Contributions to speech analysis: acoustic and prosodic aspects

Jean-Luc Rouas

université de BORDEAUX

HABILITATION À DIRIGER LES RECHERCHES

Ecole doctorale Mathématique et Informatique

Spécialité Informatique

CONTRIBUTION À L'ANALYSE DE PAROLE : ASPECTS ACOUSTIQUES ET PROSODIQUES

Jean-Luc Rouas

Soutenue le 22 mars 2022

Membres du jury :

Mme. Véronique DELVAUX	Chercheuse Qualifiée FNRS	Université de Mons	Rapportrice
Mme. Corinne FREDOUILLE	Professeure des Universités	Université d'Avignon	Rapportrice
Mme. Helena MONIZ	Assistant Professor	University of Lisbon	Rapportrice
M. Frédéric BIMBOT	Directeur de Recherche CNRS	INRIA Rennes	Président
M. Cédric GENDROT	Maître de Conférences HDR	Université Sorbonne Nouvelle	Examinateur
M. Emmanuel VINCENT	Directeur de Recherche INRIA	INRIA Nancy	Examinateur

Contents

List of Figures

List of Tables

Foreword

This document presents the works I supervised during the last ten years. My contributions deal with the whole speech analysis chain: from the recordings to the final evaluation of the automatic system, with the necessary intermediate step of perception experiments for the validation of the collected data, feature extraction from the speech signal and machine learning to build the models. As such, this interdisciplinary research has benefited from the inputs of colleagues from different fields: linguistics, cognitive science, mathematics and machine learning, medical sciences.

The main topic of my research is speech analysis, with a particular focus on the expressivity than can be found in vocal signals. Dealing with expressivity is not an easy task since speech signals convey a lot of information, the most obvious being the words than are pronounced. This information is however mixed with other ones, for instance the way the same words are said can have a great impact on the meaning of a sentence. This impact can be unvolontary (i.e., when the speaker is subject to a particular emotion) or volontary (when a singer wants a express a feeling or when a speaker wants to adopt a specific attitude towards his discussion partner). Moreover, the speech signal also carries information about the speaker identity and the mental state he is in. In that way, speech analysis may also be used as a complementary clue for medical assessment (for example pathological sleepiness diagnosis or followup).

The research I carried out were developed among the following axes which are detailled in the different parts of this document:

The study of singing styles:

I supervised the PhD thesis of Leonidas Ioannidis "The Modeling of the Singing Voice and Singing Voice Quality: Indexing, Classification and Applications" from 2011 to 2015 (unachieved). During this period we used original techniques for identifying the singing styles (breathy, flow, modal, pressed) with aim of characterising the recordings made by ethnomusicologists which were made available to us during the DIADEMS project (2012-2015). This study involved the recording of a database with a professional baritone singer. This work has been published in [Ioannidis and Rouas, 2012], [Ioannidis and Rouas, 2014] and [Rouas and Ioannidis, 2016]. It is described in detail in the first part of this document.

The cross-linguistic study of social affects:

This research was initiated during the ANR JCJC (National funding for young researchers) PADE (2011-2014) and is carried out in collaboration with CLLE-ERSSàB and LIMSI. The objective is to study the vocal expressions of social affects and their perception by naive listeners. This study begun with the design of the recording protocol and I was involved in the data collection phase in France and Japan. The next step of the study consisted in the validation of the recordings with the help of perceptual tests carried out with native subjects. The first test was a performance test, where listeners were asked to evaluate the performances of all our speakers. Once the best speakers were identified, categorisation tests were carried out with other native subjects in order to validate the human capacities to categorise social affects using only speech. The selection of audio-visual excerpts for creating social affective stimuli by native and non-native speakers has been studied in [Shochi et al., 2018] and [Shochi et al., 2020]. The automatic identification

of social affect using features inspired from the ones used in the singing style characterisation is presented in [Rouas et al., 2019]. This research forms the second part of this document.

The study of deep neural networks for automatic speech recognition in French:

During the industrial PhD thesis of Florian Boyer (2017-2021), we study the possibilities offered by the different deep neural network techniques for automatic speech recognition in French, in the absence of recent studies on the French language. A peculiar point of interest was made on the choice of the intermediary unit: since end-to-end methods can use different units from the classically used phones, a choice can be made by choosing between characters, words or parts of words. Unfortunately, due to the very competitive nature of the automatic speech recognition domain and the lack of interest from the international scientific community for the French language (neither used in reference corpora, nor considered as a low resource language), we did not manage to publish our results. We however made the report [Boyer and Rouas, 2019] available on Arxiv. Technical improvements on the ESPNET project have also recently been published [Boyer et al., 2021] [Guo et al., 2021]. A summary of the results obtained by Florian is given in the third part of this document.

The study of pathological sleepiness expression in speech:

This research is carried out in collaboration with the SANPSY "Sleep, Attention and NeuroPSYchiatry" (USR3413) and the Bordeaux Hospital. I began to work on this subject in 2017 when we started to record the database with the help of a Master intern and it continues with the PhD thesis of Vincent Martin (2019-2022).

This work differs from previous studies on sleepiness and its originality relies on three main points:

- The sleepiness diagnosis is made by clinicians since we study patients followed by the sleep clinic, whereas sleepiness is often measured only by self-administered questionnaires,
- All patients are highly phenotyped, which means that in addition to the sleepiness diagnosis we also have access to numerous information related to the subjects, whether it be on physical characteristics or life habits. This allow us to study the eventual cofactors and to eliminate the parameters that are not linked to the sleepiness state.
- We record each subject multiple time since our protocol involves recording during the whole day the patients stay at the hospital (as opposed to a unique recording for which we do not have timing information). This permits to study the influence of the circadian cycle.

Thanks to our previous experience in field recordings and with the funding of an engineer and a PhD student, we have been able to record a hundred patients suffering from Excessive Daytime Sleepiness [Martin et al., 2020b]. We then studied acoustic features [Martin et al., 2019] [Martin et al., 2020c] before completing them with original

parameters derived from reading mistakes [Martin et al., 2020a] and automatic speech recognition [Martin et al., 2021]. This work in progress is addressed in part four of this document.

Part I

The Modeling of the Singing Voice and Singing Voice Quality: Indexing, Classification and Applications (PhD thesis Leonidas Ioannidis 2011-2015)

I.1 Introduction

From 2011 to 2015, I was the thesis supervisor of Leonidas Ioannidis (co-supervised by Myriam Desainte-Catherine). The work we carried out together is resumed in this chapter. Unfortunately, due to personal reasons, Leonidas did not defend his PhD.

The main contribution of this thesis is our work on automatic classification of phonation modes on singing voice. Among the ways to characterize the singing voice, one of the most salient features is the vocal quality. Since the lyrics and the partition usually have to be respected by the singer, his identity and the feelings he wants to add to his interpretation have to be expressed through modulations of the voice quality (as opposed to speech, where a speaker may also adapt his choice of words, their duration and the intonation patterns). Phonation modes are a particular expression of voice quality in singing voice.

In the first part, we will review the main characteristics of the different phonation modes. Then, we will describe the isolated vowels databases we used, with emphasis on a new database we recorded specifically for the purpose of this work. The next section will be dedicated to the description of the proposed set of parameters (acoustic and glottal) and the classification framework. The results obtained with only acoustic parameters are close to 80% of correct recognition, which seems sufficient for experimenting with continuous singing. Therefore, we set up two other experiments in order to see if the system may be of any practical use for singing voice characterisation. The first experiment aims at assessing if automatic detection of phonation modes may help classify singing into different styles. This experiment is carried out using a database of one singer singing the same song in 8 styles. The second experiment is carried out on field recordings from ethnomusicologists gathered during the DIADEMS ANR project (2011-2015) and concerns the distinction between ”normal” singing and ”laments” from a variety of countries.

I.2 Phonation modes

This work follows and extend the work of Proutskova [**?**, Proutskova et al., 2013]. It is based on the assumption that there exists four main phonation modes in singing [Sundberg, 1987]. These phonation modes are namely: breathy, neutral (or modal), flow (or resonant) and pressed.

For example, the breathy voice may be used to express sexuality or sweetness, the most common example being Marilyn Monroe singing "happy birthday". Flow phonation may be encountered in very "active" singing as for example in the "belting" technique often used by Aretha Franklin. A good example for pressed phonation would be James Brown in "I feel good".

In a more technical understanding, phonation modes result primarily from the adjustments made at the larynx level. Both Laver [Laver, 1980] on speech and Sundberg [Sundberg, 1987] on singing voice define the phonation as having three dimensions: i) the pitch, ii) the loudness and iii) the laryngeal adjustments.

The main cues that can be used for describing the four phonation modes are:

- *Pressed* phonation is associated with an elevated larynx position which also affects the vocal tract shape. There is also a stronger muscular tension around the larynx. The pressed voice is very rich in harmonic content as it favors the rise of the harmonics rather than the fundamental. Pressed-ness in voice may be perceived as a fatigue as the phonation becomes ineffective and can, under some circumstances, be part of vocal health problems. An illustration of a simulated pressed voice spectrum and glottal signal is given on Figure I.2.1.

- In *Breathy* phonation there is a relaxing of the musculature that is responsible for the adduction/abduction of the vocal folds. There is a reduced vocal fold adduction and minimal vocal fold impact stress [Garnier et al., 2007, Sundberg, 1987]. The result is a lax voice with a high level of noise from the turbulent air that passes through. The noise to Harmonic ratio is expected to be generally higher than in other phonations modes and significantly higher in the spectral region above 2KHz [Childers and Lee, 1991]. Another strong perceptual indicator of breathiness is the sensation of excessive laryngeal airflow [Grillo and Verdolini, 2008]. An illustration of a simulated breathy voice spectrum and glottal signal is given on Figure I.2.2.

- *Flow* voice is defined more as a vocal technique as it is used exclusively in singing, and unlike the other modes it requires vocal training. Sundberg suggests that flow voice is typically produced by a lowered larynx [Sundberg, 1995], while in [Titze, 2002] it is proposed that flow voice results from the condition where the vocal tract impedance is considerably smaller than the glottal impedance in an effort to optimize the mean glottal resistance or in other words the vocal output power. The produced loudness seems to be the key issue in this vocal type where the goal is to achieve higher levels of loudness with the minimum effort. In this procedure there are three elements that characterize this phonation: 1) Formant tuning, especially

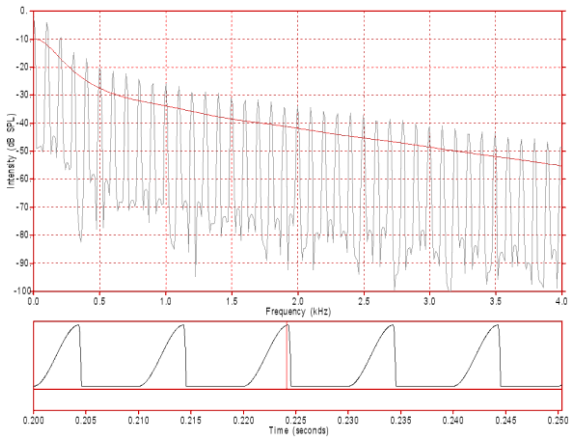


FIGURE I.2.1: The waveform (bottom) and spectrum (top) of a periodic series of synthesised glottal pulses simulating a "bright" or loud voice. Figure from www.mq.edu.au "Sound Sources in the Vocal Tract"

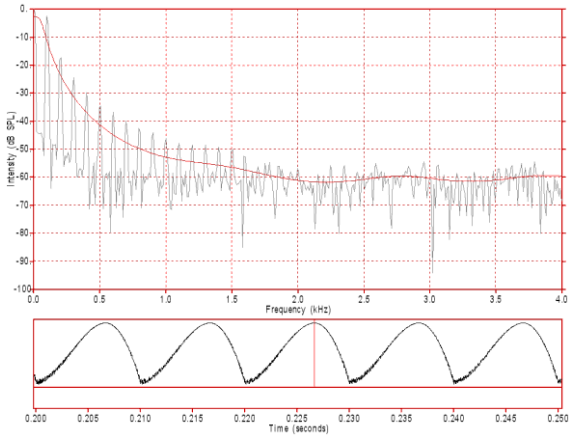


FIGURE I.2.2: The waveform (bottom) and spectrum (top) of a periodic series of synthesised glottal pulses simulating a very breathy voice. Figure from www.mq.edu.au "Sound Sources in the Vocal Tract"

the first, 2) ample harmonic content and 3) narrowing of the laryngeal vestibule [Smith et al., 2005].

- In *Modal* voice, also known as normal phonation, we find a full vibration of the vocal folds, along their entire length.

I.3 Datasets

This chapter describes the databases we used during the study: first, the two isolated vowels databases used for the study on the characterisation of phonation modes, then the two continuous singing databases we used for demonstrating the application of the characterisation of phonation modes for singing style recognition.

I.3.1 The soprano dataset

At first, we used a freely distributed database which was build specifically for the study of the phonation modes [**?**] and glottal inverse-filtering. It was recorded for the study of glottal-wave estimations and provides a public database that can be used as a reference and for comparison of existing and future methods. The authors intention was also to create a library for research on the detection of phonation modes. The lack of a reference library and training corpora in this research area was part of their motivation. This database consists of 790 short recordings of nine different vowels, found in Russian language, the singer's native language. The singer is a female soprano singer with musical training. The vowels are sung in the four phonation modes described earlier in chapter I.2. For each one of the four classes, all the vocal range of the singer was recorded wherever that was possible. The singer argues that flow and pressed voice become too unstable for the higher pitches of her vocal range, thus she discarded these samples. In Table I.3.1 the pitch-range recorded for each phonation mode are presented. The average duration of the recorded samples is 1,34 seconds, varying from 0,9 until 1,6 seconds long. The library comes with metadata including the track-number (id), pitch, vowel and phonation type of each file.

