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 23 (1), pp. 67-72.

Ito, T., Takeda, K. and Itakura, F., 2005. Analysis and recognition of whispered speech. Speech Communication, 45 (2), pp.139-152.

Iwahashi N. and Sagisaka, Y., 1994. Speech spectrum transformation by speaker interpolation. In Proc. of IEEE Int. Conf. Acoust., Speech, Signal Processing.

Iwahashi, N. and Sagisaka, Y., 1995. Speech spectrum conversion based on speaker interpolation and multi-functional representation with weighting by radial basis function networks. Speech Communication, 16(2), pp. 139-151.

Jacobson, E. 1931. Imagination, recollection, and abstract thinking involving the speech musculature. Am. J. Physiology, vol. 97, pp. 200-209.

Jasper H.H., 1958. The Ten-Twenty Electrode System of the International Federation. Electroencephalography and Clinical Neurophysiology. EEG Journal, vol. 10, pp. 371–375.

Jorgensen, C., Lee, D., and Agabon, S., 2003. Sub auditory speech recognition based on EMG signals. In Proc. of Internat. Joint Conf. on Neural Networks (IJCNN), pp. 3128–3133.

Jorgensen, C. and Binsted, K., 2005. Web browser control using EMG based sub vocal speech recognition. In Proc. 38th Annual Hawaii Internat. Conf. on System Sciences. IEEE, pp. 294c.1–294c.8.

Jou, S., T., Schultz, M., Walliczek, M. and Kraft, F., 2006. Towards continuous speech recognition using surface electromyography. In Proc. of Interspeech 2006 and 9th Internat. Conf. on Spoken Language Processing, vol. 2, pp. 573-576.

Jou, S. and Schultz, T., 2008. EARS: ElectroMyographical Automatic Recognition of Speech. Proceedings of the First International Conference on Biomedical Electronics and Devices, BIOSIGNALS 2008, vol.1, pp.3-12.

Jou, S., 2008. Automatic Speech Recognition on Vibrocervigraphic and Electromyographic Signals. PhD thesis, Carnegie Mellon University.

Jovičić, S. T., 1998. "Formant feature differences between whispered and voiced sustained vowels", Acustica-acta acustica, vol. 84, pp. 739-743.

Jovičić, S. T. and Dordevic, M., 1996. "Acoustic features of whispered speech", Acustica-acta acustica, vol. 82, p. S228.

Jovičić, S. T. and Šarić, Z., 2008. Acoustic Analysis of Consonants in Whispered Speech. Journal of voice, vol. 22(3), pp. 263-274.

Juang, B.-H., Chou, W. and Lee, C.-H., 1997. Minimum Classification Error Rate Methods for Speech Recognition. IEEE Transactions on Speech and Audio Processing, vol. 5 (3), pp. 257-265.

Kain, A. and Macon, M., 1998a. Spectral voice conversion for text-to-speech synthesis. In Proc. of ICASSP '98, vol. 1, pp. 285-288.

Kain, A. and Macon, M., 1998b. Text-to-speech voice adaptation from sparse training data. In Proc. of ICSLP '98, vol. 7, pp. 2847-2850.

Kain, A. and Macon, M., 1998c. Personalizing a speech synthesizer by voice adaptation. In Third ESCA/COCOSDA International Speech Synthesis Workshop, pp. 225-230.

Kain, A., 2001. High resolution voice transformation. PhD thesis, Oregon Health & Science University.

Kallail, K. J. and Emanuel, F. W., 1984a. Formant-frequency differences between isolated whispered and phonated vowel samples produced by adult female subjects, Journal of Speech and Hearing Research, vol. 27, pp. 245-251.

Kallail, K. J. and Emanuel, F. W., 1984b. An acoustic comparison of isolated whispered and phonated vowel samples produced by adult male subjects, Journal of Phonetics, vol. 12, pp. 175-186.

Kawahara, H., Masuda-Katsuse, I., and A.de Cheveigné, 1999. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and

instantaneous- frequency based F0 extraction: Possible role of a repetitive structure in sounds, Speech Communication. vol. 27 (3-4), pp. 187–207.

Kawahara H., Estill, J. and Fujimura, O., 2001. Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT, MAVEBA 2001, Firentze Italy, pp.13-15.

Kim, E. K., Lee, S. and Oh, Y. H., 1997, Hidden Markov Model Based Voice Conversion Using Dynamic Characteristics of Speaker, Proceedings of the EUROSPEECH, Rhodes, Greece, vol. 5, pp. 2519-2522.

Kominek, J. and Black, A. W., 2003. CMU ARCTIC databases for speech synthesis. CMU-LTI-03-177. Language Technologies Institute, School of Computer Science, Carnegie Mellon University.

Kuhn., R., Junqua, J., Nguyen, P. and Niedzielski, N., 2000. Rapid speaker adaptation in eigenvoice space. IEEE Trans. Speech and Audio Processing, vol. 8 (6), pp. 695–707.

Kuwabara, H., Y. Sagisaka, 1995. Acoustic characteristics of speaker individuality Control and conversion. Speech Communication, Elsevier, vol.16 (2), pp. 165-173.

Lam, Y.-M., Mak, M.-W. and Leong, P. H.-W., 2005. Speech Synthesis from Surface Electromyogram Signal. IEEE International Symposium on Signal Processing and Information Technology, pp. 749-754.

Laroche, J., Stylianou, Y. and Moulines, E., 1993, HNS: Speech modification based on a harmonic + noise model, Proc. IEEE-ICASSP'Y3 Internat. Conf Acoust. Speech Signal Process., Minneapolis, MN, pp. 550-553.

Lass, N. J., Hughes, K. R., Bowyer, M. D., Waters, L. T. and Bourne, V. T., 1974. Speaker sex identification from voiced, whispered, and filtered isolated vowels. J. Acoust. Soc. Am., vol. 56 (4), pp. 1180-1192.

Lee, K.-S., 2008. EMG-based speech recognition using hidden Markov models with global control variables. IEEE Transactions on Biomedical Engineering, 55(3), pp. 930–940.

Linde Y., A. Buzo and Gray, R. M., 1980. An algorithm for vector quantizer design. IEEE Trans. Comm., vol. 28, pp. 84-94.

Livesay, J., Liebke, A., Samaras, M. and Stanley, A., 1996. Covert speech behavior during a silent language recitation task. Percept. Mot. Skills, 83 (3 pt 2), 1355-1362.

Maier-Hein, L., Metze, F., Schultz, T., and Waibel, A., 2005. Session independent non-audible speech recognition using surface electromyography. In Proc. ASRU, San Juan, Puerto Rico, pp. 331-336.

Maier-Hein L., 2005. Speech Recognition Using Surface Electromyography. PhD thesis, University at Karlsruhe.

Manabe, H., Hiraiwa, A., and Sugimura, T., 2003. Unvoiced speech recognition using EMG-Mime speech recognition. In Proc. CHI, Ft. Lauderdale, Florida, pp. 794 - 795.

Manabe, H. and Zhang, Z. 2004. Multi-stream HMM for EMG-based speech recognition. In Proc. IEEE EMBS, San Francisco, California, pp. 4389-4392.

Massaro, D. W., 1998. Perceiving Talking Faces: From Speech Perception to a Behavioral Principle, MIT Press, Cambridge, Mass, USA.

Matsuda, M. and Kasuya, H., 1999. Acoustic nature of the whisper. In Proc. of Eurospeech, vol 1., pp. 137–140.

McGlone, R. E. and Manning, W. H., 1979. Role of the second formant in pitch perception of whispered and voiced vowels. Folia Phoniatrica, vol. 31(1), pp. 9-14.

McGuigan, F. J. and Dollins, A. B., 1989. Patterns of covert speech behavior and phonetic coding. Pavlov J. Biol. Sci., vol. 24 (1), pp. 19-26.

Meyer-Eppler, W., 1957. Realization of prosodic features in whispered speech. Journal of the Acoustical Society of America, vol. 29, pp. 104-106.

Moore, B. C. J., 1997. An Introduction to the Psychology of Hearing. San Diego, CA: Academic Press.

Morse, M.S., O'Brien, E.M., 1986. Research summary of a scheme to ascertain the availability of speech information in the myoelectric signals of neck and head muscles using surface electrodes. Comput. Biol. Med., vol. 16 (6), pp. 399-410.

Morse, M.S., Day, S.H., Trull, B., Morse, H., 1989. Use of myoelectric signals to recognize speech. In: Images of the Twenty-First Century Proc. 11th Annual Internat. Conf. of the IEEE Engineering in Medicine and Biology Society, Part 2, vol. 11. Alliance for Engineering in Medicine and Biology, pp. 1793–1794.

Morse, M.S. , Day, S . H ., May, J., 1990. Time Domain Analysis of The Myoelectric Signals Secondary To Speech. 12th IEEE EMBS Conference, Philadelphia, pp. 1318-1319.

Morse, M.S., Gopalan, Y.N., Wright, M., 1991. Speech recognition using myoelectric signals with neural networks. In: Proc. 13th Annual Internat. Conf. of the IEEE Engineering in Medicine and Biology Society, vol. 13 (4), Piscataway, NJ, United States. IEEE, pp. 1877– 1878.

Morris, W. R., 2003. Enhancement and recognition of whispered speech. PhD thesis, Georgia Institute of Technology.

Mouchtaris A., der Spiegel, J.V. and Mueller, P., 2006. Non-parallel training for voice conversion based on a parameter adaptation approach, IEEE Trans. Audio, Speech and Language Processing, vol. 14 (3), pp. 952-963.

Moulines E. and Charpentier, F., 1990. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. Speech Comm., vol. 9, pp. 453–467.

Moulines, E. and Sagisaka, Y., 1995. Voice conversion: State of the Art and Perspectives (Special Issue of Speech Communication). Amsterdam, The Netherlands: Elsevier, vol.16.

Moulines, E. and Laroche, J., 1995. Non-parametric techniques for pitch-scale and time-scale modification of speech, Speech Communication, vol. 16, pp.175-205.

Nakajima, Y., Kashioka, H., Shikano, K. and Campbell, N., 2003a. Non-audible murmur recognition input interface using stethoscopic microphone attached to the skin. In: Proc. of ICASSP, pp. 708–711.

Nakajima, Y., Kashioka, H., Shikano, K. and Campbell, N., 2003b. Non-audible murmur recognition. In Proc. of Eurospeech, pp. 2601–2604, Geneva, Switzerland.

Nakajima, Y., Kashioka, H., Campbell, N. and Shikano, K., 2006. Non-audible murmur (NAM) recognition. IEICE Trans. Inform. Systems E89-D (1), pp. 1-8.

Nakamura, K., Toda, T., Saruwatari, H. and Shikano, K., 2006. A speech communication aid system for total laryngectomies using voice conversion of body transmitted artificial speech. 4th Joint Meeting of the ASA and the ASJ, Hawaii, USA, vol. 120 (5), pp. 3351-3351.

Nakagiri, M., Toda, T., Kashioka, H. and Shikano, K., 2006. Improving Body Transmitted Unvoiced Speech with Statistical Voice Conversion. In Proc. of. ICSLP, pp. 2270-2273.

Narendranath M., H.A. Murthy, S. Rajendran and B. Yegnanarayana, 1995. Transformation of formants for voice conversion using artificial neural networks. Speech Communication, vol. 16(2), pp. 207-216.

Odisio, M., Bailly, G. and Elisei, F., 2004. Tracking talking faces with shape and appearance models. Speech Communication, vol. 44 (1), pp. 63–82.

Ohtani, Y., Toda, T., Saruwatari, H. and Shikano, K., 2006. Evaluation of eigenvoice conversion based on Gaussian mixture model. 4th Joint Meeting of the ASA and the ASJ, Hawaii, USA, vol. 120 (5), pp. 3036-3036.

Ohtani, Y., Toda, T., Saruwatari, H. and Shikano, K., 2007. Speaker adaptive training for one-to-many eigenvoice conversion based on Gaussian mixture model. In Proc. of Interspeech, pp. 1981-1984.

Ohtani, Y., Toda, T., Saruwatari, H. and Shikano, K., 2009. Many-to-many eigenvoice conversion with reference voice. In Proc. of Interspeech, pp. 1623-1626, Brighton, UK.

Ouni, S. and Y. Laprie, 2005. Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion. Journal of the Acoustical Society of America, vol. 118 (1), pp. 444-460.

Ouni, S., Cohen, M.-M., Ishak H. and Massaro, D. W., 2007. Visual Contribution to Speech Perception: Measuring the Intelligibility of Animated Talking Heads. EURASIP Journal on Audio, Speech, and Music Processing.

Porbadnigk, A., Wester, M., Calliess, J. and Schultz, T., 2009. EEG-based speech recognition – impact of temporal effects. In: Biosignals, Porto, Portugal, pp. 376–381.

Potamianos, G., Neti, C., Gravier, G., Garg, A. and Senior, A.W., 2003. Recent advances in the automatic recognition of audiovisual speech. In Proc. of the IEEE, vol. 91 (9), pp. 1306–1326.

Potamianos, G., Neti, C., Luettin, J., and Matthews, I. 2004. Audiovisual automatic speech recognition: an overview. In : Audiovisual speech processing. MIT Press.

Rabiner L. and B.-H. Juang, 1993. Fundamentals of Speech Recognition. Prentice hall Signal processing series.

Redner R. A. and H. F. Walker, 1984. Mixture densities, maximum likelihood and the EM algorithm. SIAM Review, vol. 26(2), pp. 195-239.

Reisberg, D., Mclean, J. and Goldfield, A., 1987. Easy to Hear but Hard to Understand: A Lip-reading Advantage with Intact Auditory Stimuli, in Dodd B. & Campbell R. (Eds.), Hearing by eye: The psychology of lip-reading, Lawrence Erlbaum Associates, Hillsdale (USA), pp. 97-114.

Reynolds, D. A. and Rose, R. C., 1995. Robust text-indipendent speaker identification using gaussian mixture speaker models. IEEE Trans. Speech and Audio Processing, vol. 3 (1), pp. 72-83.

Revéret, L. and Benoit, C., 1998. A new 3D lip model for analysis and synthesis of lip motion in speech production. In Proc. of AVSP, pp. 207-212.

Revéret, L., Bailly, G., and Badin, P., 2000. MOTHER: a new generation of talking heads providing a flexible articulatory control for video-realistic speech animation. In Proc. of ICSLP, vol. 2, pp. 755-758.

Richard, G. and Alessandro, C., 1996. Analysis/synthesis and modification of the speech aperiodic component. Speech Communication, vol. 19, pp. 221-244.

Rubin, A. D., Praneetvatakul, V., Gherson, S., Moyer, C. A. and Sataloff, R., 2006. Laryngeal hyperfunction during whispering: reality or myth? Journal of Voice; vol. 20, pp. 121–127.

Sagisaka, Y., K. Takeda, M. Abe, S. Katagiri, T. Umeda, and H. Kuwabara, 1990. A large-scale Japanese speech database. In Proc. ICSLP, pp. 1089-1092, Kobe, Japan, Nov. 1990.

Scanlon, M., Fisher, F. and Chen, S., 2002. Physiological sensors for speech recognition. Multimodal Speech Recognition Workshop, pp. 1-5.

Schwartz, J.-L., Teissier, P., and Escudier, P., 2002. La parole multimodale : deux ou trois sens valent mieux qu'un. In J. Mariani (Ed.), Traitement automatique du langage parlé - 2 : reconnaissance de la parole, pp. 141-178.