I.3.2 The baritone dataset

Working on a model that claims to successfully classify the phonation type of any given recording we came across the possibility of having a biased trained model. Having data only from a female soprano singer does not assure a good classification for other voice types. There are many different types of singers besides the soprano and they all differ

Pitch range	Samples	Mode
A3 - G5	218	Breathy
A3 - G5	196	Modal
A3 - C5	145	Pressed
A3 - G4	139	Flow

TABLE I.3.1: Table indicating the pitch range of each phonation mode in the soprano dataset.

Pitch range	Samples	Mode
A2 - G4	143	Breathy
A2 - G4	110	Modal
A2 - G4	124	Pressed
A2 - G4	108	Flow

TABLE I.3.2: Table indicating the pitch range of each phonation mode in the baritone dataset.

in terms of pitch, loudness and timbre. If we want our model to be as versatile as possible, and able to work with many types of human voices, we have to extend the training procedure to these voice types. To avoid being biased towards the high-pitched soprano voice we decided to record a male baritone singer to achieve the variability we need.

The recordings took place at SCRIME (Studio de Creation et de Recherche en Informatique et Musique Electroacoustique). The recording room is equipped with a voice booth and a mixing room.

The singer, Georgios Papaefstratiou is a professional trained baritone singer of the Choeur of the National Opera of Bordeaux. Besides his training in lyrical singing he also has experience in popular music and general music training. His singing vocal range is G2 - B4, but we recorded only the notes in the range of A2 - G4 which he claimed to be his usual working range. Before recording the database, the singer was briefed on the subject. A listening procedure followed where he listened to the recordings of the soprano database explained in 4.1 in order to be familiarized with the goals and aims of the specific project of phonation mode detection.

The recordings were made using a Neumann U87 Ai studio microphone with all the filter and attenuation switches available turned off to capture the voice with less intervention possible. The cardioid function was chosen to attenuate reflections from the recording room. The SCRIME studio is equipped with the Pro Tools mixing software, which was used for the digital recording of all the samples. The sampling rate was 44.1KHz and 24bit resolution which is enough to capture even the highest harmonics of the human singing voice. The microphone was positioned in a vertical position at the height of the singer's mouth, thus at 0° and at a 50 cm distance from it.

For this database we followed the same recording protocol held in the first dataset to keep the recording-related differences out of the picture and create a homogeneous library with two singers. Of course, we understand that this is not possible since different studio, microphone and probably other conditions where not the same. The pitch range was shorter, from A2 to G4 for all phonation types and the singer recorded only five vowels /a/, /o/, /e/, /i/, /u/ found in his native language (greek). The database consists of 487 samples of an average duration of 1.43 seconds and the distribution among the classes is seen in Table I.3.2.

This dataset is made freely available for the research community.

I.3.3 Singing styles database

The singing style dataset consists of 32 recordings which are presented in details in [Henrich and Popeil, 2003]. This database consists of recording of a unique female singer whom performs parts of the song "Amazing grace", the popular christian hymn written by the poet John Newton. The two first verses are sung several times in various singing

styles. These styles are namely: R'n'B, Belting technique, Classical, Country, Jazz, Legato, Pop, Rock. The total duration of the database is 1103 seconds with an average of 39 seconds per file.

I.3.4 the DIADEMS database

The Diadems project (Description, Indexing, Accessibility to ethnomusicological and audio documents) is a project funded by the French National Research Agency (ANR) and it aims at designing tools for the analysis and indexing of musical content adapted to the needs of ethnomusicological researchers. The main difficulties of such a database are that it consists of field recordings meaning that many things may happen in background (conversations, music, natural noises, etc.) and the recordings were made from the 1900 to nowadays using diverse hardware.

The aim of the DIADEMS project is to develop computational tools to automatically index the audio content of the CNRS - Musée de l'Homme sound archives (49,300 audio items from 5,800 collections including 28,000 items already uploaded). This ethnomusicological archive includes published and unpublished recordings of music and oral traditions from around the world, spanning a wide variety of cultural contexts. The recordings were made from the 1900 to nowadays using diverse hardware and contain a wide variety of contents (musical practice, speech, dance, ritual, interview and so on) and various settings (inside, outside, rarely in studio settings). Many sound archives in the collection include very little contextual information. The use of automatic indexation tools will help archivists to index such sound items and add new content information. It will also facilitate searches, analyses and comparison of large corpus by ethnomusicologists.

Ethnomusicologists from the Diadems project manually annotated the audio contents of a representative sample of sound items from the CNRS - Musée de l'Homme sound archives.

I.4 Features

In order to best characterise the phonation modes, we extract several audio descriptors from the audio signal. We separate these descriptors into two sub-sets, acoustic and glottal descriptors. These two sets will be evaluated separately and jointly in their ability to classify the different phonation modes. The first set consists of features extracted from the acoustic signal as recorded from the microphones and the second set consists of descriptors calculated after a glottal-wave was estimated using glottal inverse-filtering.

I.4.1 Acoustic Descriptors

A hanning window function of 25ms length has been used for all the descriptors that are presented here. A hop-size of 5ms overlapped at 1/5th of the window has been implemented. Wherever there is use of an short-term FFT (STFT) function, the FFT size is of 512 samples long, approximately 32ms, and zero-padded. One single value for each sound sample is extracted that is calculated from the mean values of all the samples in the STFT procedure. The final acoustic feature set is of dimension 24.

The different families of acoustic descriptors are:

- Harmonics Amplitude: It has been reported from many authors [Alku, 2011, **?**, Sundberg, 1987] that the difference between the two first harmonics in a source signal is an important parameter that strongly relates to specific types of phonation. Further more the third harmonic can also help determine the phonation to a lesser extent. These are calculated for the first three harmonics (H1, H2, H4).

- Formant Frequencies, Bandwidth and Amplitude: The formant frequencies are important parameters in human speech and voice signals in general. There are often used for vowel estimation and their parameters are highly tied to the vocal tract. Formants amplitudes are computed for the first three formants (A1, A2, A3), frequencies and bandwidth are calculated on the first four formants (F1, F2, F3, F4 and B1, B2, B3, B4).

- Harmonics & Formant Amplitude Differences: these features can describe the spectral shape with respect to the fundamental frequency amplitude. Amplitude differences are computed for H1-H2, H2-H4, H1-A1, H1-A2, H1-A3)

- Cepstral Peak Prominence (CPP): This method has been developed by [Hillenbrand et al., 1994] in an effort to characterize breathy voice in vocal signals for pathological speakers suffering voice disorders. Cepstral Peak Prominence (CPP) is based on periodicity measures on the cepstrum of the voice signal.

- Harmonic-to-Noise Ratio: The spectral shape is a feature that can characterize and differentiate well enough the phonation types when measured in the source signal [**?**]. HNR are computed for 4 different frequency bands: 0-0.5kHz, 0.5-1.5kHz, 1.5-2.5kHz, 2.5-3.5kHz.

I.4.2 Glottal Features

We refer to this set of features as "glottal" because they are extracted from the glottal waveforms estimated using glottal inverse-filtering.

The method used estimates vocal tract linear prediction coefficients and the glottal volume velocity waveform from a speech signal using Iterative Adaptive Inverse Filtering (IAIF) method. Analysis is carried out on a GCIsynchronous (CGI:glottal closure instant) basis and waveforms are generated using overlap and add. The method is described in [Alku, 1992]. The method is synchronized to the glottal closure instants, thus a GCI detection is needed before applying the inverse-filtering method. This method is described in [Drugman and Dutoit, 2009] and in recently published comparative review of GCI and GCO detection methods was found to be most accurate [Drugman et al., 2012].

The features extracted for the estimated glottal signal are:

- Normalized Amplitude Quotient: The normalized amplitude quotient was introduced in [Alku et al., 2002] and was presented as a time-based parametrization method for a more robust measurement of the closing phase, than the closing quotient (CQ).
- Quasi-Open Quotient (QOQ): The quasi-open-quotient is a variation of the open-quotient. The open-quotient measures the ratio of the time in which the glottis remains open during phonation. The QOQ is expected to have a big value in laxed and breathy voice types.
- H1H2: The difference between the fundamental frequency energy (H1) and its first harmonic (H2) is measured from the source signal estimated with the inverse-filtering method.
- Parabolic Spectral Parameter (PSP): Parabolic spectral parameter is a frequency domain feature devellopped by [Alku et al., 1997], for the quantification of the glottal volume velocity waveform.
- Peakslope: This feature has been proposed as an effective one for lax-tensed voice discrimination [**?**].
- Maximum dispersion quotient (MDQ): This parameter was proposed in [Kane and Gobl, 2013] for the differentiation of breathy and pressed (tense) vocal types.

I.5 Classification experiments

To assess the contribution of the feature sets for phonation mode classification, we carried out experiments on the two isolated vowels databases. Three machine learning methods were used: Logistic Model Trees (LMT), Support Vector Machines (SVM) and an instance-based classifier the KStar algorithm. The experiments are all done using the weka software [Hall et al., 2009] and libSVM [Chang and Lin, 2011].

For all the evaluation results a non-overlapping 10-fold cross-validation scheme was held for the training-testing procedure. This means that the whole dataset is divided randomly into 10 subsets. The same experiment is held 10 times, at each iteration 9 sets are used for the training and the last set is used for testing. The results are given when the 10 iterations are processed.

I.5.1 General results

In our first experiment, we merged the two datasets, resulting in a total of 1135 samples. The summary of accuracy results is displayed in Table I.5.1. The best result for the mixed feature set is obtained with the instance-based K-star classifier. Using this classifier, an accuracy of 78.59% was achieved over the whole dataset. We thus decided in the following experiments to report only the results from the K-star classifier.

I.5.2 Results on the Soprano and Baritone datasets

Table I.5.2 gives the accuracy results for the K-star classifier separately for the Soprano and Baritone datasets and for both acoustic and glottal feature sets.

The authors of the soprano database proposed in their work [Proutskova et al., 2013] a vowel-dependent scheme for the classification of the phonation. As a comparison, they reported an average accuracy of 65%, with the results varying from 52 to 75% according to the vowel. Our results are slightly better, with an overall classification rate of 81% correct on that part of the database.

The performances we obtained using only the baritone dataset are better than with the soprano dataset, the main errors for the soprano dataset are for quite high pitched

Feature set/Classifier	SVM	LMT	KStar	Average
Acoustic+Glottal	72.60	77.97	78.59	76.39
Acoustic	69.97	76.91	77.96	74.95
Glottal	52.24	60.69	59.17	57.37

TABLE I.5.1: Summary of the accuracy results from the three different classifiers. Comparison of the three features set used. All the data is used for these experiments. Results are in percentage of the true-positive predictions (%)

samples (above C4) which may be linked to the fact that using different phonation modes can become difficult when the pitch increases (see detailed analysis below).

When trying to compare the two feature sets for their ability to well characterise the four phonation modes we can see that the acoustic set gives the best performance. Adding the glottal set to the acoustic one does not bring a significant change to the overall classification rate (significance test carried out using T-test). We thus decided in the following to report only results using the K-star classifier and the acoustic feature set.

TABLE I.5.2: Summary of the accuracy results for the different feature sets. Results are in % of correct classifications.

	Acoustic	Glottal	Acoustic+Glottal
Soprano	79.74	64.90	81.62
Baritone	89.96	57.31	88.51
All	77.96	59.17	78.59

I.5.3 Confusion matrix

Accordingly, the confusion matrix is given for the acoustic feature set and K-star classifier in table I.5.3. We observe that breathy and modal phonation are best identified phonation although they are sometime confused. Flow and pressed voices are well distinguished from the breathy and modal but there is some confusion between them. In [Proutskova et al., 2013], it is also reported that the main confusions are between breathy + modal and flow + pressed, a finding that was reproduced in our experiments. This experiment shows that a system can be designed to classify phonation modes using only acoustic information.

The next section details the errors made by our automatic classification system.

TABLE I.5.3: Confusion matrix for the classification experiment over the whole dataset and the acoustic features set. Results are expressed in %

	Breathy	Modal	Flow	Pressed
Breathy	88.49	10.61	0.29	0.58
Modal	12.02	74.57	6.87	6.52
Flow	0	4.16	75.75	20.07
Pressed	0.81	2.85	26.53	69.79

I.5.4 Error analysis

First of all, we analyze the errors obtained according to their pitch and present them in Figure I.5.1, according to the octave they belong. We must take into consideration that the distribution among the octaves is not equal. Only the third and fourth octaves have an equilibrated ratio of 338 and 349 samples respectively. The second and fifth octave have 136 and 76 samples each.

We see that most of the errors are located in the fourth octave, meaning that high pitched voiced sounds are more difficult to classify. Due to the bad distribution among the octaves we cannot make any safe conclusions for the lower second octave, where all 136 samples were produced by the baritone singer.

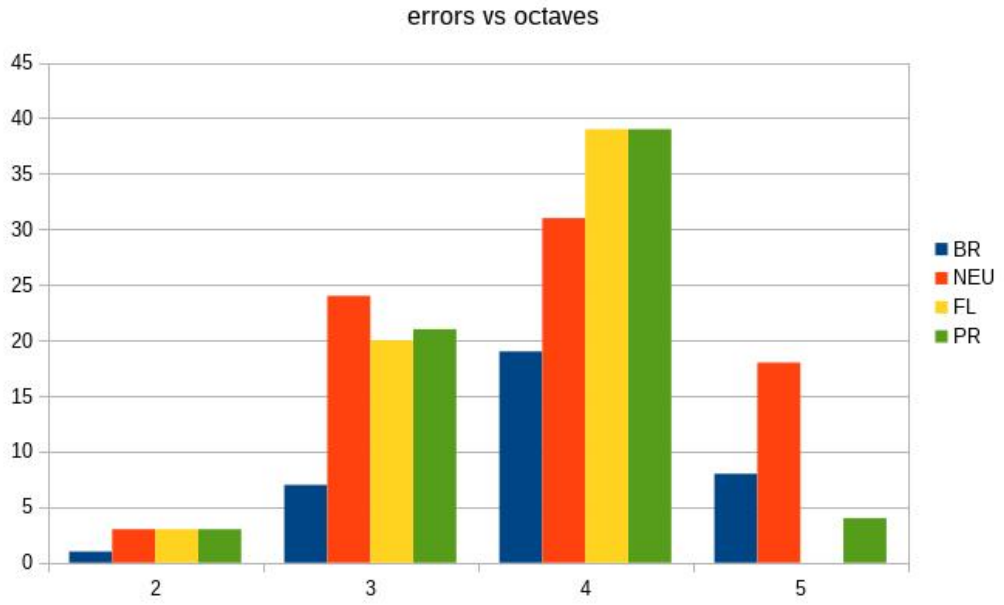


FIGURE I.5.1: Classification errors according pitch (octave) and phonation mode. BR: breathy, MOD: modal, FL: flow, PR: pressed

The information of the vowel sung can be utilized for the error analysis as well. On Figure I.5.2, the errors are displayed in a way where on the x-axis we find all the vowel present in the dataset and on the y-axis the amount of errors. Each class can be observed separately.

At first glance, we can see that most of the errors concern the modal, flow and pressed phonation and the vowel where more problem are encountered are the /o/, /oe/, and /u/. Hence, back and closed vowels are the ones that seem more difficult to detect. Back vowels are the ones produced with the tongue positioned in the back of the mouth and closed vowels are the ones where the tongue is positioned high close to the palate of the mouth cavity in addition to the opening of the mouth.

We find also a high error rate for the /e/ and /i/ vowels, more intensely for the modal phonation. They are also closed but also front vowels.

Although intelligibility of vowels is secondary to the intonation in singing voice signals, however the fact that primarily closed and secondary back vowels are not easily classified can be related to the vocal effort and vowel intelligibility. High vocal effort can be found in pressed and modal phonation modes, especially at high pitches, which can affect vowel perception [Garnier et al., 2008]. The value of the first formant frequency (F1) increases with the rise of the vocal effort along with a rise of the fundamental frequency (F0) for all closed vowels, while there is a lowering of the second formant (F2) for the front-closed and a rise for the back-closed vowels. These adjustments can affect vowel intelligibility. This should not concern 'lax' voice styles like breathy voice found here. It seems that high vocal effort and high pitch can bring similar confusion to the phonation perception. As seen in figure I.5.2 highest error rates for both pressed and modal are found in vowels /e/ (17), /i/ (15), /o/ (23), /oe/ (16) and /u/ (21).

In the next section, an experiment is carried out to assess whether this system may be useful for singing style characterisation.

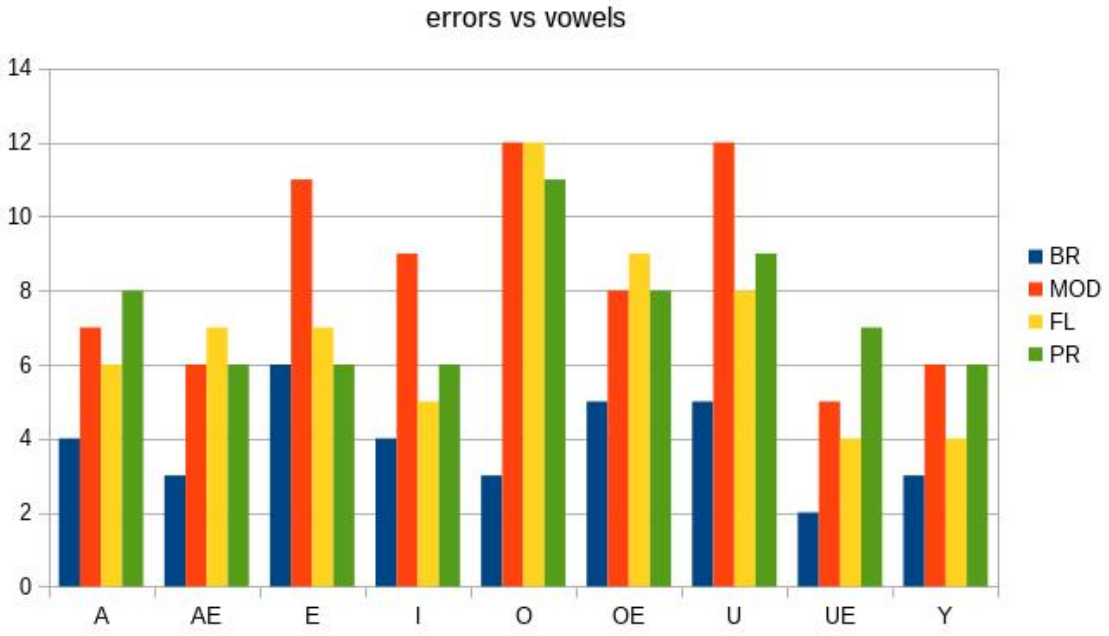


FIGURE I.5.2: Classification errors according to the vowel for the different phonation modes. BR: breathy, MOD: modal, FL: flow, PR: pressed

I.6 Towards singing style classification

The goal behind this experiment is to investigated if we could classify singing styles according to the detection of phonation modes. As training corpus we use the combined dataset of soprano and baritone (the whole database used for training). Only the acoustic features set is used. In order to form our test dataset we use the singing styles dataset (section I.3.3). The dataset is annotated according to the singing style by the singer itself and these classes are used as the ground truth.

I.6.1 Creation of the study corpus: Extraction of the voiced regions

Since our phonation mode detection system is trained on isolated vowels, we need to pre-processed the singing styles database in on order to keep only the vocalic parts.

The extraction of voiced regions is based on a forced alignment method at the phoneme level. We used the VoxForge corpus of read English for training the models. In total there are 22.897 short recordings of speech each from a variety of speaker profiles. It includes, besides the audio files, prompt files containing the text transcription and an info file containing information about the speaker age, gender and the dialect.

All the processing is done using the HTK software [Young et al., 2006]. The parameters used to train the model are MFCC plus the first and second derivatives (MFCC + MFCCD + MFCCA). The extraction of the features is carried out using 25ms hamming window with 10ms overlapping rate. The filterbank analysis made for the MFCC extraction is done in 26 channels on the mel-scale. The number of cepstral coefficients is 12 plus the zero coefficient, which finally with the derivatives gives a 39 dimensional feature space. The target phonemes are the 44 found in the English language.

The training is done using a single-state per phoneme scheme and 32 gaussians mixture models at each state. This model is then used to perform the forced alignment of all the audio files of the corpus according to the lyrics. Finally, we apply a filter to exclude the samples that are less than 0.5 seconds to ensure that there is a stable phonation. An short example of the phoneme transcription can be seen in figure I.6.1.

I.6.2 Classification

The results for the classification task are displayed in table I.6.1. We can observe that according to the singing style different phonation modes have been preferred. We can first remark that breathy voice although being the phonation that was better recognized in the above experiments has very low rates in all singing styles, probably because of its ornamental use rather than a singing style that is used constantly by the singers.

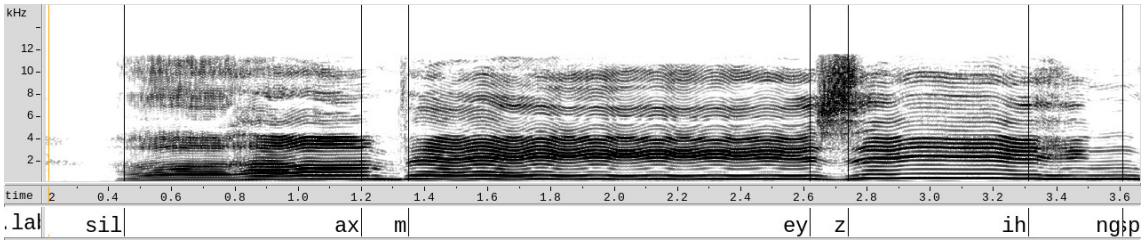


FIGURE I.6.1: Spectrogram of the beginning from a 'legato' performance of "Amazing Grace". On the lower pane the result of the forced alignment

Style	Breathy	Modal	Flow	Pressed
R&B	2.90	**42.81**	3.55	**50.72**
Belting	1.07	1.95	**63.40**	33.56
Rock	0	3.28	**76.85**	21.79
Jazz	7.54	**78.22**	2.82	11.40
Pop	**18.21**	**50.29**	**12.43**	**19.06**
Legato	4.02	**73.65**	19.09	3.21
Country	15.55	**58.41**	0	**26.03**
Classical	5.52	**93.51**	0	0.96

TABLE I.6.1: Classification of phonation mode according to singing style - results are expressed in % of total time

For R'n'B style we find that the singer chooses modal and pressed phonation equally. R'n'B music is characterized by high vocal effort so pressed phonation is not surprisingly high in this case. Belting and rock style shows a high rate of flow phonation which can be thought reasonable since flow and belting voice are considered similar voicing techniques.

Belting voice is a technique of loud singing with a voice quality similar to a yell. In the work published by [Henrich Bernardoni, 2006] we find that it is considered as an extension of chest register into the upper part. Belting sounds are produced with a high subglottal pressure which results in a long closed-phase in the glottal flow signal and high loudness levels. A tuning of the first formant with the second harmonic has been observed on open vowels. Characteristics such as long closed phase and tuning of the first formant resulting in a loud voice technique are similar to the flow voice.

In jazz performances we observe a high rate of modal voice and a relatively high breathy rate, when compared to the rest of the singing styles. Pop performances are the ones where the biggest variety in the terms of phonation modes is observed, with modal being the most detected one.

In legato and classical styles modal voice is chosen more often while in country style although modal phonation has the most instances there is a significant percentage of pressed and breathy.

However interesting these results are, they are obtained on a small database with a unique singer. Further experiments need to be carried out to validate this approach, using recordings with isolated voices from different singers in different singing styles.

I.7 Preliminary experiments on ethnomusicological recordings

One of the challenges of the DIADEMS project is to study if some distinctions can be made automatically between different "levels" of singing (such as lament, chanting or singing) and different speaking styles (storytelling, recitation, talking). As a preliminary study, we decided to use the system to classify excerpts from the lament and singing classes from the DIADEMS database described in section I.3.4. Lament is defined by the presence of several of the four common icons of crying (the cry break, the voice inhalation, the creaky voice and the falsetto vowel) proposed by [Urban, 1988]. These two categories of singing were found to be quite difficult to discriminate while they could be separated from the other classes [Feugère et al., 2015].

Even if usual acoustic parameters do not help to make the distinctions between these two categories, we assume that since laments include extreme cases of phonations, there would be a lot more instances of breathy and pressed phonations, while in "normal" singing the distribution between the four phonation modes should be more balanced.

In that experiment, we considered 10 singing segments and 6 lament segments from different countries (Albania, Turkey, Paraguay, Ethopia, Azerbaîdjan, Vietnam, Lebanon, Mexico). The singing excerpts lasted for a total of 14 minutes and the lament excerpts 7 minutes.

The proportions of the different phonation modes found are represented in figure I.7.1.

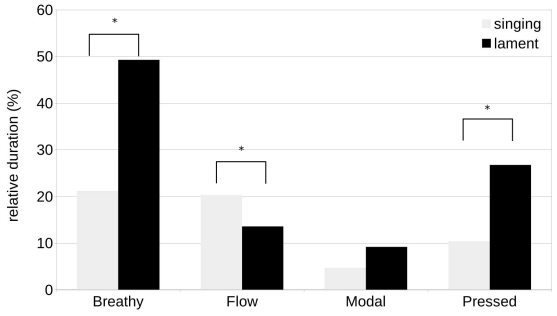


FIGURE I.7.1: Proportions of phonation modes in the DIADEMS excerpts (in % of duration). * indicates significant differences

The results seems to confirm our hypothesis since the number of "extreme" cases of phonation is much more important in lament than in singing, the proportion of breathy and pressed phonation accounting to more than 75% of the total time. Differences between the distribution of phonation modes are highly significant ($p<0.01$) except for the modal phonation were there is no significant difference between the lament and singing voice.

However, this result is to be tempered by the fact that some singing styles may also involve a good proportion of breathy and pressed phonation. Considering the classification of individual recordings may thus prove ineffective using only these cues are require further research.