Schwartz, J.-L., Berthommier, F. and Savariaux, C., 2004. Seeing to hear better: evidence for early audio-visual interactions in speech identification. Cognition, vol. 93, pp. 69-78.

Schwartz, M. F. and Rine, H. E., 1968. Identification of speaker sex from isolated whispered vowels. J. Acoust. Soc. Am., vol. 44 (6), pp. 1736-1737.

Sekimoto H., T. Toda, H. Saruwatari, K. Shikano, 2006. Improving quality of small body transmitted ordinary speech with statistical voice conversion. 4th Joint Meeting of the ASA and the ASJ, Hawaii, USA.

Shimizu S., M. Otani and T. Hirahara, 2009, Frequency characteristics of several non-audible murmur (NAM) microphones. Acoust. Sci. & Tech., vol. 30 (2), pp. 139-142.

Sokolov, Y. N., 1972. Inner speech and thought. Plenum Press, New York.

Stevens K. N., 1998. Acoustic Phonetics. MIT Press.

Stylianou, Y., Laroche, J. and Moulines, E., 1995. High-Quality Speech Modification based on a Harmonic + Noise Model. In Proc. of Eurospeech, pp. 451-454.

Stylianou, Y., Cappé, O. and Moulines, E., 1998. Continuous Probabilistic Transform for Voice Conversion. IEEE Trans. on Speech and Audio Processing, vol. 6 (2), pp. 131-142.

Sugie, N., and Tsunoda, K., 1985. A speech prosthesis employing a speech synthesizer - vowel discrimination from perioral muscle activities and vowel production', IEEE Trans. Biorned. Eng., vol. 32, pp. 485-490.

Sumby, W. H. and Pollack, I., 1954. Visual contribution to speech intelligibility in noise. Journal of Acoustical Society of America, vol. 26 (2), pp. 212–215.

Summerfield, Q., MacLeod, A., McGrath, M. and Brooke, M., 1989. Lips, teeth, and the benefits of lipreading. In A. W. Young and H. D. Ellis, editors, Handbook of Research on Face Processing, pp. 223-233. Elsevier Science Publishers.

Summerfield, Q., 1992. Lipreading and audio-visual speech perception. Philosophical transactions of the royal society of London, vol. 33(5), pp. 71-78.

Suppes, P., Lu, Z.-L. and Han, B., 1997. Brain wave recognition of words. Proc. Nat. Acad. Sci. USA 94, 14965–14969.

Tamura, M., Kondo, S., Masuko, T. and Kobayashi, T., 1999. Text-to-audiovisual speech synthesis based on parameter generation from HMM. In Proc. of Eurospeech, pp.959–962.

Tartter, V.C., 1989. What's in whisper? Journal of the Acoustical Society of America, vol. 86, pp. 1678-1683.

Tartter, V. C., 1991. Identifiability of vowels and speakers from whispered syllables, Perception & Psychophysics, vol. 49 (4), pp. 365-372.

Thomas, I. B., 1969. Perceived pitch of whispered vowels. Journal of the Acoustical Society of America, vol. 46 (2), pp. 468-470.

Toda, T., 2003. High-Quality and Flexible Speech Synthesis with Segment Selection and Voice Conversion. PhD thesis, Shikano Lab.

Toda, T., Black, A. W. and Tokuda, K., 2004. Mapping from articulatory movements to vocal tract spectrum with Gaussian mixture model for articulatory speech synthesis. International Speech Synthesis Workshop, Pittsburgh, PA.

Toda, T., Black, A.W. and Tokuda, K., 2005. Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter. In Proc. of ICASSP, vol. 1, pp. 9–12, Philadelphia, USA.

Toda, T. and Shikano, K., 2005. NAM-to-Speech Conversion with Gaussian Mixture Models. In Proc. of Interspeech, pp. 1957-1960.

Toda, T., Ohtani, Y. and Shikano, K., 2006. Eigenvoice conversion based on Gaussian mixture model. In Proc. of Interspeech, pp. 2446-2449, Pittsburgh, USA.

Toda, T., Ohtani, Y. and Shikano, K., 2007a. One-to-many and many-to-one voice conversion based on eigenvoices. In Proc. of ICASSP, pp. 1249-1252, Hawaii, USA.

Toda, T., Black A.-W. and Tokuda, K., 2007b. Voice conversion based on maximum likelihood estimation of spectral parameter trajectory. IEEE Trans. ASLP, Vol. 15, No. 8, pp. 2222–2235.

Toda, T., Black, A.W., Tokuda K., 2008. Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model. Speech Communication, vol. 50 (3), pp. 215-227.

Toda T., Nakamura, K., Sekimoto, H. and Shikano, K., 2009. Voice conversion for various types of body transmitted speech. In Proc. of ICASSP, pp. 3601-3604, Taipei, Taiwan.

Toda, T., Nakamura, K., Nagai, T., Kaino, T., Nakajima, Y. and Shikano, K., 2009. Technologies for processing body-conducted speech detected with non-audible murmur microphone. In Proc. of Interspeech, pp. 632-635, Brighton, UK (Keynote in Special Session).

Tokuda, K., Kobayashi, T. and Imai, S., 1995a. Speech parameter generation from HMM using dynamic features. Proc. of ICASSP, pp. 660–663.

Tokuda, K., Masuko, T., Yamada, T., Kobayashi, T. and Imai, S., 1995b. An algorithm for speech parameter generation from continuous mixture HMMs with dynamic features. Proc. of EUROSPEECH, pp. 757–760.

Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T. and Kitamura, T., 2000. Speech parameter generation algorithms for HMM-based speech synthesis. In Proc. of ICASSP, pp. 1315–1318, Istanbul, Turkey.

Tokuda, K., Zen, H. and Kitamura, T., 2003. Trajectory Modeling based on HMMs with the Explicit Relationship between Static and Dynamic Features. In Proc. of Eurospeech, pp. 865–868.

Tokuda, K., Zen, H. and Kitamura, T., 2004. Reformulating the HMM as a Trajectory Model. Workshop on Statictical modeling Approach for Speech Recognition (Beyond HMM), Kyoto, Japan.

Tokuda, K., Zen, H. and Kitamura, T., 2004. Reformulating the HMM as a trajectory model. In Proc. of Beyond HMM.

Tokuda, K., Zen, H., Yamagishi, J., Black, A., Masuko, T., Sako, S., Toda, T., Nose, T., Oura, K., 2008. The HMM-based speech synthesis system (HTS). <http://hts.sp.nitech.ac.jp/>.

Toutios A. and Margaritis K., 2005. Mapping the Speech Signal onto Electromagnetic Articulography Trajectories Using Support Vector Regression. In the book: Text, Speech and Dialogue, Springer Berlin / Heidelberg, vol. 3658, pp. 318-325.

Toth, A.-R., Wand, M. and Schultz, T., 2009. Synthesizing Speech from Electromyography using Voice Transformation Techniques. In Proc. of Interspeech, pp. 652-655.

Tsunoda, K., Numi, S., Hirose, H., Harris K. S. and Baer, T., 1991 .An Electromyographic Study on Laryngeal Adjustment for Whispering. Ann. Bull. RILP, vol. 25, pp. 33-38.

Valbret, H., Moulines, E. and Tubach, J.P., 1992. Voice transformation using PSOLA technique. Speech Communication, vol. 11, pp. 175-187.

Valtchev, V., Odell, J. J., Woodland, P.C. and Young, S.J., 1997. MMIE training of large vocabulary recognition systems. Speech Communication, vol.22, pp. 303–314.

Van Santen, J.P.H. and Buchsbaum, A.L., 1997. Methods of optimal text selection. In Proc. of Eurospeech, vol.2, pp. 553-556.

Von Helmholtz, H. L. F., 1954. On the sensations of tone. New York: Dover Publications.

Walliczek, M., Kraft, F., Jou, S.-C., Schultz, T. and Waibel, A., 2006. Sub-Word Unit based Non-Audible Speech Recognition using Surface Electromyography. In Proc. of Interspeech, pp. 1487-1490.

Watanabe T., Murakami, T., Namba, M., Hoya, T. and Ishida, Y., 2002. Transformation of spectral envelope for voice conversion based on radial basis function networks, International Conference on Spoken Language Processing, vol. 1, pp. 285-288.

Wester, M. and Schultz, T., 2006. Unspoken speech – speech recognition based on electroencephalography. Master's Thesis, Universität Karlsruhe, Karlsruhe, Germany.

Wolpaw, J. R., Birbaumer, N., McFarlanda, D. J., Pfurtschellere, G. and Vaughana, T. M., 2002. Brain–computer interfaces for communication and control, Clinical Neurophysiology, vol. 113 (6), pp. 767-791.

Woodland, P. C., Odell, J. J., Valtchev, V., and Young., S. J., 1994. Large vocabulary continuous speech recognition using HTK. In Proc. of ICASSP'94, pp. 125-128

Wu, Y.-J., Guo, W. and Wang, R.H., 2006. Minimum generation error criterion for tree-based clustering of context dependent HMMs. In Proc. of Interspeech, pp. 2046–2049.

Wu, Y.-J., Zen, H., Nankaku, Y. and Tokuda, K., 2008. Minimum generation error criterion considering global/local variance for HMM-based speech synthesis. ICASSP, pp. 4621-4624.

Xu, S. and Chen, L., 2008. A Novel Approach for Determining the Optimal Number of Hidden Layer Neurons for FNN's and Its Application in Data Mining. 5th International Conference on Information Technology and Applications (ICITA), pp. 683-686.

Young S., Evermann G., Gales M., Hain T., Kershaw D., Liu X.-Y., Moore G., Odell J., Ollason D., Povey D., Valtchev V., Woodland P., 2006. The Hidden Markov Model Toolkit (HTK) Version 3.4. <http://htk.eng.cam.ac.uk/>.

Zell et al., 1996. SNNS Stuttgart Neural Network Simulator User Manual, Version 4.2. http://www.ra.cs.uni-tuebingen.de/SNNS/

Zen, H., Tokuda, K., and Kitamura, T., 2004. A Viterbi algorithm for a trajectory model derived from HMM with explicit relationship between static and dynamic features. In Proc. of ICASSP, pp. 837–840.

Zen, H., Tokuda, K. and Kitamura, T., 2004. An introduction of trajectory model into HMM-based speech synthesis. Fifth ISCA ITRW on Speech Synthesis (SSW5), Pittsburgh, PA, USA, pp. 191-196.

Zen, H., Tokuda, K. and Black, A.W., 2009. Statistical parametric speech synthesis. Speech Communication, vol. 51, pp. 1039-1064.

Zeroual, C., H.Esling, J. and Crevier-Buchman, L., 2005. Physiological study of whispered speech in Moroccan Arabic. In Proc. of Interspeech, pp. 1069-1072.

Zhang Y., X. Shi, J. Lei, H. Xu, K. Huang and C. H. Chen, 2005. Gaussian Mixture Model for Underdetermined Source Separation. Neural Networks and Brain. ICNN&B '05, vol. 3, pp. 1965-1969.

Zhang, L., 2009. Modelling speech dynamics with trajectory-HMMs. PhD thesis, the university of Edinburgh.

Appendix A

Résumé en français de la thèse

Cette annexe contient un résumé détaillé en français du travail effectué dans cette thèse.

1 Introduction

La parole silencieuse ou murmurée est définie comme la production articulée de sons, avec très peu de vibration des cordes vocales dans le cas du chuchotement, et aucune vibration dans le cas du murmure, produite par les mouvements et les interactions des organes de la parole tels que la langue, le voile du palais, les lèvres, etc., dans le but d'éviter d'être entendue par plusieurs personnes. La parole silencieuse ou murmurée est utilisée généralement pour la communication privée et confidentielle ou peut être employée par les personnes présentant un handicap laryngé et qui ne peuvent pas parler normalement.

Cependant, il est difficile d'employer directement la parole silencieuse (murmurée) pour la communication face à face ou avec un téléphone portable parce que le contenu linguistique et l'information paralinguistique dans le message prononcé sont dégradés fortement quand le locuteur murmure ou chuchote. Une piste récente de recherche est donc celle de la conversion de la parole silencieuse (ou murmurée) en voix claire afin d'avoir une voix plus intelligible et plus naturelle. Avec une telle conversion, des applications potentielles telles que la téléphonie silencieuse » ou des systèmes d'aides robustes pour les handicaps laryngés deviendraient envisageables. Notre travail dans cette thèse se concentre donc sur cette piste.

Plusieurs dispositifs d'interface de parole silencieuse ont été explorés dans la littérature, notamment l'électromyographie de surface (sEMG) (Jorgensen *et al.*, 2003; Jorgensen et Binsted 2005 ; Jou, Schultz *et al.*, 2006, 2008; Walliczek *et al.*, 2006 ; Toth *et al.*, 2009), le microphone capteur de murmure inaudible (NAM) (Nakajima, 2003 ; Heracleous *et al.*, 2005; Toda et Shikano 2005), l'ultrason et l'image optique (Denby et Stone, 2004 ; Denby *et al.*, 2006 ; Hueber *et al.*, 2007, 2008ab, 2009 ; Denby *et al.*, 2009), l'articulographie électromagnétique (EMA)

(Fagan *et al.*, 2008) et l'électro-encéphalographie (EEG) (Suppes *et al.*, 1997 ; Wester et Schultz 2006 ; Porbadnigk *et al.*, 2009). Parmi ceux-ci, un dispositif qui semble particulièrement intéressant est le microphone NAM développé par des chercheurs du NAIST (Nara Institute of Science and Technology) au Japon, en raison de son usage facile et de sa taille appropriée pour la télécommunication mobile. Nakajima *et al.* (2003) ont ainsi proposé qu'il pourrait être plus efficace, en environnement bruyant, d'analyser les vibrations de parole issues directement de l'intérieur du corps, en les captant à la surface de la peau, au lieu d'analyser les sons transmis dans l'air, après avoir été émis par la bouche. Ils ont présenté une nouvelle interface de communication qui peut recueillir les vibrations acoustiques issues du conduit vocal à travers la vibration des tissus faciaux, en plaçant un capteur sur le cou, juste en dessous de l'oreille. En utilisant ce capteur microphone, Toda et Shikano (2005) ont proposé un système de conversion du murmure inaudible (NAM) vers la voix claire basé sur un modèle de mélange de gaussiennes (GMM). Il a été montré que ce système est efficace mais la qualité de la parole prédite reste insuffisante, notamment en raison des difficultés dans l'estimation du F_0 (fréquence fondamentale) à partir de la voix inaudible. La contribution principale de notre travail dans cette thèse est d'améliorer la performance d'un tel système.

Ce qui suit est un résumé de nos contributions :

Mise en correspondance directe des signaux basée sur des mixtures de Gaussiennes (GMM)

- **Pitch.** Une première amélioration proposée dans cette thèse concerne l'évaluation de la source vocale pour la parole convertie. Plusieurs approches sont explorées, notamment celles de limiter l'apprentissage uniquement aux segments voisés, de séparer l'estimation du voisement et l'évaluation de F_0 dans le processus de synthèse, d'optimiser la taille de la fenêtre de contexte du vecteur acoustique d'entrée et d'utiliser une analyse discriminante linéaire (LDA) au lieu d'une analyse en composantes principales (PCA) pour réduire la dimension du vecteur d'entrée.

- **Information audiovisuelle en entrée-sortie**. Une autre solution explorée dans cette thèse pour améliorer la performance du système de conversion est d'intégrer l'information visuelle comme complément à l'information acoustique à la fois dans les données d'entrée et de sortie. Les paramètres visuels faciaux sont estimés en utilisant un système de capture de mouvement très précis, développé pour le système de têtes parlantes conçu au Département Parole et Cognition du laboratoire GIPSA.

Mise en correspondance indirecte par pivot phonétique utilisant une reconnaissance-synthèse basée sur des modèles de Markov Cachés (HMM)

Une autre approche pour convertir la parole silencieuse en voix audible est de combiner techniques de reconnaissance et de synthèse de la parole. Cette approche a été mise en œuvre dans (Hueber *et al.*, 2007, 2008ab, 2009). En incluant des informations linguistiques, dans la reconnaissance et dans la synthèse, un tel système peut potentiellement compenser l'entrée appauvrie en intégrant de la connaissance linguistique dans le processus de reconnaissance. Nous comparons cette approche avec la technique de mise en correspondance présentée ci-dessus, pour explorer une autre solution visant à améliorer la qualité du système de conversion du chuchotement en parole claire.

Le travail présenté dans cette thèse contribue au projet CASSIS (communication assistée par ordinateur et interfaces silencieuses) impliquant la collaboration de GIPSA, de l'ENST-Paris, de l'ESPCI-Paris et de NAIST.

2 Etat de l'art

2.1 Parole silencieuse et interfaces de parole silencieuse

Parole silencieuse

La parole silencieuse est un moyen de communication que les locuteurs utilisent pour réduire la perceptibilité de la parole. Par exemple quand ils ont pour consigne de parler doucement pour ne pas gêner les autres dans une bibliothèque, dans une conférence…, quand ils sont trop faibles pour parler normalement ou quand ils communiquent des informations privées, confidentielles. Il semble que la parole silencieuse est la communication vocale la plus efficace quand seules quelques personnes autour du locuteur doivent entendre le message.

Selon le niveau de « silence » ou d'audibilité, nous pouvons définir 5 catégories :

Parole intérieure : Elle est également appelée parole imaginée ou pensée verbale. Elle se rapporte à la production silencieuse mentale de mots ou de phrases.

Parole subvocale invisible : Ce mode de la parole est articulé très doucement de sorte qu'il ne puisse pas être entendu, mais les articulateurs de la parole (langue, lèvres, peut-être mâchoire) peuvent légèrement bouger.

Parole subvocale visible : Ce mode de la parole correspond à l'articulation silencieuse. La parole est articulée mais sans émission d'air.

Murmure inaudible (NAM) : Ce mode de la parole est chuchoté doucement de sorte qu'une personne voisine ne puisse pas entendre.

Parole chuchotée : Bien qu'il soit difficile de définir avec précision les différences acoustiques entre le « chuchotement » et le « NAM », le terme « chuchotement » implique que les auditeurs voisins limités peuvent entendre le contenu du message, et qu'il peut être enregistré par un microphone externe, par la transmission dans l'air.

Interfaces de parole silencieuse

La parole peut être considérée comme un ensemble de signaux multimodaux cohérents entre eux, tels que les signaux cérébraux électiques, les signaux myoélectriques issus des muscles, les mouvements des articulateurs orofaciaux – qu'ils soient visibles ou non -, ou encore le signal acoustique.

Diverse technologies peuvent être employées pour enregistrer les signaux caractérisant l'articulation et la phonation:

- L'activité cérébrale peut être capturée par EEG non-invasive.
- Les activités musculaires peuvent être mesurées par EMG de surface.
- L'articulographie électromagnétique permet de déterminer les mouvements de points sur les articulateurs
- Les déformations de surfaces des articulateurs peuvent être déterminées à partir de cinéradiographie, d'IRM dynamique, d'échographie ultrasonique.
- Les microphones acoustiques permettent de capturer différentes sortes de signaux acoustiques.

Dans le but de réaliser une télécommunication silencieuse, où le murmure inaudible et le chuchotement sont utilisés en entrée, un microphone à condensateur NAM semble être meilleur que d'autres interfaces, y compris l'EMG, l'EMA, l'ultrason et l'EEG en raison de sa facilité d'utilisation et de du fait que le signal recueilli est directement un signal acoustique, plus directement interprétable que ceux capturés par les autres interfaces. C'est pourquoi nous nous concentrons sur le microphone NAM en tant que capteur prometteur.

Microphone NAM : Le son chuchoté, créé par les mouvements coordonnés de la langue, du vélum, des lèvres, etc., peut être capté grâce à la radiation aux lèvres mais aussi par la transmission des chocs entre articulateurs et parois ainsi que la transmission de l'onde de pression par les tissus mous. Le chuchotement peut ainsi être capté par un microphone NAM placé sur la peau, en dessous de l'oreille (Nakajima *et al.*, 2003). Le tissu peaussier et la radiation des lèvres agissent comme un filtre passe-bas et les composantes hautes fréquences sont atténuées. Toutefois, les composantes spectrales du chuchotement (et du murmure inaudible) fournissent assez d'information pour identifier les sons (Heracleous *et al.*, 2005). Le capteur NAM enregistre la parole dans une bande de fréquence allant jusqu'à 4kHz, en étant peu sensible au bruit externe.

Dans le cadre de cette thèse, ce microphone est choisi pour capturer la parole chuchotée.

2.2 Parole chuchotée

Dans la voix modale (ou claire), les sons voisés impliquent une modulation de la circulation d'air issu des poumons par la vibration des cordes vocales. Cependant, il n'y a aucune vibration des cordes vocales dans la production de voix chuchotée. Pour cette raison, les caractéristiques acoustiques du chuchotement diffèrent de celles de la voix modale. Une étude des propriétés acoustiques des voyelles (Ito *et al.*, 2005) a montré une augmentation des fréquences de formant pour les voyelles chuchotées comparées à la voix modale. Le décalage est plus grand pour les voyelles à valeurs formantiques peu élevées. Il a également été constaté que les caractéristiques du conduit vocal pour les phonèmes voisés changent plus dans le chuchotement par rapport à la voix modale que celles des consonnes non-voisées. La perception du *pitch* (hauteur du son) dans la voix modale est principalement liée à la fréquence fondamentale (F_0). Dans le chuchotement, cependant, bien qu'il n'y ait aucune vibration des cordes vocales, une certaine perception de la hauteur peut être possible. Higashikawa *et al.* (1999, 2003) ont montré que les auditeurs peuvent percevoir le *pitch* dans le chuchotement. Selon eux, les changements simultanés des formants F1 et F2 pourraient être l'un des indices qui influencent cette perception.

2.3 Conversion de la parole silencieuse en voix claire

Deux approches principales sont proposées pour produire de la parole audible et visible à partir d'articulations inaudibles :

- Les techniques de mise en correspondance basées sur un GMM peuvent être employées pour convertir directement les signaux chuchotés en parole claire, en utilisant des corpus alignés (des phrases, des mots ou d'autres unités inaudibles et audibles) sans avoir à inclure d'information phonétique: des représentations communes de la parole inaudible et de la parole claire sont stockées ou modélisées puis employées pour estimer directement la parole claire à partir de la représentation des signaux inaudibles.
- Une deuxième approche consiste à utiliser un pivot phonétique et à combiner reconnaissance de NAM avec synthèse de parole modale (Hueber *et al.*, 2007, 2008ab, 2009). Le système de reconnaissance segmente le signal de la parole en unités phonémiques, en utilisant l'information dépendante du signal et un modèle de langage. Un système de synthèse de la parole convertit alors ces unités phonétiques en voix synthétique, en utilisant la voix modale préenregistrée. La performance d'un tel système dépend principalement de la performance de la partie de reconnaissance : une reconnaissance correcte aura comme conséquence une parole reconstruite parfaite tandis que les échecs de reconnaissance ou un modèle de langage insatisfaisant pourront conduire à des dégradations fortes de la qualité de la parole convertie.

3 Contribution de cette thèse

3.1 Technique de mise en correspondance basée sur un modèle de mélange de gaussiennes (GMM)

La technique de mise en correspondance directe de signal-à-signal en utilisant des corpus alignés est très prometteuse. Toda *et al.* (2005) ont appliqué un *mapping* statistique (Stylianou *et al.*, 1998; Kain et Macon, 1998) basé sur un modèle GMM pour la conversion de parole NAM en parole claire. Dans ce système, pour synthétiser la parole, il faut estimer non seulement les traits spectraux mais aussi les traits d'excitation, y compris F_0 et les composants apériodiques. Les traits spectraux à chaque trame ont été construits en concaténant les vecteurs spectraux de plusieurs trames autour de la trame courante, afin de compenser les caractéristiques perdues sur quelques phonèmes, particulièrement les fricatives, d'énergie élevée sur les bandes à haute fréquence. Trois GMMs ont été utilisés pour convertir les traits spectraux du chuchotement en trois ensembles de traits

pour la parole claire, i.e. le spectre, le F_0 et un composant apériodique (qui capture le bruit sur chaque bande de fréquence du signal d'excitation) (cf. 4.2).

Toutefois, bien que l'intelligibilité segmentale des signaux synthétiques calculés par la mise en correspondance soit acceptable, les auditeurs ont des difficultés à découper le flux sonore pour récupérer des mots. Ce problème est dû en partie à la restauration de la mélodie synthétique. Ainsi, dans notre système, adapté de Toda *et al.*, nous nous sommes concentrés sur l'amélioration de l'estimation de la mélodie et de la détection du voisement.

3.1.1 Séparation de l'estimation du voisement et de l'évaluation de F_0

Détection du voisement

Dans le système original, Toda *et al* (2005) estiment la valeur de F_0 pour toutes les trames en utilisant un modèle GMM. Un seuil de valeur de F_0 est déterminé pour assigner une étiquette voisée/non-voisée à chaque trame. Dans notre système, nous avons testé la possibilité d'améliorer la détection du voisement en employant un Réseau de Neurones (RN) *feedforward* simple. De façon empirique, nous avons utilisé 50 neurones d'entrée (*i.e.* autant que la taille des vecteurs d'entrée du système de conversion décrit sur la figure 4.4), 17 neurones cachés et 1 neurone de sortie. Les vecteurs de paramètres d'entrée du module de conversion spectral – renouvelés toutes les 5ms et issus d'une Analyse en Composantes Principales (ACP) des cepstres de 17 trames de 20ms centrées sur la trame courante - ont été utilisés comme vecteur d'entrée pour ce réseau. Pour l'apprentissage du réseau, le classement en voisé/non-voisé de chaque énoncé chuchoté a été obtenu en l'alignant avec l'énoncé modal correspondant. Le tableau 4.2 (c.f. 4.3.3) montre l'évaluation des performances de ce RN. Comparativement à l'erreur dans le système original, nous avons une amélioration significative de cette détection.

Estimation de F0

Pour l'estimation des valeurs de F_0, au lieu de prendre tous les segments du chuchotement, seuls les segments voisés ont été utilisés pour entraîner une mixture de Gaussiennes (GMM), ceci afin d'éviter de perdre des composantes gaussiennes pour représenter les valeurs nulles de F_0 codant les segments non-voisés. De plus, un réseau de neurones (RN) est utilisé pour prédire ces segments. Pour la synthèse, on prédit donc des valeurs de F_0 continues, hachées par le voisement calculé par ce RN. La figure 4.6 montre que la courbe de F_0 synthétisé par notre système est plus proche de la F_0 cible que celle prédite par le système original.

3.1.2 Influence de la fenêtre contextuelle sur la performance du système

Dans (Toda, Black et Tokuda, 2008), les auteurs ont prouvé que l'utilisation des vecteurs de paramètre d'entrée construits en enchaînant des trames acoustiques, pour prendre en compte les contraintes dynamiques sur les paramètres acoustiques, est efficace pour améliorer la précision du *mapping*. Dans cette thèse, la taille de la fenêtre contextuelle est augmentée en enchaînant plus de trames adjacentes. Ainsi, dans notre cas, la fenêtre de contexte est élargie en sélectionnant une trame toutes les *m* trames, et en maintenant le nombre de trames pour la concaténation constant. Une ACP ou une analyse discriminante linéaire (LDA) est alors appliquée à ce vecteur de multi-trames pour réduire sa dimension.

Les résultats montrent qu'en augmentant la taille de la fenêtre, l'estimation de la F_0 est bien meilleure car les contours intonatifs sont portés par des unités de taille supérieure au phonème, soit des syllabes ou des pieds (tableau 4.3). Par contre, la distorsion spectrale augmente quand la taille de la fenêtre contextuelle augmente (tableau 4.4). L'interprétation la plus plausible est qu'une fenêtre de taille d'un phonème contient de façon optimale des traits contextuelles nécessaires pour la conversion spectrale.

3.1.3 Contribution de l'information visuelle

Un nouveau système de conversion a été construit à partir de données audiovisuelles. La base de données comprend cette fois-ci 190 phrases du japonais prononcées par un locuteur natif, en modes chuchoté et modal, enregistrées par un capteur NAM et un microphone tête (dont 150 phrases pour le corpus d'apprentissage et 40 phrases pour le corpus de test). Le système capture, à 25 Hz, les positions 3D de 142 billes collées sur le visage (c.f. figure 3.4), en synchronie avec le signal acoustique échantillonné à 16000 Hz.

Un modèle de forme est construit à partir des positions 3D des 142 points sur le visage du locuteur. La méthodologie de clonage développée dans notre département (Bailly *et al.*, 2006) consiste en une ACP itérative appliquée sur des sous-ensembles de points pertinents. Cette analyse extrait 5 paramètres articulatoires liés au mouvement de la mâchoire, des lèvres et du larynx.

Un vecteur caractéristique audiovisuel est obtenu en combinant caractéristiques audio et visuelle comme pour les AAM (Active Appearance Models) de Cootes *et al.* (2001). Chaque vecteur visuel est multiplié par un poids w avant d'être

concaténé avec le vecteur acoustique correspondant. La dimension du vecteur conjoint est ensuite diminuée grâce à une autre ACP.

La conversion utilise les vecteurs de projection des trames sur les 40 premiers axes principaux. Ici, les nombres de gaussiennes sont fixés à 16 pour l'estimation spectrale pour l'estimation de F_0 et aussi bien pour l'estimation des composantes apériodiques.

Les tableaux 4.5 et 4.6 montrent la contribution positive de l'information visuelle sur la performance du système. Le meilleur résultat est obtenu pour une dimension du vecteur visuel de 40 et avec un poids w=0.25 pour l'estimation du voisement et w=1 pour l'estimation spectrale et de la source d'excitation. L'erreur de l'estimation du voisement diminue de 5.4% alors que la distorsion spectrale entre paroles convertie et modale diminue alors de 3.7%. La différence entre le F_0 converti et celui naturel diminue aussi de 6.9 %.

3.1.4 Evaluation perceptive

Evaluation audio

Seize auditeurs français ont participé à nos tests perceptifs sur l'intelligibilité et le naturel de la parole convertie des deux systèmes. 20 phrases, qui n'étaient pas incluses dans le corpus d'apprentissage, ont été utilisées. Chaque auditeur a passé deux tests ABX. Il a entendu une phrase prononcée dans la voix modale (X) et les versions converties à partir du chuchotement par les deux systèmes. Pour chaque phrase, l'auditeur devait choisir laquelle était la plus proche de l'originale (X), en terme d'intelligibilité et de naturel. La figure 4.12 fournit les scores moyens d'intelligibilité et de naturel obtenus pour les phrases converties utilisant les systèmes original et modifié, cumulés pour tous les auditeurs. Les scores d'intelligibilité sont significativement plus élevés pour les phrases synthétisées par notre système ($F=23.41$, $p<.001$). Ceci est aussi vrai pour le naturel : le système proposé a été encore plus fortement préféré à l'original ($F = 74.89$, $p < .001$).