I.8 Overview

Unfortunately Leonidas did not manage to finish the last modifications asked by the reviewers. It can be notred that the baritone phonation modes database has since been used in several projects from other researchers across the world:

- Kadiri, S. R., & Yegnanarayana, B. (2018). Analysis and Detection of Phonation Modes in Singing Voice using Excitation Source Features and Single Frequency Filtering Cepstral Coefficients (SFFCC). In Interspeech (pp. 441-445).
- Kadiri, S. R., & Alku, P. (2019). Mel-frequency cepstral coefficients derived using the zero-time windowing spectrum for classification of phonation types in singing. The Journal of the Acoustical Society of America, 146(5), EL418-EL423.
- Kadiri, S. R., Alku, P., & Yegnanarayana, B. (2020). Analysis and classification of phonation types in speech and singing voice. Speech Communication, 118, 33-47.
- Proutskova, P. (2019). Investigating the Singing Voice: Quantitative and Qualitative Approaches to Studying Cross-Cultural Vocal Production (Doctoral dissertation, Goldsmiths, University of London).
- Sun, X., Jiang, Y., & Li, W. (2020, July). Residual attention based network for automatic classification of phonation modes. In 2020 IEEE International Conference on Multimedia and Expo (ICME) (pp. 1-6). IEEE.

Moreover, as seen in the next parts of this document, the features we used for the phonation mode classification have been reused for my other works on the speech analysis.

Part II

Social affective signals: the banana study

II.1 Introduction

The prosodic expressions of social affects, or attitudes, plays an important role in face-to-face interaction. They are used by speakers to express their intention to their communication partner. Such choices are partly linked with the speaker's own proficiency in the spoken language, her/his personality, gender and the communication context which are also constrained at the linguistic level. Thus, each language has specific formulae or conventional prosodic variations for specific interaction contexts. Usually, studies investigating such kinds of prosodic variations rely on stereotypic stimuli [de Moraes, 2008, Fujisaki and Hirose, 1993, Gu et al., 2011, Morlec et al., 2001]. One common difficulty in studies which aim to compare the prosody of social affects is linked to a trade-off between the high sound quality required for acoustic analysis, the need for a neutral lexical content of the studied sentences (ideally identical sentences for all the studied affects), the search for spontaneity of the expressions, and a clear labeling of the communicative goals of the speaker. Most of the cited studies use laboratory corpora. Typically, adhoc sentences are recorded by speakers trying to read a sentence and reproduce a given expressivity. To enhance the spontaneity of these expressions and to facilitate the speaker's task, [Grichkovtsova et al., 2012] proposes to place target sentences in affectively loaded texts. Similarly, [Gu et al., 2011] recorded attitudinally-neutral sentences embedded into dialogues that prepare the speaker to perform an adequate expression for the target sentence. The approach used during this research builds on these works. In order to study the expressive strategies used by speakers of varying linguistic backgrounds, communicative situations have been set-up so they can be plausibly used in different languages.

II.2 Social affects database

II.2.1 Recording procedure

In order to immerse subjects in the context, a scenario was set up for each attitude, and the subject was requested to engage in a short dialogue that would lead to the production of target sentences with the native speaker. For the current experiment, 16 contexts have been selected, corresponding to a set of attitudes used in [Rilliard et al., 2009, Shochi et al., 2009] for different languages. Some of these contexts do not have lexical equivalents in all languages, as the corresponding communication situations have not been conventionalised in that particular culture. It is the case for example of the Japanese notion of *kyoshuku*, described by [Sadanobu, 2004] as "corresponding to a mixture of suffering ashamedness and embarrassment, which comes from the speaker consciousness of the fact his/her utterance of request imposes a burden to the hearer". For instance, *Kyoshuku* has no lexical equivalent in English. Meanwhile, "walking on eggs" corresponds to a certain extent to this concept.

II.2.2 List of social affects

The following 16 social affects were used in the present corpus: Admiration (ADMI), Arrogance (ARRO), Authority (AUTH), Contempt (CONT), Doubt (DOUB), Irony (IRON), Irritation (IRRI), Neutral declarative sentence (DECL), Neutral question (QUES), Obviousness (OBVI), Politeness (POLI), Seduction (SEDU), Sincerity (SINC), Surprise (SURP), Uncertainty (UNCE), Walking on eggs (WOEG). They are defined by prototypical situations with the social relationship of the two interlocutors specified – as well as the communicative goal of the speaker (see [Rilliard et al., 2013] for details). For all situations, a short neutral target sentence has been used to record the respective prosodic expressions: "A banana". In order to elicit these target sentences in each context, small dialogues were written (cf. [Gu et al., 2011]), that take place in the prototypical context described above, and that end with the target sentence.

During the recordings, each speaker (A) has an active interlocutor (B) who interacts with her/him in order to enhance the naturalness of the communication situation, and to ease the production of realistic expressions. Speakers are indeed not asked to produce an isolated sentence with an identified attitude (e.g. seduction or authority), but rather to immerse in a scenario.

- Admiration (ADMI): A & B are almost the same age and know each other well. Both love French cuisine, and talk about the very delicious food they ate yesterday at a famous French restaurant. The scene is at a coffee shop.
- Arrogance (ARRO) : both A & B are from the same university, but A is older and A's father is head of the university and A is a bit of a snob. Both know each other, but are not friends. A organized a social party, and B was not invited to the party,

but A is aware of his/her presence during the party. The scene is a party room, and A says to B that only his friends are invited.

- Authority (AUTH): Speaker A is a custom agent; speaker B is a traveler. B is in front of A, requesting permission to enter the country; A needs to impose his authority; the scene is at a custom counter at the airport.
- Contempt (CONT): both A & B are from the same university, but A is older; both know each other, but are not friends. In fact, A really hates B. A organized a social party, and speaker B was not invited, but A is aware of his/her presence. The scene is at a party room
- Doubt (DOUB): A & B are colleagues, same age. A knows that his colleague B didn't go to the baseball game yesterday, but B pretends he went to the game, and A doesn't believe it. The scene is at a coffee shop.
- Irony (IRON): A & B are friends, same age; A is going to Boston to see an important baseball game, and B, who is living in Boston calls A. Unfortunately, the weather in Boston is rainy and B says its wonderful; the scene is at an airport.
- Irritation (IRRI): A & B are almost the same age and know each other. A is sitting next to B. Suddenly, B starts to smoke, and A is very angry; he wants him/her to stop, expressing his irritation toward speaker B. The scene is a public place.
- Neutral declarative sentence (DECL): A & B are colleagues, same age; A giv es information without any personal perspective; the scene is at a coffee shop.
- Neutral question (QUES): A & B are colleagues, same age. A asks for information, without any personal perspective, awaiting a simple answer. The scene is at a coffee shop.
- Obviousness (OBVI): A & B are colleagues, same age; everyone knows B doesn't like French movies, but A asks B if he likes French movies or not; the scene is at a coffee shop.
- Politeness (POLI): A & B are almost the same age and don't know each other well, but work together professionally. A is sitting next to B; both start social talk. The scene is at a formal party.
- Seduction (SEDU): A loves B and they have an intimate relationship. A gives a compliment to B in a sexually provocative way. The scene is at a club house.
- Sincerity (SINC): B is chief of the section which A belongs to; B is older than A. The chief (B) wants A to take on a big project; A is pleased to be asked to do this, and expresses his enthusiasm, honesty and sincerity for this task. The scene is at B's office.
- Surprise (SURP): A & B are friends, same age. A didn't know that B can sing well. One day, B makes A listen to his beautiful voice. The scene is at friend's home.
- Uncertainty (UNCE): A & B are colleagues, same age. A saw B at the baseball game yesterday, but is not 100sure if it was really B; the scene is at a coffee shop.

- "Walking on eggs" (WOEG): B is chief of the section which A belongs to; B is older than A. The chief (B) wants A to do a task which is a lot of work, and it seems to A it is impossible to do this, so A tries to reject this request by trying to make sure her/his boss (B) doesn't get angry for refusing. The scene is at B's office.

Currently, these situations have been adapted to three languages: American English, Japanese and French.

Here I will present only the data and the results we obtained with the japanese speakers.

II.2.3 Speakers

A set of 19 Japanese native speakers (11 females, 8 males) have been recorded. Most speakers were recruited amongst university students and were paid for their performance. The recordings took place in a sound-treated room at Waseda University, Japan. The sound was captured by an *Earthworks QTC1* omnidirectional microphone, placed at one meter from the mouth of the speaker (this distance was chosen to limit the influence of the speaker movements on the sound level). The microphone level was calibrated before each recording session using a *Bruel & Kjaer* acoustical calibrator, thus the sound pressure level can be corrected after recording to a level comparable across all speakers. The target sentence "banana" was then manually searched for across the recorded corpus, isolated and extracted into individual files. Any speech utterances from speaker *B* occurring during the expressive gesture of speaker *A* performing the target sentence were removed from the sound track (none overlapped with their speech). Due to the interactive nature of the recording, some spontaneous changes were observed on the target sentences: typically "banana" sentence with interjections, such as "hmm", "er", "oh", etc., together with the target sentences. Each speaker recorded one utterance of the word for each of the 16 attitudes, resulting in a total of 304 stimuli. These were stored as 16 kHz, 16-bit WAV files. Each stimulus was trimmed to discard the beginning and the ending silence. The wave file of each stimulus was hand-labeled at a phonetic level using the PRAAT software [Boersma and Weenink, 2012].

II.3 Perceptual experiments: Japanese

In order to validate the data we collected during the recording sessions, we devised two perceptual experiments:

- The first one is a performance test, aiming at evaluating the performances of each speaker. This experiment aim is to validate (or not) the performances of each speaker and is further used to select the best performing speakers for the next experiment.
- The second experiment uses the two best performing speakers and is a categorical test. Listeners are asked to determine which social affect is used in a sentence. The aim here is to evaluate the human capabilities of social affect recognition from speech.

These two test are detailed below.

II.3.1 Performance test

An evaluation test was designed in order to evaluate how well each speaker expresses each social affect. This test is carried out with the same protocol as the one used in [Rilliard et al., 2013]. 26 subjects (12 females, 14 males, mean age 31), all Japanese native speakers from the Tokyo area listened to 608 stimuli (19 speakers performing 16 attitudes with two sentences) and evaluated the speakers performance on a scale from 1 to 9 (from "very bad performance" to "very good performance"). In order to make sure that all subjects agree on the definition of each affect, a sheet with a description of the target affects was provided to the subjects before the experiment. Stimuli are audio-visually presented in a random order. Before playing each stimulus, the name of the target social affect is first presented during 2.5 seconds. Then, the subject has 10 seconds to rate the performance using the keyboard. The next stimulus is automatically played after 10 seconds or after the answer of the subject if the answer is given before the 10 seconds.

An ANOVA analysis was performed on the results using the SPSS software. The factors for the statistical analysis are: the Social affect (Affect, 16 levels), the Sentence (2 levels) and the Speaker (19 levels). Results of the ANOVA are presented in table II.3.1. All factors have a significant effect on the listeners perception as well as their relative interaction.

Figure II.3.1 represents the mean performance reached by each speaker in a decreasing order. A posthoc Tukey test shows that one male and one female performance is significantly different from the other speakers: Subject 21 (mean performance: 7,34) performs best among the other speakers while Subject 13 perform worst (mean performance: 4,69).

	Sum square	Df	Mean square	D	Sig.	Partial $\eta 2$
Speaker	3594,585	18	199,699	73,599	,000	,110
Affect	4914,693	15	327,646	120,754	,000	,145
Sentence	29,531	1	29,531	10,884	,001	,001
Speaker * Affect	4381,363	270	16,227	5,981	,000	,131
Speaker * Sentence	106,413	18	5,912	2,179	,003	,004
Affect * Sentence	978,581	15	63,239	24,044	,000	,033
Speaker * Affect * Sentence	2301,338	270	8,523	3,141	,000	,073

TABLE II.3.1: ANOVA table of the performance evaluation

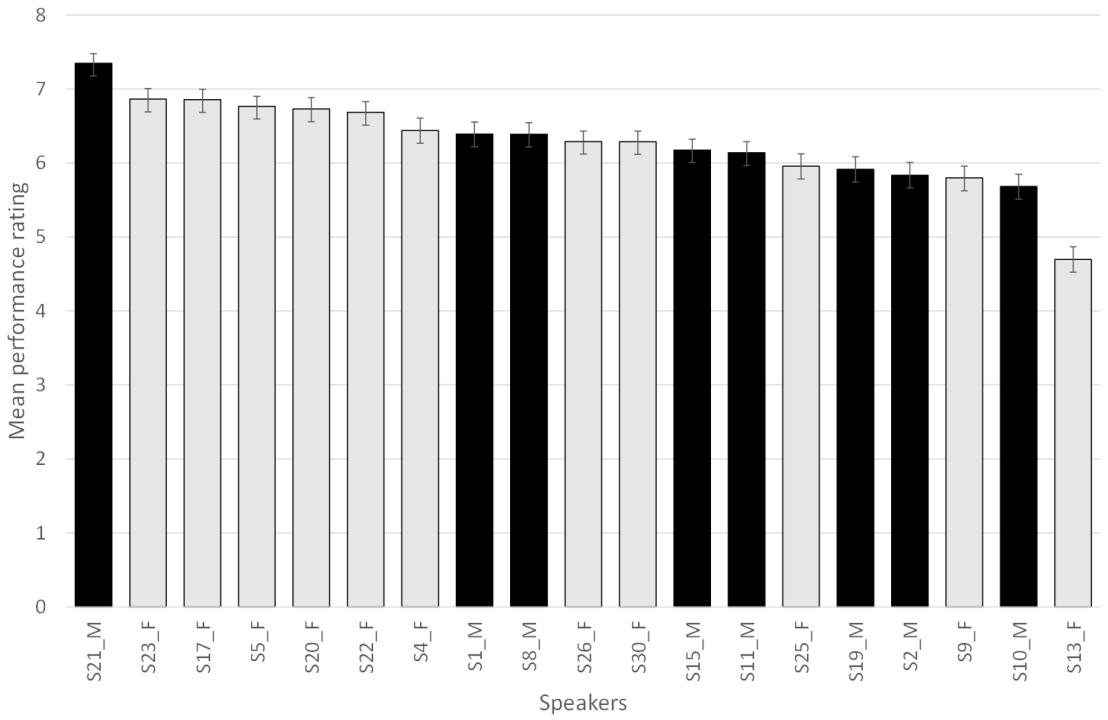


FIGURE II.3.1: Mean performance ratings for each speaker in decreasing order. Performance rating for male and female speakers are respectively displayed in black and grey. Error bars represent the 95% confidence interval

Figure II.3.2 shows the relative performance rate for each affect. Affects concerning unassertive expression are higher rated with a mean of 7,43 for DOUB, 7,24 for SURP and 6,89 for QUES. Among the assertive expressions, IRON was significantly less perceived correctly than all the other expressions. The interaction between Speaker and Affective expression is also significant and reveals that the performances of individual have an impact on the listener's perceptual behaviour.