Evaluation audiovisuelle

Pour confirmer la contribution positive de l''information visuelle, huit sujets japonais ont participé à nos tests perceptifs sur la parole audiovisuelle convertie. Les stimuli étaient composés des VCVs japonais (le V étant choisi parmi cinq voyelles et C parmi vingt-sept consonnes) dans quatre conditions :

- Parole produite à partir de l'acoustique chuchotée (appelée condition A à partir d'A) (1)

- Parole produite à partir du chuchotement audiovisuel (condition A à partir d'AV) (2)
- Parole et animation faciale produites à partir de l'acoustique chuchotée (condition AV à partir d'A) (3)
- Parole et animation faciale produites à partir du chuchotement audiovisuel (AV à partir d'AV) (4)

Chaque participant a entendu et a regardé aléatoirement une liste de VCV audio et audiovisuels synthétiques. Pour chaque VCV, il a été invité à choisir quelle consonne il avait entendue parmi une liste de 27 consonnes japonaises possibles.

La figure 4,13 fournit les taux moyens d'identification pour tous les participants. Les paramètres visuels améliorent de manière significative l'identification des consonnes. En particulier, les rapports d'identification sont de 28.37% pour la condition 2, de 30.56% pour la condition 3 et de 36.21% pour la condition 4 qui sont tout sensiblement plus hauts que celle de la condition 1 (23.84 %).

Bien que l'information visuelle augmente considérablement les taux d'identification sur tous les VCVs examiné, les taux sont toujours assez bas (moins de 40%). Il convient de noter cependant que l'identification de logatome VCV est une tâche difficile, particulièrement dans une langue avec beaucoup de consonnes. Il pourrait donc être argumenté que les taux d'identification sur des mots ou des phrases pourraient être beaucoup plus élevés dans une tâche de d'intelligibilité sur des mots ou des phrases, avec l'aide de l'information lexicologique, syntactique et contextuelle. Il pourrait donc être intéressant de vérifier si l'information visuelle a fourni en fait quelques traits liés au lieu d'articulation, même si la détection précise de chaque phonème était difficile. Par conséquent nous avons également groupé les consonnes dans des catégories correspondant à différent lieux d'articulation, pour examiner quels groupes consonantiques ont bénéficié le plus de l'addition d'information visuelle. La figure 4.14 montre que les consonnes bilabiales, ce qui sont intrinsèquement plus saillantes visuellement, ont été mieux identifiées comme bilabiales, quand l'information visuelle a été prise en entrée du système de conversion et quand des stimulus audiovisuels ont été présentés aux participants.

3.2 Conversion basée sur un modèle de Markov caché (HMM)

Afin de comparer la performance de la technique de conversion de voix basée sur GMM proposée par Toda et Shikano (2005) à l'approche consistant à combiner la

reconnaissance de la parole NAM et la synthèse de la parole, un système de conversion du chuchotement-vers-la parole claire basé sur HMM est étudié. Ce système combine 2 modules, un module de reconnaissance basé sur HMM et un module de synthèse basé sur HMM.

Au lieu d'utiliser la synthèse concaténative basée sur diphone proposée dans (Hueber *et al.*, 2008, 2009), nous avons employé une synthèse basée sur HMM, proposée par (Tokuda *et al.,* 2000) parce que l'intégration d'un modèle paramétrique offre quelques avantages. Tout d'abord, le système entier de synthèse est paramétré plutôt que les formes d'onde pré-enregistrées stockées. Le système est donc orientable. Il a été suggéré qu'un modèle paramétrique basé sur HMM pourrait apporter un couplage plus intime entre la reconnaissance et la synthèse, en abordant les deux problèmes dans un cadre statistique logique unifié (Tokuda *et al.*, 2004 ; Zen *et al.*, 2004, 2009 ; Zhang, 2009 ; Wu *et al.*, 2006, 2008).

Les densités de probabilité communes des paramètres du chuchotement et de la parole claire sont modélisées par des HMMs sur la taille d'un phonème. À partir de la distribution gaussienne simple à chaque état, monophone indépendant du contexte, le modèle a été enrichi de deux manières pendant l'apprentissage :

- La distribution monogaussienne à chaque 'état est remplacée par des modèles de mélange de gaussiennes (GMM). Le nombre de Gaussiennes varie de 1 à 2, à 3 et à 4.

- En raison des effets de coarticulation, il est peu probable qu'un HMM indépendant du contexte pourrait de façon optimale représenter un allophone donné. Par conséquent les HMMs dépendant du contexte sont employés pour enrichir le modèle. En raison de données d'apprentissage limitées, nous avons seulement employé le contexte de biphone pour les modèles acoustiques en groupant les phonèmes dans des classes de contexte. En ,plus d'utiliser nos connaissances phonétiques *a priori* pour définir de telles classes, des arbres de confusion (visuels et acoustiques) ont été construits en se basant sur la matrice des distances euclidiennes des paramètres visuels et acoustiques entre chaque paire de phonème.

Le tableau 5.2 montre la contribution positive de l'information visuelle pour la tâche d'identification. En moyenne, les mouvements faciaux améliorent le taux d'identification de 13,2% (61,35% à 74,52%). En particulier, dans le cas d'identification des voyelles, la précision obtenue en utilisant l'information visuelle est 78,28%, montrant une amélioration de 7% comparée avec l'information

acoustique seule. Dans le cas de l'identification de consonnes, cette amélioration est de 15,7% (56,62% à 72,35%).

Pour la synthèse, on a comparé ce système avec le système basé sur le modèle GMM, présenté dans le chapitre 4. Bien que les mouvements faciaux aient une contribution positive dans les deux systèmes, la performance du système basé sur HMM est actuellement inférieure à celle du système direct de signal-à-signal basé sur le modèle de GMM (tableau 5.4). Ce phénomène peut être expliqué par deux raisons. D'abord, la covariance diagonale actuellement utilisée pour chaque état dans le système basé sur HMM ne prend pas en compte la covariance entre les paramètres chuchotés et ceux de la parole claire, tandis que le système basé sur GMM le fait, en employant une matrice pleine de covariance. Deuxièmement, la synthèse et la reconnaissance sont employées séparément : les modèles HMM tendent à réduire au minimum l'erreur de reconnaissance, mais pas l'erreur de reconstruction finale.

4 Conclusion et perspectives

L'objectif de cette thèse est la conversion de la parole chuchotée en voix claire en employant un cadre statistique. Si cette conversion pouvait être effectuée, les téléphones cellulaires silencieux, la communication de la parole en conditions défavorables, les applications pour aider les personnes présentant des handicaps laryngés ou de nouvelles interfaces homme-machine deviendraient envisageables.

En utilisant un microphone à condensateur attaché sur le cou, au-dessous d'une oreille du locuteur pour capturer la parole chuchotée, nous proposons diverses solutions pour améliorer le naturel de la parole produite par le système de conversion de chuchotement-vers- la parole claire basé sur un modèle GMM : séparer la détection du voisement et l'estimation de F_0 ; augmenter la taille de la fenêtre de contexte du vecteur d'entrée jusqu'à la taille d'une syllabe et utiliser un analyse discriminante linéaire (LDA) au lieu d'une analyse en composantes principales (PCA) pour réduire la dimension de ce vecteur. Des tests subjectifs perceptifs ont été menés pour montrer l'amélioration apportée par ces méthodes.

Une autre solution examinée dans cette thèse pour améliorer le système est celle d'utiliser les mouvements visuels comme complément à l'information acoustique. Les mouvements orofaciaux liés aux gestes des lèvres, de la mâchoire et du larynx contribuent à l'intelligibilité de la parole convertie, surtout quand ces informations sont ajoutées en sortie du système de conversion.

Dans la lignée de ce travail de thèse, quelques pistes de recherche sont proposées pour l'avenir :

Couplage entre la reconnaissance et la synthèse

La difficulté principale pour le système de conversion chuchotement - parole claire basé sur les HMMs est d'aborder la reconnaissance de la parole et la synthèse dans un cadre unifié, car il manque un modèle statistique qui puisse être employé pour les deux problèmes. Les HMMs qui ont été utilisés avec succès pour la reconnaissance sont insatisfaisants pour la synthèse (Wu *et al.*, 2006). Dans notre système, nous employons le critère d'estimation avec maximum de vraisemblance (MLE) pour apprendre les modèles HMM. Cependant, le critère de MLE évalue seulement la pertinence du modèle aux données dans le sens des probabilités qui ne reflète pas la distance finale entre les paramètres synthétisés et les vecteurs cibles. Le critère de MLE n'est clairement pas approprié au but de la synthèse de la parole basée sur HMM. Nous pourrions utiliser le critère de la racine de la moyenne du carré(RMS) proposé par Zhang (2009) ou l'erreur de génération minimum (MGE) (Wu et Wang, 2006) pour apprendre les modèles HMM. Cela nous permet d'éliminer cette contradiction entre l'apprentissage et la génération. En outre, le problème de lissage des paramètres générés pourrait être corrigé en incorporant une contrainte de restauration de la variance globale (GV) (Toda et Tokuda, 2007; Wu *et al.*, 2008).

Elimination du bruit

Le signal capturé par le microphone NAM est bruité, ce qui influence la performance du système de conversion (cf. section 1.3.4). La réduction du bruit est évidemment une tâche importante. Afin d'augmenter le rapport signal-sur-bruit (SNR) du microphone NAM, nous pourrions tirer parti des corrélations des signaux capturés par deux microphones NAM sur le cou du locuteur ou nous pourrions utiliser d'autres corrélations multimodales, c.-à-d. la corrélation de la parole silencieuse avec les 'informations visuelles.

Données multimodales

La parole est naturellement multimodale. La recherche sur l'exploitation des signaux visuels du visage du locuteur est essentiellement concentrée sur la robustesse dans le bruit pour les systèmes de reconnaissance de la parole. L'introduction de visages ou d'avatars parlants dans les systèmes de synthèse de la parole améliore également leur naturel et intelligibilité. Généralement la prise en compte de l'aspect visuel de la parole en s'inspirant des mécanismes

biophysiologiques de production et de perception de la parole peut bénéficier au traitement automatique de la parole et aux interfaces homme-machine.

Dans ce contexte, nous avons utilisé des signaux visuels extraits à partir du visage du locuteur comme information complémentaire pour le signal acoustique capturé par le microphone NAM, dont les composantes à haute fréquence notamment sont atténuées. Bien que les résultats montrent la contribution positive de cette information pour nos systèmes de conversion, les mouvements visuels du visage (mouvements des lèvres et de la mâchoire) rendent seulement compte d'un appareil articulatoire partiel. Cette information est insuffisante. L'inclusion des mouvements des organes intérieurs de la parole (c.-à-d. ceux de la langue recueillis par EMA, ou par imagerie ultrason), l'activité laryngée (EEG, EMG) etc., est évidemment très prometteuse.

Temps réel

Le traitement en temps réel est également nécessaire afin d'utiliser le microphone NAM pour la communication silencieuse dans la vie quotidienne. Les algorithmes de mise en correspondance travaillent souvent sur des phrases isolées et utilisent des informations dépendantes du contexte. La quantité d'information contextuelle a un impact fort sur la performance. Des applications réalistes exigeront que le contexte soit tronqué à une durée maximum pour permettre la conversation. Les spécialistes de télécommunication estiment qu'un retard maximal de 200ms est toléré pour permettre une conversation en duplex (Guéguin, 2006 ; Guéguin *et al.*, 2008). Il pourra ainsi être intéressant d'étudier l'impact de la limitation d'information contextuelle sur la performance et la qualité subjective.

Appendix B

Publications

Journals

Tran, V.-A., Bailly, G., Loevenbruck, H. and Toda, T., (in press). Improvement to a NAM-captured whisper-to-speech system, Speech Communication - special issue on Silent Speech Interfaces.

Heracleous, P., Tran, V.-A., Nagai, T. and Shikano, K., 2009. Analysis and recognition of NAM speech using HMM distances and visual information. IEEE Transactions on audio, speech and language processing.

Heracleous, P., Beautemps, D., Tran, V.-A., Loevenbruck, H. and Bailly, G., 2009. Exploiting visual information for NAM recognition. IEICE Electronics Express 6(2): 77-82.

Conferences

Tran, V.-A., G. Bailly, H. Loevenbruck and T. Toda (2009). Multimodal HMM-based NAM-to-speech conversion. In Procs of Interspeech, Brighton, 656-659.

Tran, V.-A., Bailly, G., Loevenbruck, H. and Jutten, C., (2008a). Improvement to a NAM captured whisper-to-speech system. In procs of Interspeech, Brisbane, Australia, 1465-1468.

Tran, V.-A., Bailly, G., Loevenbruck, H. and Toda, T., (2008b). Predicting F0 and voicing from NAM-captured whispered speech. In Procs of Speech Prosody, Campinas, Brazil, 107-110.

Tran, V.-A., Bailly, G., Loevenbruck, H. and Jutten, C., (2008c). Amélioration de système de conversion de voix chuchoté vers la voix audible. JEP (Journées d'Etudes sur la Parole). Avignon, France.

Article 1: Speech prosody 2008

Predicting F0 and voicing from NAM-captured whispered speech

Viet-Anh Tran, Gérard Bailly*, Hélène Loevenbruck* & Tomoki Toda***

* GIPSA-Lab, Speech & Cognition dpt., 46, av. Félix Viallet, 38031 Grenoble Cedex, France
** Graduate School of Information Science, Nara Institute of Science and Technology, Japan
{viet-anh.tran,gerard.bailly,helene.loevenbruck}@gipsa-lab.inpg.fr, tomoki@is.naist.jp

Abstract

The NAM-to-speech conversion proposed by Toda and colleagues which converts Non-Audible Murmur (NAM) to audible speech by statistical mapping trained using aligned corpora is a very promising technique, but its performance is still insufficient, mainly due to the difficulty in estimating F_0 of the transformed voice from unvoiced speech. In this paper, we propose a method to improve F_0 estimation and voicing decision in a NAM-to-speech conversion system based on Gaussian Mixture Models (GMM) applied to whispered speech. Instead of combining voicing decision and F_0 estimation in a single GMM, a simple feed-forward neural network is used to detect voiced segments in the whisper while a GMM estimates a continuous melodic contour based on training voiced segments. The error rate for the voiced/unvoiced decision of the network is 6.8% compared to 9.2% with the original system. Our proposal benefits also to F_0 estimation error.

Keywords: voice conversion, F_0 estimation, neural network, non-audible murmur, whispered speech.

1. Introduction

Speech conveys a wide range of information. Among them, the linguistic content of the message being uttered is of prime importance. However, paralinguistic information such as the speaker's mood, identity or position with respect to what he/she says also plays a crucial part in oral communication [10]. Unfortunately, when a speaker murmurs or whispers, this information is degraded.

To solve this problem, Nakajima *et al.* [5] found that acoustic vibrations in the vocal tract can be captured through the soft tissues of the head with a special acoustic sensor called a NAM microphone attached to the surface of the skin, below the ear. Using this stethoscopic microphone to capture non-audible murmur, Toda *et al.* [1] proposed a NAM-to-Speech conversion system based on the GMM model in order to convert "non-audible speech" to ordinary speech. It was shown that this system effectively works but its performance is still insufficient, especially in the naturalness of the converted speech. This is due to the difficulties in F_0 estimation from unvoiced speech. These authors claimed that it is inevitable to improve the performance of NAM-to-Speech systems. Nakagiri *et al.* [4] proposed another system which converts NAM to whisper. F_0 values do not need to be estimated for converted whispered speech because whisper is another type of unvoiced speech, just like NAM, but more intelligible.

Another direction of research consists in using a phonetic pivot by combining speech recognition and synthesis techniques as in the Ouisper project [12]. By introducing higher linguistic levels, such systems can potentially predict a phonological structure that can be used in speech resynthesis. But no results have been reported yet and such an approach seems unsuitable for applications with open domain.

In this paper, we propose to improve signal-based GMM mapping by a better estimation of the voiced source of the converted speech. Whisper-to-speech was used because of difficulties in getting accurate phonetic segmentation in NAM.

In the training stage, whispered speech and ordinary speech utterance pairs were carefully aligned using phonetic transcription information. The main difference between our system and the original system proposed in [1] is that only voiced segments were provided as input to train the GMM model which maps spectral vectors of whisper to F_0 values of converted speech. In the conversion stage, we use a feed-forward neural network to detect the voiced segments in whispered utterance and then compute F_0 for these segments only instead of computing these values for all segments.

Another innovative aspect of this paper is the language: we have applied voice conversion techniques to French, which has much more complex syllabic structures and a larger phonemic inventory than Japanese. Degraded performance is thus expected with respect to original published results.

The paper is organized as follows. Section 2 describes some characteristics of whispered speech. Section 3 describes the frameworks of our NAM-to-Speech conversion system. Section 4 describes our experimental evaluations and finally, conclusions are drawn in Section 5.

2. Whispered speech

In recent years, advances in wireless communication technology have led to the widespread use of mobile phones for private communication as well as information access using speech. Speaking loudly to a mobile phone in public places may be a nuisance to others, however, whispered speech can only be heard by a limited set of listeners surrounding the speaker and can therefore effectively be used for quiet and private communication over mobile phones [7].

2.1. Acoustic features

In normal speech, voiced sounds involve a modulation of the air flow from the lungs by vibrations of the vocal folds. However, there is no vibration of the vocal folds in the production of whispered speech. Exhalation of air is used as the sound source, and the shape of the pharynx is adjusted such that the vocal folds do not vibrate. Due to this difference in production mechanism, the acoustic characteristics of

whisper differ from those of normal speech. A study on the acoustic properties of vowel sounds [7] has shown an upward shift of the formant frequencies for vowels in whispered speech compared to normal speech. The shift is larger for vowels with low formant frequencies. The authors also found that the cepstral distances between normal and whispered speech for vowels and voiced consonants are higher than those of unvoiced consonants. This means that the vocal tract characteristics of vowels and voiced consonants change more significantly in whisper relative to ordinary speech than those of unvoiced consonants.

The perception of vowel pitch in normal speech is related mainly to the fundamental frequency (F_0) which corresponds to periodic pulsing. In whispered speech, however, although there is no periodic pulsing, some pitch-like perception may occur. Higashikawa *et al.* [8] have shown that listeners can perceive pitch during whispering and formant frequency could be one of the cues used in perception. More precisely, the authors in [9] indicated that "whisper pitch" is more influenced by simultaneous changes in F1 and F2 than by changes in only one of the formants.

2.2. NAM microphone

Nakajima *et al.* [5] proposed a new communication interface which can capture acoustic vibrations in the vocal tract from a sensor placed on the skin, below the ear. This position is shown in Fig. 1. This position allows a high quality recording of various types of body transmitted speech such as normal speech and whisper. Body tissue and lip radiation act as a low-pass filter and the high frequency components are attenuated. However, the non-audible murmur spectral components still provide sufficient information to distinguish and recognize sound accurately [6]. Currently, the NAM microphone can record sound with frequency components up to 4 kHz while being little sensitive to external noise.

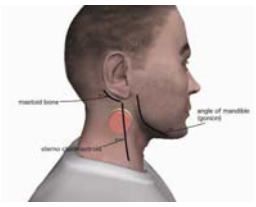


Figure 1: *Position of NAM microphone*.

3. NAM-to-Speech conversion system

Several approaches have been proposed to convert non audible speech to modal voice. The most trivial system consists in chaining NAM recognition with speech synthesis. Direct signal-to-signal mapping using aligned corpora is also very promising: Toda *et al.* [1] applied statistical feature mapping [10][11] to NAM-to-speech conversion.

Although the segmental intelligibility of synthetic signals computed by statistical feature mapping is quite acceptable, listeners have difficulty in chunking the speech continuum into meaningful words. A large part of this problem is due to impoverished synthetic intonation. In this study, we focus on improving the estimation of pitch and voicing of the converted speech. Fig. 2 shows the conversion used in our system.

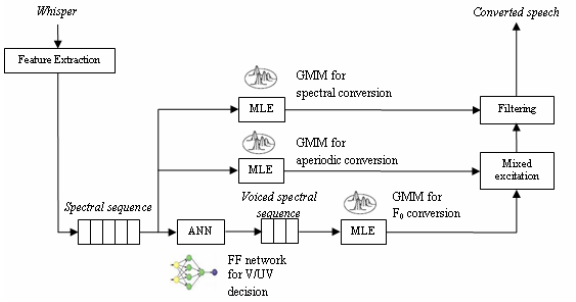


Figure 2*: Conversion process of NAM-to-Speech system. Spectral Estimation.*

Before training the models for spectral estimation and F_0 estimation, the pairs of whisper and speech uttered by a speaker must be aligned because of different speaking rates. Transcription information was used for this task to get a better alignment compared to blind dynamic time warping (DTW).

3.1. Spectral Estimation

We use the same schema for spectral estimation as the one proposed by Toda [1]. As described in [1], feature vector X_t of whisper consisting of spectral feature vectors of several frames around a current frame t was aligned with a target speech feature $Y_t = [y_t, \Delta(y_t)]$ consisting of static and dynamic features. These vectors were then used to train a GMM for representing the joint probability density $p(X_t, Y_t \mid \Theta)$, where Θ denotes a set of GMM parameters. Another model for representing the probability density of global variance (GV) of the target static features $p(v(y) \mid \Theta_v)$ was trained where Θ_v denotes a set of parameters of a Gaussian distribution, $v(y)$ denotes global variance over the time sequence y of target static feature. This global variance information is used to alleviate the over-smoothing, which is inevitable in the conventional ML-based parameter estimation [2].

In the conversion, the target static feature y was estimated from source feature $X = [X_1, X_2, \ldots, X_T]$ so that a likelihood $L = p(Y|X, \Theta)^w p(v(y) \mid \Theta_v)$ was maximize where w is a weight and the vector Y is represented as $W\,y$, where W denotes a conversion matrix from the static feature sequence to the static and dynamic feature sequence.

3.2 Excitation Estimation

The mixed excitation is defined as the frequency-dependent weighted sum of white noise and a pulse train with phase manipulation. The weight is determined based on an aperiodic component in each frequency band [14].

Aperiodic estimation was done in the same way as the spectral estimation except that global variance (GV) was not used because GV does not cause any large difference to the converted speech in the aperiodic conversion [3].

ML-based conversion method was used the F_0 estimation. Static and dynamic features Y_t of F_0 are used while keeping the same feature vector of whisper X_t as that used for the spectral conversion. However, instead of using all the segments in each pair of utterance, only voiced segments X_t, Y_t were extracted to train a GMM on the joint probability in a similar way as the spectral estimation in order to avoid loosing some Gaussian components for representing the zero values of F_0 set for unvoiced segments. A feed-forward neural network is used to predict these segments from X. For synthesis,

continuous F_0 values are predicted that are paced by the voicing parameter computed by the network.

4. Evaluation

In order to show our improvement in F_0 estimation and voicing attribution, two evaluations were done, comparing our system with the original system proposed in [1].

The training corpus consists in 200 utterance pairs of whisper and speech uttered by a French male speaker and captured by a NAM microphone and head-set microphone. Respective speech durations are 4.9 minutes for whisper (9.7 minutes with silences) and 4.8 minutes for speech (7.2 minutes with silences). The 0^{th} through 24^{th} mel-cepstral coefficients were used as a spectral feature at each frame. The spectral segment feature of whisper was constructed by concatenating feature vectors at current ± 8 frames, and then the vector dimension was reduced to 50 using a PCA technique. Log-scaled F_0 extracted by STRAIGHT [13] was used as the target feature.

4.1. Voicing estimation

In the original system [1], first, the authors took all the frames in each utterance and estimated F_0 value at each frame by using the trained GMM model. Then, they used a threshold to assign the voiced/unvoiced label for this frame. The F_0 values in every unvoiced frames were then set to zero. We applied this original technique to our data.

Then to compare with our technique, we created a feed-forward neural network with 50 input nodes, 17 hidden nodes and 1 output node. The segmental features at each frame of the whispered utterances were used as input vector for this network. The voiced/unvoiced label for each segment in the training whispered data was obtained from the voiced/ unvoiced label of the corresponding speech utterance by aligning the two utterances. All the whispered utterances used for training the GMM were also used to train this network.

Table 1: *Voicing error using neural network or GMM.*

Type of error	Feed-fwd NN (%)	GMM (%)
Voiced error	2.4	3.3
Unvoiced error	4.4	5.9
Total	~ 6.8	~ 9.2

Table 1 shows the evaluation of this network. Compared with the error in the original system, the result shows that we have a slight improvement of the voiced/unvoiced detection.

4.2. Parameter evaluation

We also compared the two systems with different number of mixtures for estimating F_0 on both the training and the test data. The number of mixtures of mel-cepstral mapping function was set to 32. Full covariance matrices were used for both GMMs. The test corpus consisted of 70 utterance pairs not included in the training data. The error was calculated as the normalized difference between synthetic F_0 and natural F_0 in the voiced segments that were well detected by the two systems. The error is given by the following formula: $Err = (synthetic\_F_0 - natural\_F_0)/ natural\_F_0$. Fig. 3 shows that the proposed framework outperforms the original system. In addition, when the number of Gaussian mixtures increase, the errors of both systems on the training data decrease, but these errors on test data are little sensitive to the number of mixtures. This is further illustrated in Table 2 which provide correlation coefficients for the two systems.

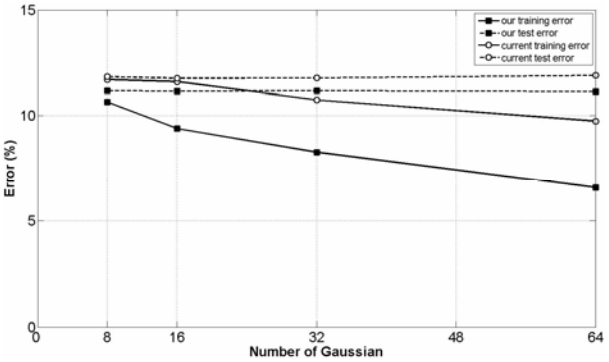


Figure 3: *Natural and synthetic F_0 curve.*

Fig. 4 shows an example of whispered-, converted- (our system) and natural-speech. As can be seen the formant patterns of converted speech are flatter than those of natural speech. Global variance was used to attenuate this difference.

Table 2: *Correlation coefficient between natural F_0 and converted F_0 by the two systems.*

# of Gaussian mixtures	Our system		Original system	
	train	test	train	test
8	0.565	0.495	0.494	0.451
16	0.667	0.493	0.513	0.456
32	0.746	0.498	0.603	0.460
64	0.834	0.499	0.682	0.444

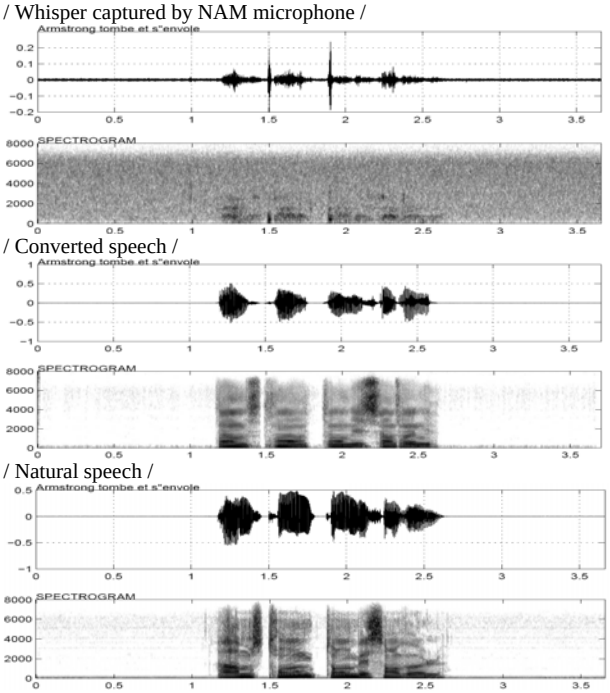


Figure 4: *Whispered speech captured by NAM sensor, converted speech and ordinary speech for the same utterance: "Armstrong tombe et s'envole".*

Figure 5 shows an example of a natural (target) F_0 curve and the synthetic F_0 curves generated by the two systems. It

shows that our new system is closer to the natural F_0 curve than the original system.

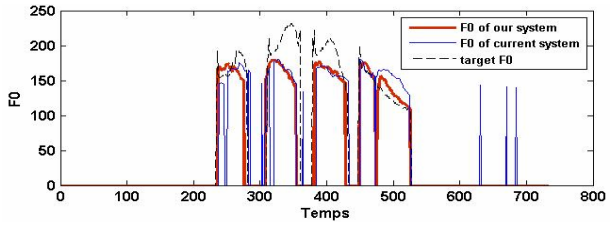
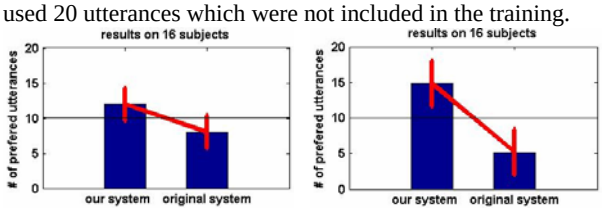


Figure 5: *Natural and synthetic F_0 curve for the same utterance : "Armstrong tombe et s'envole".*

4.3. Perceptual evaluation

Sixteen French listeners who had never listened to NAM participated in our perceptual tests on intelligibility and naturalness of the converted speech from the two systems. We used 20 utterances which were not included in the training.



(a) *(b)*

Figure 6*: Intelligibility score (a) and Naturalness score (b).*

4.3.1. Evaluation on intelligibility

For this evaluation, we used synthetic speech with the estimated mel-cepstrum, estimated aperiodic component and estimated F_0 by both systems.

Each listener heard an utterance pronounced in modal speech and the converted utterances obtained from the whispered speech with both systems. For each utterance, they were asked which one was closer to the original one, in terms of intelligibility (ABX test). Fig. 6a provides the mean intelligibility scores for all the listeners for the converted sentences using the original system and our new system. The intelligibility score is higher for the sentences obtained with our new system ($F = 23.41$, $p < .001$).

In the evaluation in [1], it was shown that F_0 estimation (compared with using a constant F_0) improves intelligibility. This might be caused by a better detection of the voiced/ unvoiced property in the F_0 estimation. In our case, a feed-forward neural network was used instead of using GMM.

4.3.2. Evaluation on naturalness

We used here the version of the test utterances produced with the modal voice. These utterances are considered as ideal desired targets of the mapping. We thus further processed the synthetic utterances by warping their time scale to the targets using the warping procedure used for training.