II.3.2 Categorisation test

Then, in order to investigate the perceptual capacity of listeners to interpret the prosodic expressions of affect (audio alone), we conducted a perceptual test on Japanese native

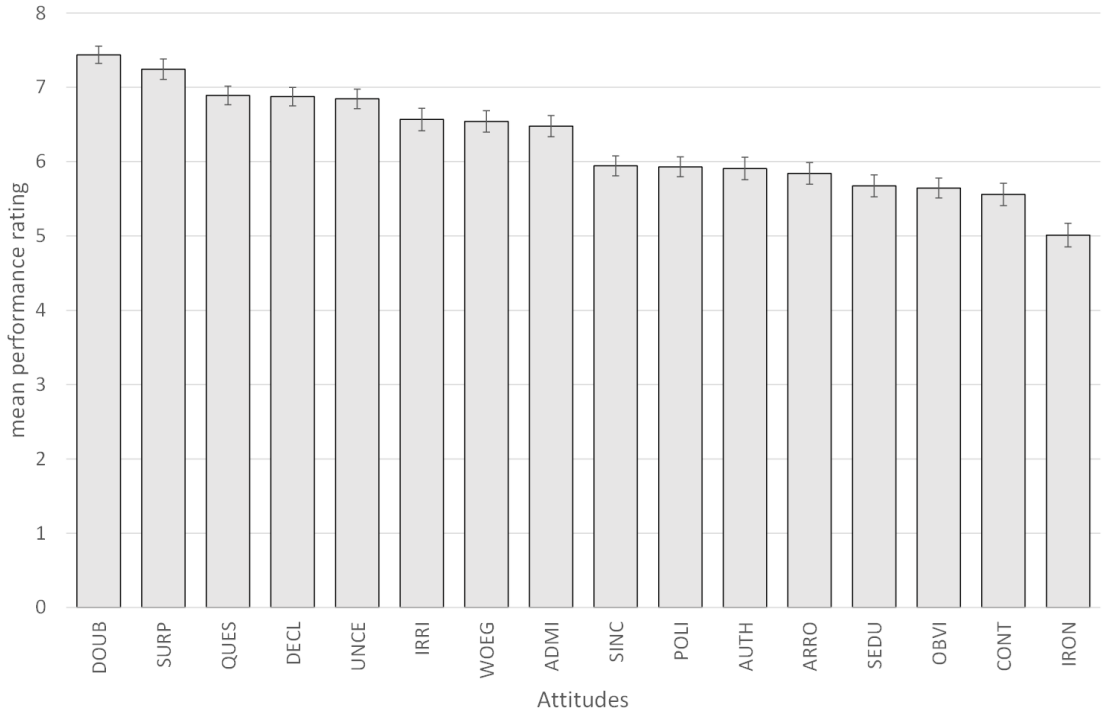


FIGURE II.3.2: Mean performance ratings for each affect in decreasing order. Error bars represent the 95% confidence interval

speakers based on a forced choice paradigm derived from the corpus used for the performance test. Accordingly to results described in [Guerry et al., 2016], we selected for the purpose of this study 9 contexts which can be further regrouped in 4 clusters: The first cluster is composed of Contempt (CONT), Irony (IRON), Irritation (IRRI) and Obviousness (OBVI) and corresponds to expressions of imposition. The second cluster is composed of Politeness (POLI), Seduction (SEDU) and Sincerity (SINC) which are polite and submissive expressions. The third category is composed only of Surprise (SURP) which is a potentially universal affect. The last category contains only "Walking on eggs" which is a dubitative expression. We decided to label these broad categories as respectively Imposition, Politeness, Surprise and Dubitative.

A total of 36 stimuli (9 expressions x 4 speakers) were presented in a random order. The 4 selected speakers are the 2 best performing male and female speakers as identified by the previous experiment. 29 listeners, native speakers of the Tokyo dialect (19 F/ 10 M) participated in this test. The subjects listened to each stimulus only once and were asked to answer the perceived affective expressions amongst 9 possible responses. The interface and instruction were displayed in Japanese.

The results of this experiment are shown on Table II.3.2. Although the overall correct identification rate is not very high (33.6%), it is much higher than the chance level (11.1%). The best recognised social affects are Surprise (94.8%) and Irritation (61.6%). The worst recognised affects are Contempt (9.8%), Sincerity (11.2%) and Seduction (13.8%).

As can be seen on Table II.3.2, most confusions seem to occur within the theoretical categories. For example, Contempt is confused with either Irritation or Obviousness, while most expressions of Seduction and Sincerity are identified as Politeness. Using the broad categories, the results are shown on Table II.3.3.

TABLE II.3.2: Perceptual categorisation (29 listeners)

	CONT	IRON	IRRI	OBVI	POLI	SEDU	SINC	SURP	WOEG
CONT	9.8	5.4	45.5	31.3	5.4	0.0	1.8	0.0	0.9
IRON	34.8	14.3	6.3	7.1	10.7	17.0	2.7	7.1	0.0
IRRI	1.8	2.7	61.6	15.2	12.5	0.9	2.7	2.7	0.0
OBVI	5.4	5.4	29.5	14.3	5.4	0.0	5.4	34.8	0.0
POLI	1.8	0.0	0.0	9.8	33.0	12.5	16.1	6.3	20.5
SEDU	3.4	4.3	2.6	19.8	36.2	13.8	12.9	2.6	4.3
SINC	3.4	2.6	0.9	12.9	42.2	10.3	11.2	3.4	12.9
SURP	4.3	0.0	0.0	0.0	0.0	0.9	0.0	94.8	0.0
WOEG	6.0	2.6	4.3	6.0	15.5	6.9	2.6	7.8	48.3

TABLE II.3.3: Perceptual categorisation (broad classes)

	Imposition	Politeness	Surprise	Dubitative
Imposition	72.5	16.1	11.2	0.2
Politeness	20.6	62.8	4.1	12.5
Surprise	4.3	0.9	94.8	0.0
Dubitative	19.0	25.0	7.8	48.3

With this clustering, the global correct identification rate is 70%. The most important confusion occurs between the Dubitative and Politeness expressions.

II.4 Automatic classification of social affects in Japanese

Although many researches are addressing the problem of ”emotion” or speaker state automatic recognition (see [Valstar et al., 2016] for example), none to our knowledge try to deal with social affective meaning in speech. There are however experiments aiming at measuring acoustic distances between social affects as for example in [Mixdorff et al., 2017]. Furthermore, the results of our perceptual experiment lead us to believe that discrimination between social affects may be carried out to a certain extend using automatic classification. We use the same database that was used in the perceptual tests carried out in section II.3 but the tests are carried out using all the 19 speakers (i.e. not with only the 4 best performing ones).

II.4.1 Experimental design

As not much data is available – the total duration of the corpus is 16 minutes – we devised the experiment as a cross validation one. This means that throughout the experiment, to assure speaker independence of the models, we use all the data from all speakers except the one we test for training the models (i.e. 18 speakers are used for training while 1 speaker is used for the test). This procedure is repeated until all the speakers have been tested.

II.4.1.1 Preprocessing

Since phonetic transcriptions of all the excerpts have been done manually, we decided to use them as a basis for our analysis. All the parameters are then computed on the vocalic segments with the exception of duration, which is computed at the syllable level. As our target sentence is /banana/, we then have 3 points of measure, one for each /a/ while duration is computed for /ba/, /na(1)/ and /na(2)/.

II.4.1.2 Features

The features we decided to use are mainly coming from the matlab toolbox COVAREP [Degottex et al., 2014] which we modified to add some features and to extract features on selected segments. Out of the 37 computed features, 31 are related to acoustic measurements: fundamental frequency (F0, F0SLOPE, FOVAR), the intensity (NRJ, NRJSLOPE, NRJVAR), duration (DUR), harmonics amplitude (H1, H2, H4), formants amplitude (A1, A2, A3), frequencies (F1, F2, F3, F4) and bandwidth (B1, B2, B3, B4), differences between harmonics amplitude (H1-H2, H2-H4), differences between amplitude of harmonics and formants (H1-A1, H1-A2, H1-A3), cepstral peak prominence (CPP), harmonics to noise ratios on different frequency bands (HNR05, HNR15, HNR25, HNR35). 5 features are glottal features that are computed using inverse filtering (IAIF method): normalised

amplitude quotient (NAQ [Alku et al., 2002, Airas and Alku, 2007]), quasi-open quotient (qOQ), difference between amplitude of harmonics 1 and 2 in the estimated glottal signal (H1H2aiff), parabolic spectral parameter (PSP [Alku et al., 1997]), PeakSlope [**?**], maximum dispersion quotient (MDQ [Yanushevskaya et al., 2015]).

As these features are computed on each vowel, we thus have three measurements per social affect per speaker. The number of features per social affect per speaker is thus 111. Some of the features are normalised in order to remove the effect of gender whenever possible. A further normalisation is carried out using the "declarative" sentence, which is considered as reference. All values coming from the reference sentence are then subtracted from the values for each social affect.An example of the resulting normalised fundamental frequency curves can be seen on Figure II.4.1. On this figure, the mean value of normalised fundamental frequency is displayed as the black line, while the inter-speaker variation is represented by the grey area.

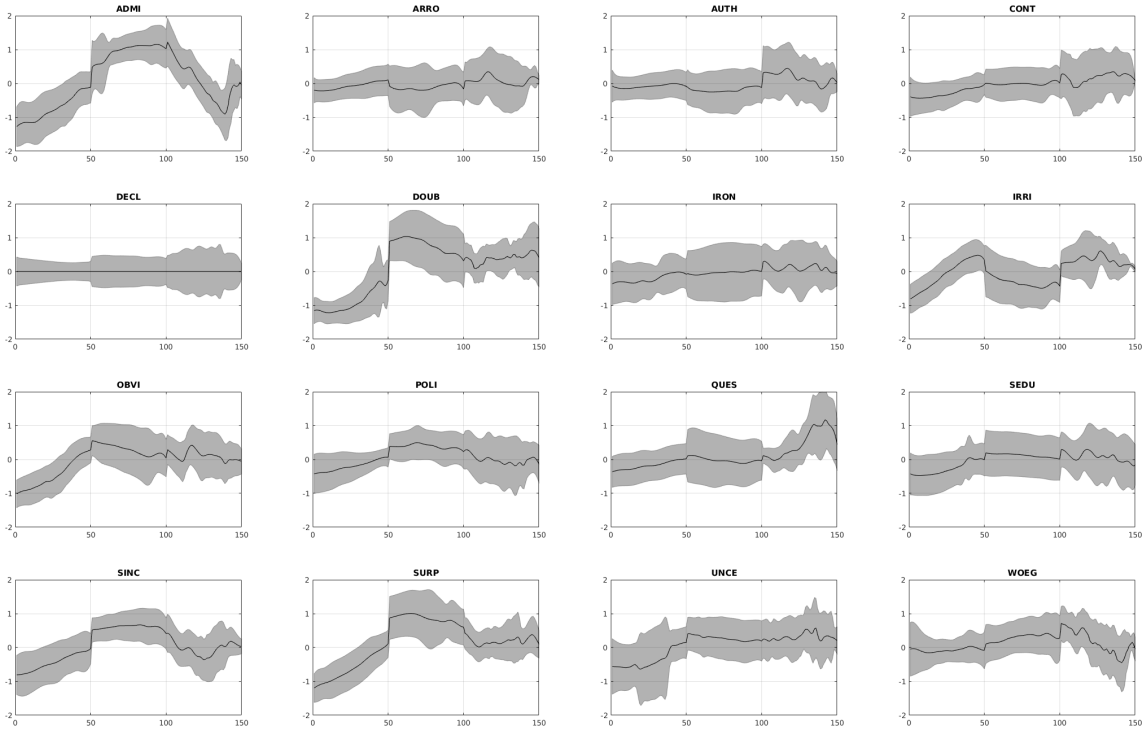


FIGURE II.4.1: Normalised fundamental frequency for 16 attitudes (speaker variation is denoted by the grey area)

Given the dynamic nature of speech, incorporating some kind of dynamic measure may help discriminating between the social affects. That is why we decided to compute the differences between the values on successive vowels, for each of the features described above. This results in having, for example for the F0 feature, instead of [$F0_1$, $F0_2$, $F0_3$] a vector containing [$F0_1$, $F0_2$, $F0_3$, $F0_2 - F0_1$, $F0_3 - F0_2$]. Integrating this information adds 74 more features to the original 111 features set, resulting in a total of 185 features.

II.4.1.3 Features selection

Given this quite important number of features, a feature selection algorithm is used to keep only the most relevant ones. In this work, we decided to use the IRMFSP algorithm,

described in [Peeters, 2003]. It consists in maximising the relevance of the descriptors subset for the classification task while minimising the redundancy between the selected ones.

This iterative method ($\iota \leq p$) is composed of two steps. The first one selects at iteration ι the non-previously selected descriptor which maximizes the ratio between inter-class inertia and the total inertia expressed as follow:

$$\hat{d}^{(\iota)} = \arg\max_{d} \frac{\sum_{k=1}^{K} n_k (\mu_{d,k} - \mu_d)(\mu_{d,k} - \mu_d)^T}{\sum_{i=1}^{n} (f_{d,i}^{(\iota)} - \mu_d)(f_{d,i}^{(\iota)} - \mu_d)^T}, \tag{II.4.1}$$

where $f_{d,i}^{(\iota)}$ denotes the value of descriptor $d \in [1, p]$ affected to the individual i. $\mu_{d,k}$ and μ_d respectively denote the average value of descriptor d into the class k and for the total dataset. The second step of this algorithm aims at orthogonalizing the remaining data for the next iteration as follows:

$$f_d^{(\iota+1)} = f_d^{(\iota)} - \left(f_d^{(\iota)} \cdot g_{\hat{d}}\right) g_{\hat{d}} \ \forall d \neq \hat{d}^{(\iota)}, \tag{II.4.2}$$

where $f_{\hat{d}}^{(\iota)}$ is the vector of the previously selected descriptor $\hat{d}^{(\iota)}$ for all the individuals of the entire dataset and $g_{\hat{d}} = f_{\hat{d}}^{(\iota)} / |f_{\hat{d}}^{(\iota)}|$ is its normalized form.

This algorithm has the advantage of not trying to combine the different features (as what would occur when using a PCA for instance) and of providing a ranking of the supposed discriminant power of the features, allowing to explore the compromise to be made between number of features and expected classification performance.

II.4.2 Experiments

The classification is carried out using cross validation (leaving one speaker out as described above) and a varying number of features (from 1 to 185 ranked using the IRMFSP algorithm). Thus, the exact process for each step of the cross validation is as follows, until all speakers have been used for testing:

- Select a test speaker (all the other speakers will be used for training).
- Carry out the feature ranking process (IRMFSP) on the training data only.
- For a fixed number of features, train a Linear Discriminant model
- Estimate the class of all the recordings made by the test speaker
- Evaluate the performance of the system for the speaker

II.4.2.1 Performance vs. number of features

The first experiment aims at finding the most relevant features. Given the framework described above, we can evaluate the performance of the system, using a varying number of features, starting with the most relevant one according to the IRMFSP algorithm.

The optimal number of features is found to be 8 (see Figure II.4.2). Since the IRMFSP feature ranking is computed at each step of the cross validation process, we kept the ranking at each step and computed the median ranking for each feature across all speakers.

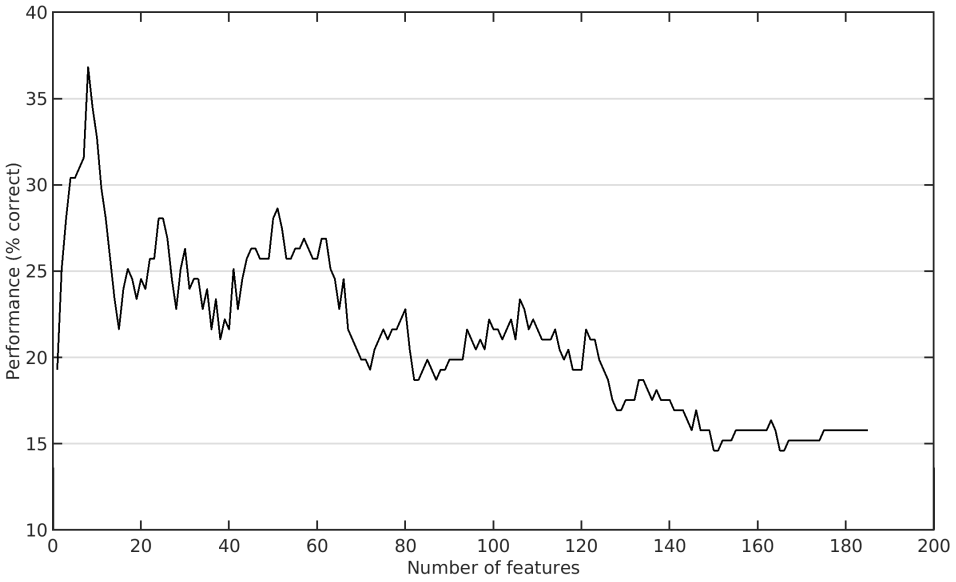


FIGURE II.4.2: Performance vs. number of features ranked using the IRMFSP algorithm

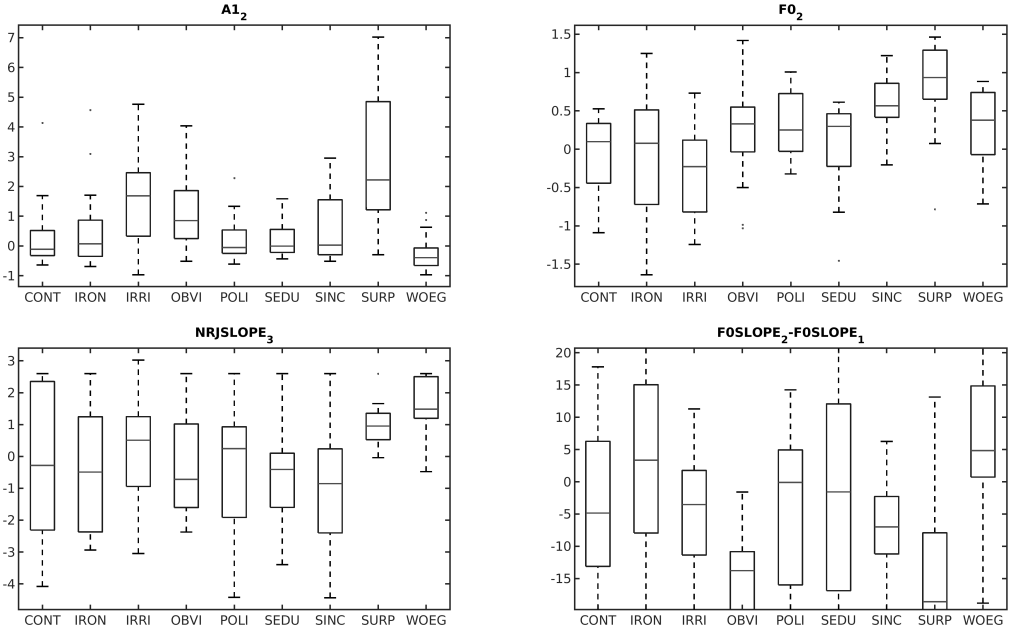
This way, the eight features that have the best median ranking are: (1) $A1_2$: amplitude of the first formant on the second vowel, (2) $F0_2$: mean value of the fundamental frequency on the second vowel, (3) $NRJSLOPE_3$: slope of the energy curve on the third vowel, (4) $F4_1$: frequency of the fourth formant on the first vowel, (5) $F0SLOPE_2 - F0SLOPE_1$: difference between the slope of the fundamental frequency on the second and the first vowel, (6) $NRJ_3 - NRJ_2$: difference on the mean value of intensity between the third and the second vowel, (7) MDQ_1: maxima dispersion quotient on the first syllable, (8) $F0VAR_3 - F0VAR_2$: difference in the variance of fundamental frequency between the third and second syllable.

Among those selected features, we can observe that the first two, i.e. the most discriminant ones, are measurements made on the middle vowel. On figure II.4.3, it can be seen that the values of the first formant amplitude on the second vowel are higher for Surprise, Obviousness and Irritation, while the fundamental frequency on the second vowel is higher for positive expressions (SURP, POLI, SEDU, SINC) and OBVI and WOEG than for CONT, IRON and IRRI. The approximated slope of energy on the third and last vowel show that ending the sentence with rising energy happens for SURP and WOEG. The difference between the slopes of F0 on the first and second vowel aims at focusing on the contrasts between rising/falling or falling/rising patterns and continuing patterns. In that respect, it seems that the intonation patterns are continuous for most expressions except OBVI, SURP and WOEG.

II.4.2.2 Results for the best feature set

The results of the automatic classification experiment are given on Table II.4.1. Overall, the classification achieves a performance of 38.6% of correct identifications. As for the perceptual test, while not being a great performance, this is much higher than chance.

FIGURE II.4.3: Boxplots of some relevant features (a) $A1_2$ (b) $F0_2$ (c) $NRJslope_3$ (d) $F0slope_3 - F0slope_2$



While looking more closely at the results, we can observe that the most easily classified affect is Surprise (78.9%) followed by "Walking on eggs" (57.9%). Some other affects are mildly recognised, such as Irritation (52.6%), Obviousness (47.3%) and Sincerity (47.4%).

TABLE II.4.1: Results of the automatic classification for 9 attitudes (% correct)

	CONT	IRON	IRRI	OBVI	POLI	SEDU	SINC	SURP	WOEG
CONT	36.8	15.8	5.3	10.5	0.0	15.8	5.3	0.0	10.5
IRON	15.8	15.8	15.8	5.3	5.3	10.5	10.5	5.3	15.8
IRRI	5.3	10.5	52.6	10.5	0.0	10.5	5.3	5.3	0.0
OBVI	5.3	0.0	10.5	47.4	5.3	10.5	10.5	10.5	0.0
POLI	5.3	10.5	0.0	5.3	5.3	10.5	26.3	5.3	31.6
SEDU	0.0	31.6	15.8	10.5	10.5	5.3	15.8	0.0	10.5
SINC	5.3	0.0	5.3	10.5	5.3	5.3	47.4	5.3	15.8
SURP	0.0	0.0	0.0	5.3	0.0	0.0	15.8	78.9	0.0
WOEG	10.5	5.3	0.0	5.3	10.5	5.3	5.3	0.0	57.9

With the same 8 features, we reproduced the whole experiment using only the 4 theoretical classes, with the same cross-validation procedure. The results are displayed on Table II.4.2. While achieving a overall identification rate of 60%, the system performs poorly for politeness and dubitative categories. Politeness is often confused with imposition while the dubitative expression is confused with both politeness and imposition expressions. The best recognised expressions are the expression of surprise and the expressions of imposition.

TABLE II.4.2: Automatic classification in 4 broad theoretical classes (% correct)

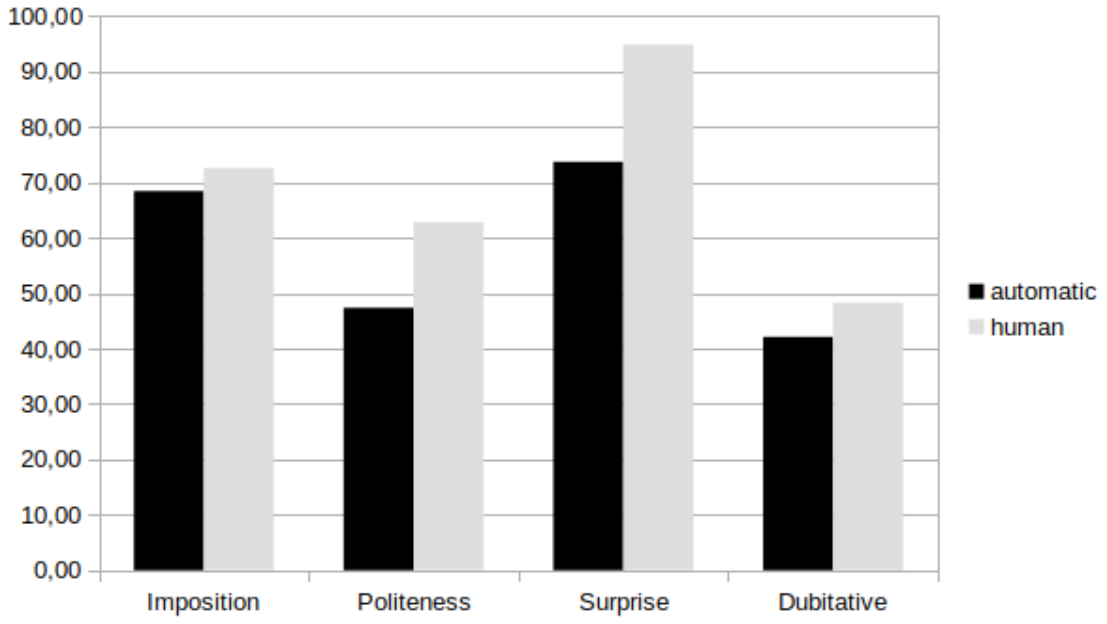
	Imposition	Politeness	Surprise	Dubitative
Imposition	68.4	19.7	3.9	7.9
Politeness	38.6	47.4	3.5	10.5
Surprise	5.3	21.1	73.7	0.0
Dubitative	26.3	31.6	0.0	42.1

II.4.3 Discussion and perspectives

Unfortunately, we were only able to evaluate the four best speakers in the perceptual experiment. This is due to the fact that we need to keep the experiment simple to avoid cognitive overload and that we need to replicate the experiment with many listeners to assess their global behaviour. Concerning the automatic classification experiment, the design is of course different: we do not need many trials because the classification produces the same result each time, but we need to have as many speakers as possible to assess the generalisation of the approach. In that study, we can therefore consider the machine as a particular listener which is used to evaluate all the speakers.