We then conducted an ABX test. Because of the warping, all utterances have almost the same temporal organization. For each sentence, subjects choose A or B as the nearest to the natural X in terms of naturalness. Fig. 6b shows the mean naturalness that all the listeners rate as the nearest. Again the proposed system was strongly preferred to the original one ($F = 74.89$, $p < .001$).

5. Conclusions

This paper described the improvement in F_0 estimation and voicing decision we propose for NAM-to-speech conversion system applied to whispered speech. GMM models were used to estimate the spectra, aperiodic component and F_0 of the converted speech from spectral segments obtained from NAM-captured whispered speech, based on a maximum likelihood criterion. To estimate the F_0 features in the whispered utterances, only voiced segments were used. They were detected using a simple feed-forward neural network. Although the performance of the system is improved compared to that of the original system, the estimated pitch is still flat due to the GMMs. In the future, we will investigate how to obtain audible speech from whisper by using a HMM which is appropriate for modelling a time sequence of speech parameters. Also, we plan to use complementary information such as video in the aim of obtaining other useful parameters.

Acknowledgment: The authors are grateful to Coriandre Vilain and Alain Arnal for data acquisition, Prof. Hideki Kawahara of Wakayama University in Japan for the permission to use the STRAIGHT system.

6. References

[1] Toda, T.; Shikano, K., 2005. NAM-to-Speech Conversion with Gaussian Mixture Models. In *Proc. Interspeech*. Lisboa, 1957-1960.

[2] Toda, T.; Black, A.W.; Tokuda, K., 2007. Voice Conversion Based on Maximum Likelihood Estimation of Spectral Parameter Trajectory. In *IEEE Transactions on Audio, Speech and Language Processing*. Vol. 15, No. 8, 2222-2235.

[3] Ohtani , Y.; Toda, T.; Sarawatari, H.; Shikano, K., 2006. Maximum Likelihood Voice Conversion Based on GMM with STRAIGHT Mixed Excitation. In *Proc. Interspeech – ICSLP*. Pittsburgh, USA. 2266-2269.

[4] Nakagiri, M. ; Toda, T. ; Kashioka, H.; Shikano, K., 2006. Improving Body Transmitted Unvoiced Speech with Statistical Voice Conversion. In *Proc. Interspeech-ICSLP*. Pittsburgh, USA. 2270-2273.

[5] Nakajima, Y.; Kashioka, H.; Shikano, K.; Campbell N., 2003. Non-audible murmur recognition. In *Proc. Inter-speech(Eurospeech)*. Geneva, Switzeland, 2601-2604.

[6] Heracleous, P.; Nakajima, Y., 2004. Audible (normal) speech and inaudible murmur recognition using NAM microphone. In *EUSIPCO*, Vienna, Austria.

[7] Ito, T.; Takeda, K.; Itakura, F., 2005. Analysis and recognition of whispered speech. In *Speech Communication*. Lisboa. Vol. 45, Issue 2, 139-152.

[8] Higashikawa, M.; Nakai, K.; Sakakura, A; Takahashi, H., 1996. Perceived Pitch of Whispered Vowels – Relationship with formant frequencies: A preliminary study. *Journal of Voice*, 155-158.

[9] Higashikawa, M; Minifie, F .D., 1999. Acoustical-perceptual correlates of "whispered pitch" in synthetically generated vowels. In *Journal of Speech, Language, and Hearing Research*. Vol 42, 583-591.

[10] Stylianou, Y. ; Cappé O. ; Moulines E, 1998. Continuous probabilistic transform for voice conversion. In *IEEE Trans. Speech and Audio Processing*, Vol. 6, No.2, 131-142.

[11] Kain, A.; Macon M. W., Spectral voice conversion for text-to-speech synthesis. In *Proc. ICASSP*. Seattle, U.S.A. Vol 1, 285-288.

[12] Hueber, T.; Chollet, G.; Denby, B.; Dreyfus G.; Stone M, 2007. Continuous-Speech Phone Recognition from Ultrasound and Optical Images of the Tongue and Lips. In *Proc. Interspeech*. Antwerp, Belgium,

[13] Kawahara, H.; Masuda-Katsuse, I.; Cheveigné, A., 1999. Restructuring speech representations using a pitch-adaptive time frequency smoothing and instantaneous-frequency- based F_0 extraction: Possible role of a repetitive structure in sounds. In *Speech Communication*. Vol. 27, No. 3-4, 187-207.

[14] Kawahara, H.; Estill, J.; Fujimura, O., 2001. Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT. *MAVEBA*, Firentze, Italy.

Article 2: Interspeech 2008

Improvement to a NAM captured whisper-to-speech system

Viet-Anh Tran [1], *Gérard Bailly* [1] *Hélène Lœvenbruck* [1] *Christian Jutten* [2]

1. Département Parole & Cognition, GIPSA-lab, UMR 5216 CNRS/INPG/UJF/U. Stendhal
2. Département Image & Signal, GIPSA-lab, UMR 5216 CNRS/INPG/UJF/U. Stendhal

{viet-anh.tran, gerard.bailly, helene.loevenbruck, christian.jutten}@gipsa-lab.inpg.fr

Abstract

In this paper, new techniques to improve whisper-to-speech conversion are investigated, in the framework of silent speech telephone communication. A preliminary conversion method from Non-Audible Murmur (NAM) to modal speech, based on statistical mapping trained using aligned corpora has been proposed. Although it is a very promising technique, its performance is still insufficient due to the difficulties in estimating F_0 from unvoiced speech. In this paper, two distinct modifications are proposed, in order to improve the naturalness of the synthesized speech. In the first modification, LDA (Linear Discriminant Analysis) is used instead of PCA (Principal Component Analysis) to reduce the dimensionality of the input spectral vectors. In addition, the influence of long-term variation of spectral information on pitch estimation is examined. The second modification is an attempt to integrate visual information as a complementary input to improve spectral estimation, F_0 estimation and voicing decision.

Index Terms: audiovisual voice conversion, non-audible murmur, whispered speech.

1. Introduction

Speech conveys a wide range of information. Among them, the linguistic content of the message being uttered is of prime importance. However, paralinguistic information such as the speaker's mood, identity or position with respect to what he/she says also plays a crucial part in oral communication [12]. Unfortunately, when a speaker murmurs or whispers, this information is degraded.

To solve this problem, Nakajima *et al.* [10] found that acoustic vibrations in the vocal tract can be captured through the soft tissues of the head with a special acoustic sensor called a Non-Audible Murmur (NAM) microphone attached to the surface of the skin. Using this stethoscopic microphone to capture non-audible murmur, Toda *et al.* [14] proposed a NAM-to-Speech conversion system based on the GMM model in order to convert "non-audible speech" to modal speech. It was shown that this system effectively works but the naturalness of the converted speech is still unsatisfactory. This is due to the poor F_0 estimation from unvoiced speech. These authors conclude that it is necessary to improve the performance of NAM-to-Speech systems. Nakagiri *et al.* [9] propose to simply convert NAM to whisper. F_0 values do not need to be estimated for converted whispered speech because whisper is another type of unvoiced speech, just like NAM, but more intelligible.

Another direction of research consists in using a phonetic pivot by combining speech recognition and synthesis techniques as in the Ouisper project [6]. By introducing higher linguistic levels, such systems can potentially predict a phonological structure that can be used in speech resynthesis. Although excellent results have been reported for Japanese [3], NAM recognition for languages with a richer phonological structure such as French, using spontaneous speech and open domain is unrealistic.

In this paper, we propose two different methods to improve signal-based GMM mapping from whisper to speech. Whisper is used instead of NAM because of the difficulties in getting accurate phonetic NAM segmentation. The first method consists in using Linear Discriminant Analysis (LDA) instead of using Principal Component Analysis (PCA) to obtain the spectral vector for whisper. Classes clustering different F_0 ranges and phonemes are used for LDA. Furthermore, we compare different sizes of the context window to study the influence of spectral variation on the pitch estimation performance. In the second method, visual information is integrated as a complement to the audio information. The visual parameters are obtained by the face cloning methodology developed at ICP [11].

The paper is organized as follows. Section 2 describes some characteristics of whispered speech. Section 3 briefly describes the framework of our Whisper-to-speech conversion system already explained in [15]. Modifications in this system concerned pitch estimation. Although the results of the modifications lead to satisfactory improvements, we also investigate other direction to improve the quality of the converted speech by adding other source of information. For this reason, section 4 presents our preliminary study on the promising contribution of visual information to the conversion system proposed by Toda [14]. Finally, conclusions are drawn in Section 5.

2. Whispered speech

In recent years, advances in wireless communication technology have led to the widespread use of mobile phones for private communication as well as information access using speech. Speaking loudly to a mobile phone in public places may be a nuisance to others. Whispered speech, however, can only be heard by a limited set of listeners surrounding the speaker and can therefore effectively be used for quiet and private communication [7]. However, it is hard to directly use whispered speech as a medium for human communication because of its lesser intelligibility and unfamiliar perception.

The conversion of whispered speech to modal voice is necessary for the realization of a "silent speech telephone".

2.1. Acoustic features

In normal speech, voiced sounds involve a modulation of the air flow from the lungs by vibrations of the vocal folds. However, there is no vibration of the vocal folds in the production of whispered speech. Exhalation of air is used as the sound source, and the shape of the pharynx is adjusted such that the vocal folds do not vibrate. Due to this difference in the production mechanism, the acoustic characteristics of whisper differ from those of normal speech. A study on the acoustic properties of vowels [7] has shown an upward shift of the formant frequencies for vowels in whispered speech compared to normal speech. The shift is larger for vowels with low formant frequencies. The authors also found that the cepstral distances between normal and whispered speech for vowels and voiced consonants are higher than those of unvoiced consonants: vocal tract characteristics of vowels and voiced consonants change more significantly in whisper relative to ordinary speech than those of unvoiced consonants.

The perception of vowel pitch in normal speech is mainly related to the fundamental frequency (F_0) which corresponds to periodic pulsing. In whispered speech, however, although there is no periodic pulsing, some pitch-like perception may occur. Higashikawa *et al.* [4] have shown that listeners can perceive pitch during whispering and formant frequency could be one of the cues used in perception. More precisely, the authors in [5] indicate that "whisper pitch" is more influenced by simultaneous changes in F1 and F2 than by changes in only one of the formants.

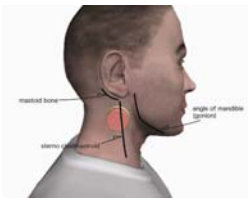


Figure 1: *Position of NAM microphone.*

2.2. NAM microphone

Nakajima *et al.* [10] proposes a new communication interface which can capture acoustic vibrations in the vocal tract from a sensor placed on the skin, below the ear (figure 1). This position offers a high quality recording of various types of body transmitted speech such as normal speech, whisper and NAM. Body tissue and lip radiation act as a low-pass filter and the high frequency components are attenuated. However, the non-audible murmur spectral components still provide sufficient information to distinguish and recognize sound accurately [3]. Currently, the NAM microphone can record sound with frequency components up to 4 kHz while being little sensitive to external noise.

3. Using LDA for whisper-to-speech

Toda *et al.* [14] proposed a NAM-to-Speech conversion system based on GMM model [12][8] in order to convert "non-audible speech" to ordinary speech. Although the segmental intelligibility of synthetic signals computed by statistical feature mapping is quite acceptable, listeners have difficulty in chunking the speech continuum into meaningful words. This is mainly due to impoverished synthetic intonation. In this study, we focus on improving the pitch estimation of the converted speech. We use the same schema as in the system we proposed in [15] except that the dimensionality of the spectral sequence is reduced by an LDA instead of a PCA. The diagram is shown in figure 2. In order to synthesize speech, we need to estimate not only spectral features but also excitation features, including F_0 and aperiodic components.

The spectral segment feature at each frame is constructed by concatenating spectral vectors for several frames around the current frame, in order to compensate for the impoverished phonetic features (especially for unvoiced fricatives losing their high frequency bands). Three GMMs are used to convert the segment features of whisper to three speech features, i.e., the spectrum, the F_0 and an aperiodic component which captures the noisy strength on each frequency band of excitation signal. Only voiced segments are used to train the model of F_0 estimation in order to avoid wasting Gaussian components for representing zero or undefined values of F_0 of unvoiced segments. These voiced segments are detected by a small feed-forward neural network. Estimated F_0 and aperiodic components are passed through a mixed excitation module before being combined with estimated spectra to compute the converted speech.

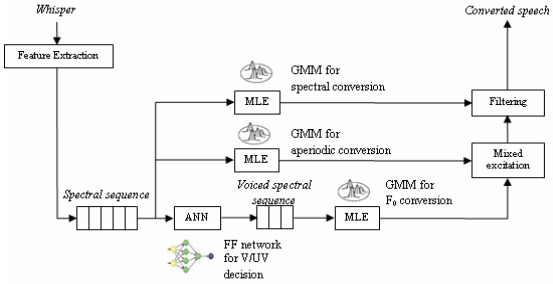


Figure 2: *Whisper-to-Speech conversion process.*

3.1. Evaluation

Two evaluations have been conducted, comparing this system with the system we proposed in [15].

The training corpus consists of 200 utterance pairs of whisper and speech uttered by a French male speaker and captured by a NAM microphone and a head-set microphone. The spectral characteristics of each frame are the 0^{th} through 24^{th} mel-cepstral coefficients. The context-dependent spectral feature of whispered frames are constructed by concatenating the spectral vectors at current ± 8 frames (context window). This vector is then reduced to 50 by LDA. We test the impact of the size of the context window by choosing one frame every 2, 3, 4 and 5 frames to combine with the current frame (windows varied from phoneme size ~100 ms to syllable size ~350 ms). Log-scaled F_0 characterize the target speech.

The test corpus consists of 70 utterance pairs not included in the training data which were uttered by the same speaker.

3.1.1. F_0 estimation

For this evaluation, the F_0 values of the target speech are classified into 13 classes: unvoiced frames are set to 0 Hz and voiced frames fall into 12 intervals, from 70Hz to 300 Hz. The class of a whispered frame is deduced from the class of the corresponding speech frame by aligning the two utterances. The number of Gaussian mixtures for F_0 estimation varies from 8 to 64 (8, 16, 32, 64). The size of the context window is also varied from the phoneme size (~100

ms) to the syllable size (~350 ms) (by picking one frame every 1-5 frames).

Table 1 shows that LDA improves the precision of pitch estimation with respect to PCA. Larger window sizes also improve the prediction. The F_0 error decreases by 16% compared to the system proposed in [15] (10.90% → 9.15%).

Figure 3 shows an example of a natural (target) F_0 curve and the synthetic F_0 curves generated by the two systems (LDA + large context window *vs.* PCA + small context window). It shows that our new system is closer to the natural F_0 curve than the old one.

Table 1. *F_0 errors (%) between converted and target speech*

method	window size (frame interval)	Number of Gaussian mixtures			
		8	16	32	64
PCA	1	10.96	10.90	10.92	**10.90**
	2	10.77	10.41	10.29	10.44
	3	10.33	9.98	10.08	10.28
	4	9.90	9.58	9.47	9.82
	5	9.44	9.17	9.32	9.31
LDA	1	10.85	10.58	10.56	10.64
	2	10.36	10.23	10.11	10.36
	3	9.98	9.94	9.93	10.29
	4	9.45	9.43	9.62	9.67
	5	**9.15**	9.22	9.25	9.37

3.1.2. Spectral estimation

We also test the influence of LDA and long-term spectral variation to the spectral estimation. This time, we use phonetic information to get the label for whispered data to train the LDA. Each whispered frame is classified in one of 34 allophones, depending on what phoneme it belongs to.