Considering only the broad classes of affects, when looking at the confusion matrix for the perceptual test (Table II.3.3) and the for automatic classification (Table II.4.2), we can observe a rather similar behaviour: Seduction is in both case poorly identified, while the other classes of affect are mostly correctly classified. As a graphic way of comparison, Figure II.4.4 presents the performance obtained separately for each social affect for the automatic classification system and for a categorisation perceptual experiment.

FIGURE II.4.4: % correct for broad classes of social affects by human and machine



These results show the similar behaviour between human perception and automatic

classification of the broad class of social affects. We will need to confirm these results using more data, particularly by testing complete utterances rather than a single word. In the future, we will also reproduce the same experiment using different languages such as French and English.

II.4.4 Overview

While this research is not directly linked to a PhD student supervision on my side, it is to be noted than most of the perceptual experiments were carried out during the course of the PhD thesis of Marine Guerry [Marine Guerry, 2019] who was supervised by my close collaborator Takaaki Shochi from CLLE-ERSSàB. The article [Shochi et al., 2020] was awarded the best paper of the year award.

Part III

Speech recognition in French and integration in a Natural Language Understanding system (PhD thesis Florian Boyer 2017-2021)

III.1 Introduction

From 2017 to 2021, I was the thesis supervisor of Florian Boyer (CIFRE thesis with the Airudit start-up[1]). Although I admit that I am not a speech recognition specialist, I took the opportunity offered with the funding of this thesis so that we could benefit from a state-of-the-art speech recognition system for the French language using the most recent neural network techniques.

The title of Florian's thesis is "Reconnaissance de parole pour le français et intégration dans un système de compréhension du langage parlé" which can be translated to "Speech recognition for the French Language and integration in a Natural Language Understanding system".

The aim of the work carried out by florian was to study the possibilities offered by speech recognition systems for interacting with a natural language understanding algorithm developed and used by the Airudit company. To do so, he first studied the performances of speech recognition systems on French and then proposed a linking method with the understanding system.

During this thesis, some original contributions have been made in the field of automatic speech recognition and spoken language understanding. They can be either of academic or industrial nature. These contributions are resumed in the following list:

- Evaluation of all methods and architectures currently used for ASR in French
- Study on the minimal unit to be used to model the ASR problem in French, in an isolated manner and in consideration of the NLU problem.
- Formalisation of errors produced by end-to-end ASR systems in French and evaluation of their impact for the NLU problem.
- Contributions to several open-source projects and toolboxes used for ASR. These contributions related to several aspects such as: training, inference (offline and online/streaming modes), new fonctionalities and recipes for ASR systems building in several languages.

In this document, I propose to provide an overview of speech recognition performances for the French language. Florian Thesis has taken place at the time where most research on speech recognition switched from what we may call "classical" approaches, involving acoustic models, lexicon and language models and thus linguistic knowledge, to "end-to-end" approaches which do not require any linguistic knowledge at all. This is illustrated by Figure III.1.1 where the performances on English conversational telephony were reported in 2018. I propose here as an illustration of Florian's work to provide the same kind of picture for the French language.

[1] www.airudit.com

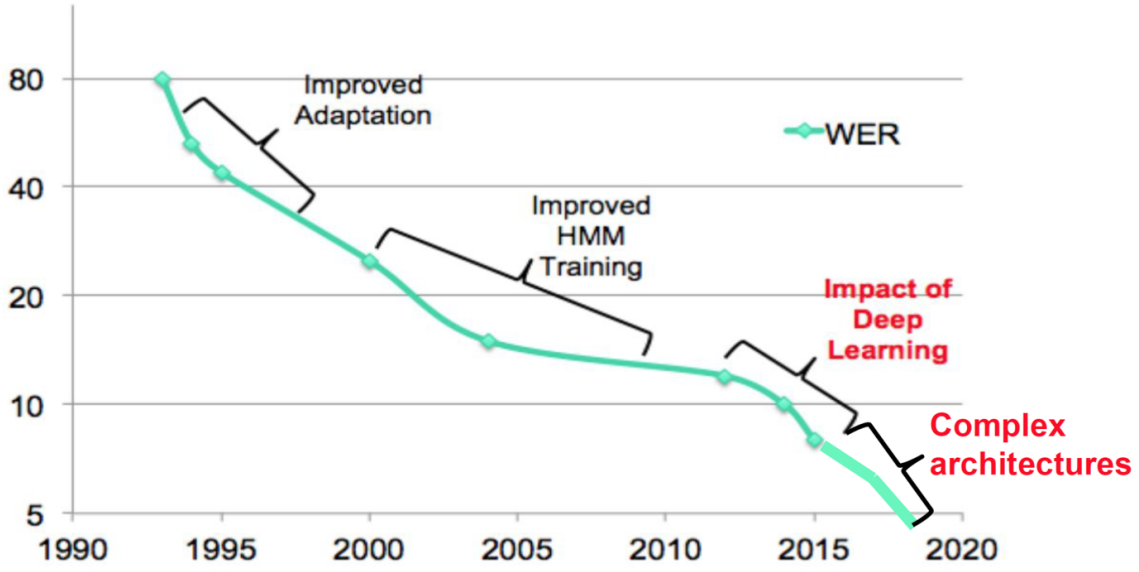


FIGURE III.1.1: ASR Performance on English Conversational Telephony (Switchboard) - Image from Bhuvana Ramabhadran' s presentation at Interspeech 2018

III.2 Database

We carried out our experiments using the data provided during the ESTER evaluation campaign (*Evaluation of Broadcast News enriched transcription systems*) [Galliano et al., 2009] which is one of the most commonly used corpus for the evaluation of French ASR. Evaluations are done on test set. The details of the dataset are described in [Galliano et al., 2009]. We use the same normalization and scoring rules as in the evaluation plan of the ESTER 2 campaign except that we do not use equivalence dictionary and partially pronounced words are scored as full words.

To train the acoustic models we use the 90h of the training set from ESTER2 augmented by 75h from ESTER1 training set and 90h from the additional subset provided in ESTER1 with their transcriptions provided in the corpus EPAC [Estève et al., 2010]. We removed segments containing less than 1,5 seconds of transcribed speech and we excluded the utterances corresponding to segments with more than 3000 input frames or sentences of more than 400 characters for the end-to-end models. Because some irregulars segment-utterance pairs remained, we re-segmented the training data using the GMM-HMM model (with LDA-MLLT-SAT features) we build our phone-based chain model upon. During re-segmentation, only the audio parts matching the transcripts are selected. This brings the training data to approximately 231h. 3-fold speed perturbation [Ko et al., 2015] and volume perturbation with random volume scale factor between 0.25 and 2 may also be applied, leading to a total of training data of 700h.

For language modeling, we use the manual transcripts from the training set. We extend this set with manually selected transcriptions from other speech sources (BREF corpus [Lamel et al., 1991], oral interventions in EuroParl from '96-'06 [Koehn, 2005] and a small portion of transcriptions from internal projects). The final corpus is composed of 2 041 916 sentences, for a total of 46 840 583 words.

III.3 Systems

III.3.1 Baseline system

The baseline system is based on the usual pipeline for speech recognition (illustrated on Figure III.3.1) with the following steps:

- optional data augmentation step at the signal level (speed/volume perturbations)
- feature extraction (usually MFCCs)
- acoustic modelling
- pronounciation modelling (lexicon)
- language modelling

This "classic" approach has been built using the Kaldi toolbox [Povey et al., 2011] with the chain model variant.

The data used for training the models is detailed in Table III.3.1

Corpus	Transformation	Audio data AM	Text data
ESTER1 + ESTER2 + EPAC	Re-alignment +*Speed-Perturbation* (x3) + *Volume Perturbation*	231 hours (x3) = env. 700 hours	2,041,916 sentences (46,840,583 words)

TABLE III.3.1: Summary of the data used for training the models

As features for our models, we use a 40-dimensional high resolution MFCC vector (*i.e.* linear transform of the filterbanks) and CMVN. We also trained separately a phone-based chain model with the previous 40-dimensional MFCC vector concatenated with a 100-dimensional i-vector [Gupta et al., 2014] as input to assess the impact of speaker-dependent features.

The acoustic model is a TDNN-HMM model trained with the LF-MMI objective function. The neural network is based on a sub-sampled time-delay neural network (TDNN)

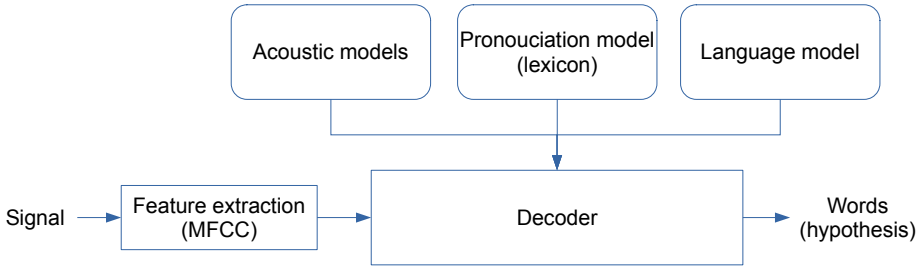


FIGURE III.3.1: Illustration of the components of a "classic" ASR system

with 7 TDNN layers and 1024 units in each, time stride value being set to 1 in the first three layers, 0 in the fourth layer and 3 in the final ones.

For the linguistic part, we also trained a word 3-gram language model using SRILM's n-gram counting method [Stolcke, 2002] with KN discounting. As lexicon we use the phonetic dictionary provided by the LIUM, thus the vocabulary of our language model is limited to the most frequent 50k words found in our training texts and also present in their dictionary.

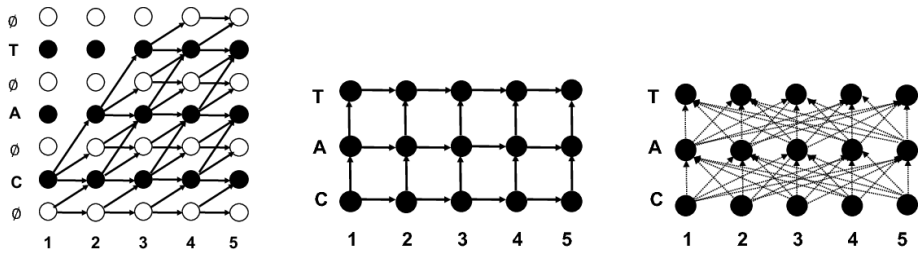
The performances obtained using the baseline system are reported in Table III.3.2.

Acoustic model	Lexicon	Language Model	WER (%)
HMM-GMM	50,000 words	3-gram	16.8
HMM-TDNN (+ i-vectors)	-	-	14.2 (13.7)

TABLE III.3.2: Results for the baseline system on the ESTER2 test set

III.3.2 "End-to-end" approaches

The baseline approach requires each step of the pipeline to be learned separately, using at least some linguistic knowledge for the building of the pronounciation model. The end-to-end approaches have the advantage that they do not require any linguistic knowledge for building the models. They can therefore be trained without the need of language experts and also take advantage of the availability of new data to extend their vocabulary without any additional cost.



(a) CTC (b) RNN-Transducer (c) Attention

FIGURE III.3.2: Illustration of transition probabilities for the three end-to-end methods on 5-frame input with "CAT" labels [Battenberg et al., 2017]. The node at (t, u) represents the output probability for the u first elements of the output sequence at time t. Vertical arrows indicate the prediction of several labels at time t (not allowed for CTC). Horizontal arrows represent the prediction of repetitive characters (for CTC) or the prediction of null labels (RNN-Transducer) Bold arrows represent the hard alignments (CTC or RNN-Transducer) and soft alignments (Attention)

We decided to study the three main types of architectures for end-to-end systems that are commonly used: 1) the *Connectionnist Temporal Classification* (CTC) [Graves et al., 2006] which uses Markov assumptions (*i.e.* conditional independence between predictions at each time step) to efficiently solve sequential problems by dynamic programming, 2) The Attention-based encoder-decoder approach [Bahdanau et al., 2016, Chorowski et al., 2015] which rely on an attention mechanism to perform non-monotonic alignment between acoustic frames and recognized acoustic units. 3) RNN-transducers

[Graves, 2012] which extends CTC by additionally modeling the dependencies between outputs at different steps using a prediction network analogous to a language model.