Table 2. *Spectral distortion (dB) between converted speech and target speech*

method	window size (frame interval)	Number of Gaussian mixtures	
		8	16
PCA	1	7.23	6.96
	2	7.20	7.01
	3	7.42	7.26
	4	7.25	7.55
LDA	1	6.96	6.83
	2	6.98	7.01
	3	7.03	7.17
	4	7.19	7.34

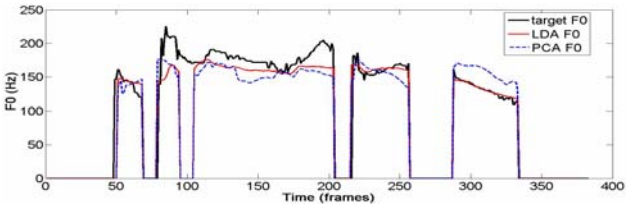


Figure 3: *Comparing natural and synthetic F_0 curves*

Table 2 shows again that LDA is slightly better than PCA. Contrary to the evaluation on the F_0 estimation, when the size of the context window increases, the spectral distortion increases. In this case, the discontinuities in the input whispered vector probably degrade the performance of the system. Another plausible interpretation is that a phoneme-sized window optimally contains necessary phonetic cues for conversion.

4. Preliminary study of audiovisual whisper-to-speech conversion

To convey a message, humans produce various linguistic sounds by controlling the configuration of oral cavities. The articulators determine the resonance characteristics of the vocal tract during speech production. Therefore, speech can be characterized not only by acoustic properties but also by articulatory properties. The articulatory parameters, which vary much slower than acoustic parameters, can effectively characterize speech [13]. Some important articulators are the lips, which significantly contribute to the intelligibility of visual speech face-to-face human interaction. In the field of man and machine communication, the visual signal corresponding to speaking lips can be helpful both in input and output modalities [1].

4.1. Audiovisual conversion system

The conversion system is built using audiovisual data. The system captures, at a sampling rate of 50 Hz, the 3D positions of 142 colored beads glued on the speaker's face (Figure 4) synchrony with the acoustic signal sampled at 16000 Hz.

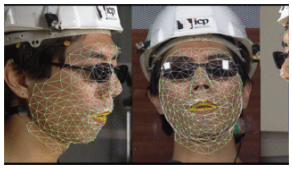


Figure 4: *Characteristic points used for capturing the movements.*

The shape model is built using a so-called guided Principal Component Analysis (PCA) where a priori knowledge is introduced during the linear decomposition. We compute and iteratively subtract predictors using carefully chosen data subsets [11]. For speech movements, this methodology extracts 5 components that are directly related to the rotation of the jaw, to lip rounding, upper and lower lip vertical movements and movements of the throat linked underlying movements of the larynx and hyoid bone. The resulting articulatory model also includes components for head movements and facial expressions but only components related to speech articulation are considered here.

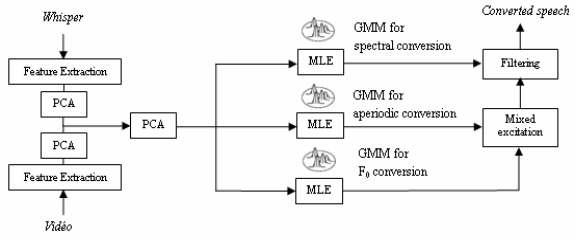


Figure 5: *Diagram of the audiovisual conversion system.*

The audiovisual feature vector is obtained by combining whispered spectral and visual feature vectors in an identical way to the AAM (Active Appearance Models) introduced by Cootes [2]: each articulatory vector is multiplied with a weight *w* before concatenation with the corresponding acoustic vector. The dimension of the joint vector is further decreased by an additional PCA (see Figure 5).

4.2. Preliminary results

The database consists of 120 sentences for training and 25 sentences for the testing, pronounced by a native Japanese speaker. The 0^{th} through 19^{th} mel-cepstral coefficients are used as spectral features at each frame. The input feature vector for computing speech spectrum is constructed by concatenating feature vectors at current ±8 frames and further reduced to a 40-dimention vector by a PCA. Similarly to the processing of the acoustic signal, each visual frame is interpolated at 200 Hz – so as to be synchronous with the audio processing – and characterized by a feature vector obtained by concatenating and projecting ±8 frames centered around the current frame on the first n principal components. The dimension of the visual vector n is set to 10, 20, 40 or 50. The weight *w* was also changed from 0.25 to 2. The conversion system uses the first 40 principal components of joint audiovisual vector. In this evaluation, the number of Gaussian was fixed at 16 for the spectral estimation, 8 for the F_0 estimation and 8 for the aperiodic components estimation.

Table 3 shows the positive contribution of visual information on the performance of the conversion. The best results are obtained with $w = 1$ and a dimension of the visual vector of 20. The spectral distortion between the converted speech and the modal speech is decreased by 2.3% while the error decreases by 16.5 % for voiced/unvoiced detection and 10.3% for F_0 estimation. With visual information only, the performance of the system is significantly degraded.

Table 3. *Contribution of visual information to (a) spectral estimation; (b) voiced/unvoiced decision; (c) F0 estimation. Best performance (bold) is obtained for a balanced contribution of audio and visual parameters.*

		Visual weight									
Distorsions	Visual dimension	**Audio-only**	0.25	0.5	0.75	**1**	1.25	1.5	1.75	2	**Video-only**
(a) Cepstral distortion in dB	10		5.656	5.632	5.608	5.576	5.595	5.634	5.619	5.654	
	20		5.676	5.630	5.596	**5.564**	5.598	5.606	5.598	5.623	
	40	**5.687**	5.676	5.630	5.596	5.611	5.568	5.605	5.596	5.618	**9.894**
	50		5.676	5.630	5.596	5.597	5.568	5.609	5.592	5.619	
(b) Voiced/ unvoiced detection (%)											
	10		13.786	13.484	12.898	12.669	20.707	20.670	20.972	21.219	
	20		13.236	13.575	12.559	**12.358**	20.734	20.276	20.569	19.534	
	40	**14.811**	13.236	13.575	12.559	12.696	20.450	20.532	20.358	20.258	**31.335**
	50		13.236	13.575	12.559	13.383	20.450	20.505	20.743	20.258	
(c) F0 estimation (%).											
	10		18.39	18.14	17.54	17.27	24.58	24.52	24.77	25.53	
	20		17.85	18.21	**17.14**	17.47	24.62	24.15	24.51	23.53	
	40	**19.48**	17.85	18.21	17.14	17.28	24.39	24.41	24.47	24.21	**36.31**
	50		17.85	18.21	17.14	17.93	24.39	24.37	24.63	24.21	

5. Conclusions

This paper describes our modifications to improve the intelligibility and the naturalness of the converted speech of the whisper-to-speech system based on GMM model. First, the use of LDA with a large context window significantly improved the converted speech, compared with using PCA with a small window. Secondly, the preliminary results on the contribution of visual information on a Japanese corpus encourage us to continue in this direction using a larger audiovisual corpus. Although the performance of the system is improved and the difference is clearly audible, the estimated pitch is still too flat due to the GMMs. In the future, we will investigate how to obtain audible speech from whisper by using a HMM which is more appropriate for modelling a time sequence of speech parameters.

6. Acknowledgements

The authors are grateful to C. Vilain, C. Savariaux, A. Arnal, K. Nakamura & T. Toda for data acquisition, to T.Toda for letting us use the NAM system, to Prof. Hideki Kawahara of Wakayama University in Japan for the permission to use the STRAIGHT system.

7. References

[1] Benoît, C., "Intelligibilité audio-visuelle de la parole, pour l'homme et la machine", Institut National Polytechnique : Grenoble, 1998.

[2] Cootes, T.F., Edwards, G.J. and Taylor C.J. "Active Appearance Models", IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(6):681-685, 2001.

[3] Heracleous, P. et al., "A tissue-conductive acoustic sensor applied in speech recognition for privacy", International Conference on Smart Objects & Ambient Intelligence, Grenoble-France, 93-98, 2005.

[4] Higashikawa, M., Nakai, K., Sakakura, A. and Takahashi, H., "Perceived Pitch of Whispered Vowels – Relationship with formant frequencies: A preliminary study", 155-158.

[5] Higashikawa, M., Minifie, F.D., "Acoustical perceptual correlates of whispered pitch in synthetically generated vowels", In Journal of Speech, Language, and Hearing Research, 42, 583-591, 1999.

[6] Hueber, T. et al., "Continuous-speech phone recognition from ultrasound and optical images of the tongue and lips", Interspeech, 2007.

[7] Ito, T., Takeda, K. and Itakura, F., "Analysis and recognition of whispered speech", Speech Communication, Lisboa, 45(2), 139-152, 2005.

[8] Kain, A. and Macon M.W., "Spectral voice conversion for text-to-speech synthesis", ICASSP, Seattle, 285-288, 1998.

[9] Nakagiri, M. et al., "Improving body transmitted unvoiced speech with statistical voice conversion", Interspeech, Pittsburgh, 2270-2273, 2006.

[10] Nakajima, Y., et al., "Non-audible murmur recognition", Eurospeech, Geneva, Switzerland, 2601-2604, 2003.

[11] Revéret, L., Bailly, G. and Badin, P., "MOTHER: a new generation of talking heads providing a flexibile articulatory control for video-realistic speech animation", International Conference on Speech and Language Processing, Beijing, China, 755-758, 2000.

[12] Stylianou, Y., Cappé, O. and Moulines, E., "Continuous probabilistic transform for voice conversion", IEEE

Transactions on Speech and Audio Processing, 6(2), 131-142, 1998.

[13] Toda, T., Black, A.W. and Tokuda, K., “Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model”, Speech Communication, 50, 215-227, 2008.

[14] Toda, T. and Shikano, K., “NAM-to-Speech conversion with gaussian mixture models”, Interspeech, Lisbon-Portugal, 1957-1960, 2005.

[15] Tran, V-A., Bailly, G., Loevenbruck, H. and Toda, T., “Predicting F_0 and voicing from NAM-captured whispered speech”, Speech Prosody, 2008

Article 3: Interspeech 2009

Multimodal HMM-based NAM-to-speech conversion

Viet-Anh Tran [1], *Gérard Bailly* [1], *Hélène Lœvenbruck* [1], *Tomoki Toda* [2]

[1] Département Parole & Cognition, GIPSA-lab, UMR 5216 CNRS/INPG/UJF/U. Stendhal
[2] Graduate School of Information Science, Nara Institute of Science and Technology, Japan
{viet-anh.tran, gerard.bailly, helene.loevenbruck}@gipsa-lab.inpg.fr, tomoki@is.naist.jp

Abstract

Although the segmental intelligibility of converted speech from silent speech using direct signal-to-signal mapping proposed by Toda *et al.* [1] is quite acceptable, listeners have sometimes difficulty in chunking the speech continuum into meaningful words due to incomplete phonetic cues provided by output signals. This paper studies another approach consisting in combining HMM-based statistical speech recognition and synthesis techniques, as well as training on aligned corpora, to convert silent speech to audible voice. By introducing phonological constraints, such systems are expected to improve the phonetic consistency of output signals. Facial movements are used in order to improve the performance of both recognition and synthesis procedures. The results show that including these movements improves the recognition rate by 6.2% and a final improvement of the spectral distortion by 2.7% is observed. The comparison between direct signal-to-signal and phonetic-based mappings is finally commented in this paper.

Index Terms: audiovisual voice conversion, non-audible murmur, whispered speech, silent speech interface, HMM-based conversion.

1. Introduction

Silent speech consists in articulating sounds with no or little vibration of the vocal cords in order to avoid being overheard [2]. Silent speech is commonly used in situations where private and confidential communication is required. However, it is hard to use it directly in telecommunication, especially with a cellular phone because of its poor intelligibility and unfamiliar perception. This problem challenges researchers with two questions: how to better capture silent speech/ articulation and how to convert it to audible voice? To cope with these challenges, several silent speech interfaces (SSI) have been proposed in the literature: motion capture of fleshpoints on the main speech articulators using Electromagnetic Articulography (EMA) sensors [3], real−time characterization of the vocal tract using ultrasound (US) and optical imaging of the tongue and lips [4][5], digital transformation of signals from a Non Audible Murmur (NAM) microphone [2][1][6][7], surface electromyography (sEMG) of the muscles or the larynx [8][9]. Together with these technologies, two main different approaches have been proposed to generate audible – and visible – speech from signatures of non audible articulation:

1. Plugging a speech synthesis system to a speech recognizer [4][5]. The generation is quite straight-forward: the recognizer segments the speech flow into phonemic units using both signal-dependent information and a more or less sophisticated language model. A standard speech synthesis system then converts this phonetic string into a synthetic voice either using the pre-recorded modal voice of the speaker or built-in available resources. The performance of such a system is mainly dependent on the recognition performance: correct recognition will result in a perfect reconstructed speech while recognition failures or inadequate language models result in drastic degradations.
2. Mapping technique based on GMM model [10] [11][1] can be used to directly convert these signals into sound using aligned corpora: joint multi-frame representations of subvocal signals and speech are either stored or modeled and then used to perform direct estimation – or inversion – of speech given the sole representation of subvocal signals. This can be seen as a quantization or optimization process that estimates the most probable speech signal given the subvocal signals and an *a priori* joint model of the combination. The overall quality of the generated speech signals is more homogenous here since the active perception of the listener may compensate for impoverished output signals. No decision is made by the mapping system concerning the phonetic content of the message. Top-down constraints driving speech intelligibility are all provided by the human perceiver.

In both cases, a remaining challenge is the generation of voicing decision and melody – speaker-specific and language-specific tones, accents and intonational patterns – that need to be estimated from non-modal phonation characterized by the absence of vocal folds vibration. Although subvocal articulation seems to still recruit motor neurons driving movements of laryngeal effectors resulting in observable EMG or small displacements of the larynx [12], this “phantom” activity has to be captured and transformed into meaningful melodic movements. So far most systems generate flat melody. Systems combining recognition and synthesis should rely either on language models or recognition of prosodic constituents to drive an intonation model. No such attempts have been reported in the literature so far. Although not completely flat, the synthetic melody

computed by voice conversion techniques has a reduced dynamics. First attempts to focus on this generation step have been performed by Tran *et al.* [6]. We notably used large windows over the subvocal signals to estimate suprasegmental features. The naturalness score is noticeably better but there is still much space left for improvement.

In this paper, we focus on the segmental intelligibility of converted speech. We first study the impact of visual information for the HMM-based speech conversion system, for both recognition and synthesis tasks. Then, this system is compared with the GMM-based system proposed by Toda *et al* [1].

The paper is organized as follows. Section 2 describes some characteristics of the NAM microphone. Section 3 describes the HMM-based whisper-to-speech conversion system, the promising contribution of visual information to this system and the comparison between the two approaches mentioned above. Finally, conclusions are drawn in Section 4.

2. Non-audible murmur microphone

Nakajima *et al.* [2] proposed a new communication interface which can capture acoustic vibrations in the vocal tract from a sensor placed on the skin, below the ear, called a NAM microphone. This microphone offers a high quality recording of various types of body transmitted speech such as normal speech, whisper and NAM. Body tissue and lip radiation act as a low-pass filter and the high frequency components are attenuated. However, the recorded spectral components still provide sufficient information to distinguish and recognize sound accurately. Currently, the NAM microphone can record sound with frequency components up to 4 kHz. Although this microphone is little sensitive to noise when using simulated noise, its performance decreases in real noise environment because of the Lombard reflex effect [13]. Figure 1 shows an example of whispered speech captured by this microphone. Note that the signal delivered by the NAM microphone is highly sensitive to bursts of stop consonants.

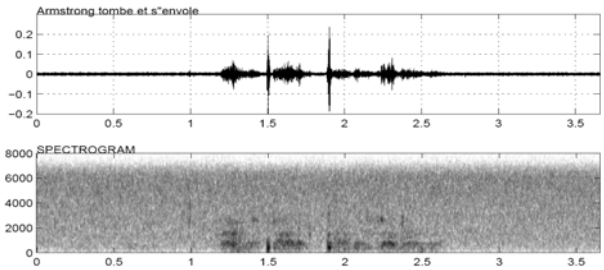


Figure 1: *Whispered speech captured by a NAM sensor for the French utterance: "Armstrong tombe et s'envole" ([amstRõg tõb e sãvol]).*

3. Audiovisual HMM-based conversion

During speech production, humans produce sounds by controlling the configuration of oral cavities. The speech articulators determine the resonance characteristics of the vocal tract. Movements of visible articulators such as the jaw and lips are known to significantly contribute to the intelligibility of speech during face-to-face communication. In the field of person-machine communication, visual information can be helpful both as input and output modalities, especially in the case of silent speech [6][7].

3.1. Audiovisual corpus

The conversion system is built using audiovisual data pronounced by a native Japanese speaker (the corpus is described in [6]). Two speech modes were recorded: whisper and normal (modal) speech. The system captures, at a sampling rate of 50 Hz, the 3D positions of 142 coloured beads glued on the speaker's face (see Figure 2) in synchrony with the acoustic signal sampled at 16000 Hz.

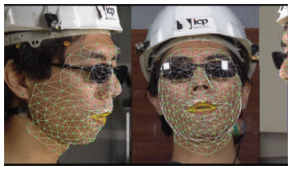


Figure 2*: Characteristic points used for capturing the movements.*

3.2. Visual parameters extraction

A shape model is built using a so-called guided Principal Component Analysis (PCA) where *a priori* knowledge is introduced during a linear decomposition. We compute and iteratively subtract predictors using carefully chosen data subsets [14], for a given speaker and a given language. For speech movements and for our particular Japanese speaker, this methodology extracts 5 components that are directly related to the rotation of the jaw, to lip rounding, to upper and lower lip vertical movements and to movements of the throat associated with underlying movements of the larynx and hyoid bone. The resulting articulatory model also includes components for head movements and facial expressions but only components related to speech articulation are considered here.

3.3. Conversion system overview

In order to compare the performance of the GMM-based voice conversion technique [1][6] with the approach of combining NAM recognition and speech synthesis, a multi-streams HMM-based whisper-to-speech conversion system was developed. It combines 2 modules, namely HMM recognition and HMM synthesis: instead of the corpus-based synthesis proposed in [5], we use HMM-based synthesis, as described in [15]. The voice conversion is performed in three steps:

1. Using aligned training utterances, the joint probability densities of source and target parameters and duration probability distribution are modeled by context-dependent phone-sized HMM. Static and dynamic acoustic and visual parameters of source and target are stored separately in 4 streams (whispered spectral stream, whispered visual stream, speech spectral stream and speech visual stream). Because of limited training data, we only used the right context for the acoustic models, where subsequent phonemes are classified coarsely into 3 groups for vowels ({/a/}, {/i/,/e/}, {/u/,/o/} without distinguishing between long and short vowels) and 7 groups for consonants: bilabials ({/p/,/pj/},{/b/,/bj/},{/m/,/mj/}), alveolars (/d/,/t/,/n/,/nj/,/s/,/ts/,/z/,/j/), palatals (/Σ/,/tΣ/,/Z/), velars ({/k/,/kj/},{/g/,/gj/}), /f/, /w/ and others

({/h/,/hj/}, {/r/,/rj/}). We add /f/ and /w/ to the context because they are visually distinguished from other consonants (see figure 3). Silences are also classified into 2 groups for utterance-final and internal silences. Gaussian mixtures with two Gaussians and diagonal covariance matrices are used to model the joint observations of each HMM state.

2. HMM-based recognition is performed using the source streams (acoustic and visual) with the HTK toolkit [16]. The linguistic model is limited to phone bi-grams learnt on the training corpus.

3. HMM-based synthesis of the recognized context-dependent phone sequence and target streams (separately acoustic and visual) is performed using the HTS software [15][17].

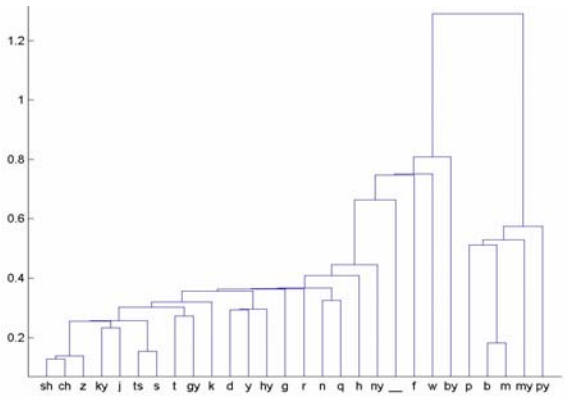


Figure 3: *Confusion tree of whispered visual movements of consonants (the smaller the ordinate, the more confused the two categories are)*

3.4. Experiments and results

The Japanese data consists in 150 utterances for training and 40 utterances for the test. The 0^{th} through 19^{th} mel-cepstral coefficients extracted by STRAIGHT [18] and their first deltas are used as spectral features while 5 visual parameters and their first deltas are used to characterize the movements of the jaw and lips, for both aligned modal speech and whisper.

3.4.1. Impact of visual information for recognition

Table 1 provides the recognition scores for all phones as well as separately for all vowels and consonants presented in the test corpus. These results show the positive contribution of visual information for the recognition task. On average, all phones considered, the input facial movements improve recognition rate by 6.2 % (65.24 % to 71.43 %). In the case of vowel recognition, the accuracy obtained by using the visual information is 76.45 %, showing an improvement of 8.7 % compared with using acoustic information only. In the case of consonant recognition, this improvement is of 7%. The lesser improvement of consonants compare to that of vowels can be attributed to the large number of labial doubles for Japanese consonants.

Table 2 shows the contribution of facial movements to the recognition of consonants considering the place of articulation. The consonants are classified into 4 groups: bilabials, alveolars, palatals and velars. The bilabials benefit from a very significant improvement (27.6%) while alveolars display only a slight improvement (4.5%). Note that facial movements also benefit surprisingly to the other consonants (17.4% improvement for velars and 14.4% degradation for palatals respectively). The small number occurrences of velars and palatals in the test corpus probably cause this phenomenon. The small facial movements cueing these phones should in fact have no significant impact on their recognition.

Table 1. *Recognition ratio for all vowels, consonants and all the phones represented in the test corpus.*

Phones	AU (%)	AUVI (%)
Vowels	67.79	76.45
Consonants	61.65	68.68
All phones	65.24	71.43

Table 2. *Recognition ratio with different places of articulation.*

Phones	AU (%)	AUVI (%)
Bilabials	**53.27**	**80.83**
Palatals	74.98	60.6
Alveolars	67.06	71.51
Velars	63.25	80.65

Table 3. *Cepstral distortion between converted speech and target speech (dB).*

System	AU	AUVI
GMM	5.99	5.77
HMM	6.58	6.4

3.4.2. Impact of visual information for synthesis

The GMM-based system that we used as a reference for this comparison is described in [1][6]. A GMM with 16 gaussians, full covariance matrix is used for the spectral estimation. Global variance is also used to reduce the over-smoothing, which is inevitable in the conventional ML-based parameter estimation [19].

Table 3 compares the contribution of visual information for the intelligibility of converted speech in terms of cepstral distortion between target speech and synthesized speech, with the two systems. Although facial movements have a positive contribution in both systems (cepstral distortion relatively decreases by 2.7% from 6.58 dB to 6.4 dB), the performance of the HMM-based system is currently inferior compared with the direct signal-to-signal system based on GMM model. This inferior score could be explained by two reasons. First, the diagonal covariance currently used for each state of the models in the HMM-based system does not take into account the covariance between whispered speech parameters and speech parameters, but the GMM-based system does, by using a full covariance matrix. Second, synthesis and recognition are used separately, therefore the trained HMM models tend to minimize the recognition error, but not the final reconstruction error.

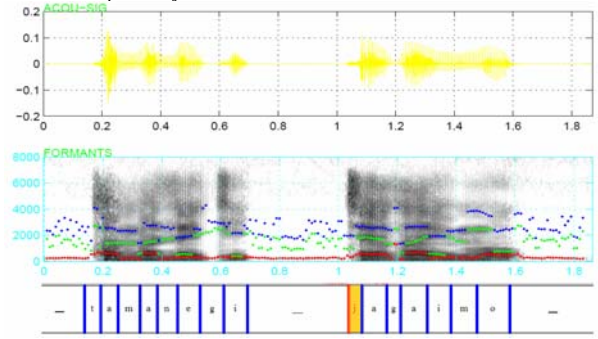
Figure 4 shows an example of converted speech by the two systems. The formant structures of the GMM-based converted speech is clearer than the other one.

4. Conclusions

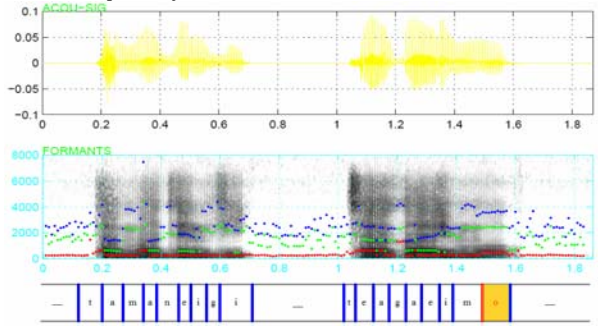
This paper describes audio-visual whisper to speech conversion that couples a speech synthesis system with a speech recognizer. The facial movements act as a compensation for lip radiation loss in the signal captured by the NAM microphone. This noticeably improves the performance of such a system, especially for the recognition task. The experimental results also show that this influence depends on place of articulation. Although the performance of such a system is currently inferior to the GMM based system, we hope that by modeling the covariance between whispered speech parameters and speech parameters, using more data, extending the acoustic models as well as the linguistic model, and by using global variance, the performance of this system will further improve.

In particular, we think that a more intimate coupling of recognition and synthesis – obtained for example by considering trajectory formation accuracy in HMM training or by considering N-best solutions in the synthesis process – should overcome the limitation of the proposed approach.

/converted speech by GMM/



/converted speech by HMM/



/target speech/

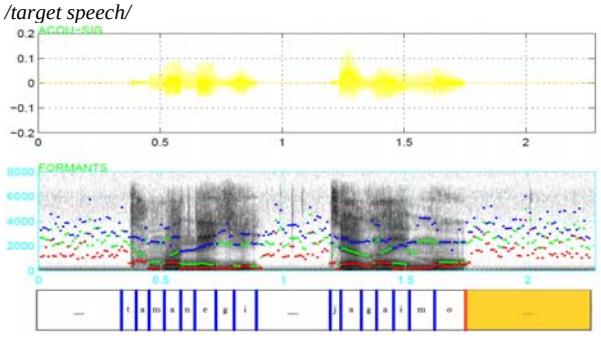


Figure 4: *Whispered speech captured by a NAM sensor for the utterance: "tamanegi jagaimo".*

5. Acknowledgements

The authors are grateful to C. Vilain, C. Savariaux, A. Arnal, K. Nakamura for data acquisition, to Prof. Hideki Kawahara of Wakayama University in Japan for the permission to use the STRAIGHT analysis-synthesis system.

6. References

[1] Toda, T. and Shikano, K., "NAM-to-Speech Conversion with Gaussian Mixture Models". InterSpeech, Lisbon - Portugal, 1957-1960, 2005.

[2] Nakajima, Y., Kashioka, H., Shikano, K. and Campbell, N. (2003). "Non-audible murmur recognition Input Interface using stethoscopic microphone attached to the skin". International Conference on Acoustics, Speech and Signal Processing, 708-711.

[3] Toda, T., Black A.W., Tokuda K., "Acoustic-to-Articulatory Inversion Mapping with Gaussian Mixture Model", ICSLP, 2004.

[4] Hueber, T., Aversano, G., Chollet, G., Denby, B., Dreyfus, G., Oussar, Y., Roussel, P. and Stone, M., "EigenTongue feature extraction for an ultrasound-based silent speech interface". IEEE International Conference on Acoustics, Speech and Signal Processing, Honolulu, Hawaii, 1245-1248, 2007.

[5] Hueber, T., Chollet, G., Denby, B., Dreyfus, G. and Stone, M., "Visual phone recognition for an ultrasound-based silent speech interface". Interspeech, Brisbane, Australia, 2008.

[6] Tran, V-A., Bailly, G., Loevenbruck, H. and Toda, T., "Improvement to a NAM captured whisper-to-speech system", Interspeech, 1465-1468, 2008.

[7] Heracleous, P., Beautemps, D., Tran, V.-A., Loevenbruck, H., Bailly, G., "Exploiting visual information for NAM recognition", IEICE Electronics Express, **6**(2): 77-82, 2009.

[8] Jorgensen, C. and Binsted, K., "Web browser control using EMG-based subvocal speech recognition". Proceedings of the 38th Annual Hawaii International Conference on System Sciences, Hawaii, 294c, 2005.

[9] Jou, S.-C., Schultz, T., Walliczek, M., Kraft, F. and Waibel, A., "Towards continuous speech recognition using surface electromyography". InterSpeech, Pittsburgh, PE, 573-576, 2006.

[10] Stylianou, Y., Cappé, O. and Moulines, E., "Continuous probabilistic transform for voice conversion", IEEE Transactions on Speech and Audio Processing, 6(2), 131-142, 1998.

[11] Kain, A. and Macon M.W., "Spectral voice conversion for text-to-speech synthesis", ICASSP, Seattle, 285-288, 1998.

[12] Coleman, J., E. Grabe and Braun, B., "Larynx movements and intonation in whispered speech". Summary of research supported by British Academy (2002).

[13] Heracleous, P., Kaino T., Saruwatari, H., Shikano, K., "Investigating the Role of the Lombard Reflex in Non-Audible Murmur (NAM) Recognition", in Proceedings of Interspeech2005 -EUROSPEECH, pp. 2649–2652, 2005.

[14] Revéret, L., Bailly, G. and Badin, P., "MOTHER: a new generation of talking heads providing a flexible articulatory control for video-realistic speech animation". International Conference on Speech and Language Processing, Beijing, China, 755-758, 2000.

[15] Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T. and Kitamura, T., "Speech parameter generation algorithms for HMM-based speech synthesis". ICASSP,Turkey, 1315–1318, 2000.

[16] Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V. and Woodland, P., "The HTK Book". Cambridge, United Kingdom, Entropic Ltd, 1999.

[17] Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A. and Tokuda, K., "The HMM-based speech synthesis system version 2.0". Speech Synthesis Workshop, Bonn, Germany, 294-299, 2007.

[18] Kawahara, H., Masuda-Katsuse, I. and Cheveigné, A. de, "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(3-4): 187-207, 1999.

[19] Toda, T. and Tokuda, K., “Speech parameter generation algorithm considering global variance for HMM-based speech synthesis”. Interspeech, Lisbon, Portugal, 2801-2804, 2005.