The differences between the three approaches can be summarised as follow:

- The CTC approach considers a conditional independence between predictions at different time steps. RNN-Transducers and Attention model the inter-dependence between the predicted labels.
- The alignments between input and output sequences are monotonous for CTC.
- The CTC and RNN-Transducers approaches consider alignments between inputs and outputs as latent variables and marginalise over all alignements whereas the Attention-based Encoder-Decoder approach models a soft alignement between each input and output step.

These differences are illustrated on Figure III.3.2 where are represented the alignment carried out between acoustic observation sequence inputs and orthographic labels outputs.

We then compare out results with the baseline system described in section III.3.1 and the end-to-end variant of this system [Hadian et al., 2018].

III.3.2.1 Connectionist Temporal Classification

The CTC [Graves et al., 2006] can be seen as a direct translation of conventional HMM-DNN ASR systems into *lexicon-free* systems. Thus, the CTC follows the general ASR formulation, training the model to maximize $P(Y|X)$ the probability distribution over all possible label sequences:

$$\hat{Y} = arg \max_{Y \in \mathcal{A}*} p(Y|X)$$

Here, X denotes the observations, Y is a sequence of acoustic units of length L such that $Y = \{y_l \in \mathcal{A} | l = 1, ..., L\}$, where $\mathcal{A}$ is an *alphabet* containing all distinct units. As in traditional HMM-DNN systems, the CTC model makes conditional independence assumptions between output predictions at different time steps given aligned inputs and it uses the probabilistic chain rule to factorize the posterior distribution $p(Y|X)$ into three distributions (*i.e.* framewise posterior distribution, transition probability and prior distribution of units). However, unlike HMM-based models, the framewise posterior distribution is defined here as a framewise acoustic unit sequence B with an additional blank label $<blank>$ such as $B = \{b_t \in \mathcal{A} \cup <blank> | t = 1, ..., T\}$.

$$p(Y|X) = \underbrace{\sum_{b=1}^{B} \prod_{t=1}^{T} p(b_t|b_{t-1}, Y) p(b_t|X)}_{p_{\text{ctc}}(Y|X)} p(Y)$$

Here, $<blank>$ introduces two contraction rules for the output labels, allowing to repeat or collapse successive acoustic units.

III.3.2.2 Attention-based model

As opposed to CTC, the attention-based approach [Bahdanau et al., 2016, Chorowski et al., 2015] does not assume conditional independence between predictions at different time steps and does not marginalize over all alignments. Thus the

posterior distribution $p(Y|X)$ is directly computed by picking a soft alignment between each output step and every input step as follows:

$$p_{\text{att}}(Y|X) = \prod_{l=1}^{U} p(y_l|y_1, ..., y_{l-1}, X)$$

Here $p(y_l|y_1, ..., y_{l-1}, X)$ – our attention-based objective function – is obtained according to a probability distribution, typically a softmax, applied to the linear projection of the output of a recurrent neural network (or long-short term memory network), called decoder, such as:

$$p(y_l|y_1, ..., y_{l-1}, X) = \text{softmax}(\text{lin}(\text{RNN}(\cdot)))$$

The decoder output is conditioned by the previous output y_{l-1}, a hidden vector d_{l-1} and a context vector c_l. Here d_{l-1} denotes the high level representation (*i.e.* hidden states) of the decoder at step $l-1$, encoding the target input, and c_l designate the context – or symbol-wise vector in our case – for decoding step l, which is computed as the sum of the complete high representation h of another recurrent neural network, encoding the source input X, weighted by α the attention weight:

$$c_l = \sum_{s=1}^{S} \alpha_{l,s} h_s \quad , \quad \alpha_{l,s} = \frac{\exp(e_{t,s})}{\sum_{s'=1}^{S} \exp(e_{l,s'})}$$

where e_t, also referred to as *energy*, measures how well the inputs around position s and the output at position l match, given the decoder states at decoding step $l-1$ and h the encoder states for input X. In the following, we report the standard content-based mechanism and its location-aware variant which takes into account the alignment produced at the previous step using convolutional features:

$$e_{l,s} = \begin{cases} \text{content-based:} \\ \quad w^T \tanh(W d_{l-1} + V h_s + b) \\ \text{location-based:} \\ \quad f_u = F \star \alpha_{-1} \\ \quad w^T \tanh(W d_{l-1} + V h_s + U f_{l,s} + b) \end{cases}$$

where w and b are vectors, W the matrix for the decoder, V the matrix for the high representation h and U the matrix for the convolutional filters, that takes the previous alignment for location-based attention mechanism into account.

III.3.2.3 RNN transducer

The RNN transducer architecture was first introduced by [Graves, 2012] to address the main limitation of the proposed CTC network: it cannot model interdependencies as it assumes conditional independence between predictions at different time steps.
To tackle this issue, the authors introduced a CTC-like network augmented with a separate RNN network predicting each label given the previous ones, analogous to a language model. With the addition of another network taking into account both encoder and decoder outputs, the system can jointly model interdependencies between both inputs and outputs and within the output label sequence.

Although the CTC and RNN-transducer are similar, it should be noted that unlike CTC which represent a loss function, RNN-transducer defines a model structure composed of the following subnetworks :

- The encoder or transcription network: from an input value x_t at timestep t this network yields an output vector h_t of dimension $|\mathcal{A}+1|$, where $+1$ denotes the $<blank>$ label which acts similarly as in CTC model.
- The prediction network: given as input the previous label prediction $y_{u-1} \in \mathcal{A}$, this network compute an output vector d_u dependent of the entire label sequence $y_0, ..., y_{u-1}$.
- The joint network: using both encoder outputs h_t^{enc} and prediction outputs d_u^{dec}, it computes $z_{t,u}$ for each input t in the encoder sequence and label u in prediction network such as:

$$h_{t,u}^{joint} = tanh(h_t^{enc} + h_u^{dec})$$
$$z_{t,u} = lin(h_{t,u}^{joint})$$

The output from the joint network is then passed to a softmax layer which defines a probability distribution over the set of possible target labels, including the blank symbol.

It should be noted that we made a small modification compared to the last proposed version [Graves et al., 2013]: instead of feeding the hidden activations of both networks into a separate linear layer, whose outputs are then normalised, we include another linear layer and feed each hidden activations to its corresponding linear layer which yields a vector of dimension J, the defined *joint-space*.

Similarly to the CTC, the marginalized alignments are local and monotonic and the label likelihood can be computed using dynamic programming. However, unlike CTC, RNN transducer allows prediction of multiple characters at one time step, alongside their vertical probability transitions.

III.3.2.4 Other notable approaches

Joint CTC-attention The key idea behind the joint CTC-Attention [Kim et al., 2017] learning approach is simple. By training simultaneously the encoder using the attention mechanism with a standard CTC objective function as an auxiliary task, monotonic alignments between speech and label sequences can be enforced to reduce the irregular alignments caused by large jumps or loops on the same frame in the attention-based model. The objective function below formulates the multi-task learning of the network, where $0 \leq \lambda \leq 1$ is a tunable parameter weighting the contribution of each loss function:

$$\begin{aligned}\mathcal{L}_{MTL} &= \lambda \mathcal{L}_{\text{ctc}} + (1-\lambda)\mathcal{L}_{\text{att}} \\ &= \lambda \log p_{\text{ctc}}(Y|x) + (1-\lambda)\log p_{\text{att}}(Y|x)\end{aligned}$$

The approach proposed in [Hori et al., 2017b] introduced a joint-decoding method to take into account the CTC predictions in the beam-search based decoding process of the attention-based model. Considering the difficulty to combine their respective scores, the attention-based decoder performs the beam search character-synchronously whereas the CTC performs it frame-synchronously, two methods were proposed.

The first one is a two-pass decoding process where the complete hypotheses from the attention model are computed and then rescored according to the following equation, where $p_{\text{ctc}}(Y|x)$ is computed using the standard CTC forward-backward algorithm:

$$\hat{Y} = \arg\max_{C \in A*} \{\lambda \log\ p_{\text{ctc}}(Y|x) + (1-\lambda) \log\ p_{\text{att}}(Y|x)\}$$

The second method is a one-pass decoding method where the probability of each partial hypothesis in the beam search process is computed directly using both CTC and attention model such as, given h the partial hypothesis and α the score defined as the log probability of the hypothesized sequence:

$$\alpha(h) = \lambda \alpha_{\text{ctc}}(h) + (1-\lambda)\alpha_{\text{att}}(h)$$

End-to-end lattice-free MMI The end-to-end Lattice-Free MMI [Hadian et al., 2018] is the end-to-end version of the method introduced by Povey et al. [Povey et al., 2016]. In this version, a flat-start manner is adopted in order to remove the need of training an initial HMM-GMM for alignments and the tree-building pipeline. Although the approach seems more like a flat-start adaptation of the state-of-art method than end-to-end in terms of pipeline and it does not benefit from the open-vocabulary property to construct unseen words compared to previously presented methods, we use it in our experiments as it showed small degradation over the original lattice-free MMI with different acoustic units. We can therefore contrast the orthographic differences in productions between open systems and more constrained ones where the relationship between acoustic units and a word-level representation is restricted.

RNN-transducer with attention The RNN transducer architecture augmented with attention mechanisms was first mentioned, to the best of our knowledge, in [Prabhavalkar et al., 2017]. Here, the prediction network described in III.3.2.3 is replaced by an attention-based decoder similar to the one described in III.3.2.2 and used in the joint CTC-attention. This modification allows the decoder to access acoustic information alongside the sequence of previous predictions. As the decoder output computation is not affected by this change (the decoder and joint outputs computation are not dependent on a particular choice of segmentation), the architecture can be trained with the same forward-backward algorithm used for standard RNN-transducer. Finally, unlike the previous hybrid procedure, the inference procedure can be performed frame-synchronously with an unmodified greedy or beam search algorithm.

III.3.3 Implementations

All our systems share equivalent optimization – no rescoring technique or post-processing is done – as well as equivalent resource usage. Each system is kept to its initial form (*i.e.* no further training on top of the reported system).

III.3.3.1 Acoustic units

For our experiments, three kind of acoustic units were chosen: phones, characters and subwords. The baseline phone-based systems use the standard 36 phones used in French. The CTC, attention and hybrid systems each have two versions: one for characters with 41 classes (26 letters from the Latin alphabet, 14 letters with a diacritic and apostrophe) and another version for subwords where the number of classes is set to 500, the final

set of subword units used in our training being selected by using a subword segmentation algorithm based on a unigram language model [Kudo, 2018] and implemented in Google's toolkit SentencePiece [Kudo and Richardson, 2018]. For the end-to-end variant of the chain model, characters units are used with the 41 classes set.

III.3.3.2 Models

We use the ESPNET toolkit [Watanabe et al., 2018] to train the five end-to-end systems. For each method two acoustic units are used: character and subword. Ten epochs are used to train each model.

The acoustic models for all methods share the same architecture composed of VGG bottleneck [Hori et al., 2017b] followed by a 3-layer bidirectional LSTM with 1024 units in each layer and each direction. For the models using attention mechanism we use a 1-layer LSTM with 1024 units and location-based mechanism with 10 centered convolution filters of width 100 for the convolutional feature extraction as decoder. When training jointly CTC and attention, λ was set to 0.3 based on preliminary experiments. For RNN-transducer the joint space between encoder and decoder was set to 1024 dimensions.
The input features for these models are a 80-dimensional raw filterbanks vector with their first and second derivatives with cepstral mean normalization (CMN).

For the experiments involving language models, we trained three different models using the RNNLM module available in ESPNET: one with characters, another with subwords and the last one with full words for multi-level combination when dealing with characters as units. Each model is incorporated at inference time using shallow fusion [Kannan et al., 2018], except for the word-LM relying on multi-level decoding [Hori et al., 2017a]. The main architecture of our RNNLMs is a 1-layer RNN, the number of units in each layer depending of the target unit: 650 units for subwords and characters, and 1024 units for words. Unlike the systems described above, the vocabulary for the word-based RNNLM was limited according to the training texts only.

In order to directly compare the baseline systems to the end-to-end systems relying on different word-based LM (*i.e.* N-gram and RNN-based), another RNNLM was trained using available tools in Kaldi. The language model shares the same architecture as the word-RNNLM described in this subsection and was trained with equivalent training parameters. Following lattice rescoring approach proposed in [Xu et al., 2018a], decoding was then performed with the RNNLM for all baseline systems. We observe a maximal WER improvement of 0.12% on the dev set and 0.16% on the test set compared to the systems relying on the original 3-gram. Adding to that a difference of less than 1.3% between words in language model vocabularies for baseline and end-to-end systems, we thus consider minimal the impact for our comparison.

III.3.3.3 Decoding

To measure the best performance, we set the beam size to 30 in decoding under all conditions and for all models. When decoding with the attention-only model, we do not use sequence length control parameters such as coverage term or length normalization parameters [Wu et al., 2016]. When joint-decoding, λ is set to 0.2 based on our preliminary experiments. For CTC and attention experiments involving a RNNLM, the language model weight during decoding is set to respectively 0.3 for character and subword LM, and 1.0 for the word LM. For RNN-transducer, we downscale the use of external language model when performing multi-level LM decoding, setting the value to 0.3.

III.4 Results

AM	Unités	LM	CER	WER
HMM-TDNN	phones	3-gram (word)		**14.2**
e2e HMM-TDNN [Hadian et al., 2018]	phones	4-gram (phon.) + 3-gram (mot)		14.4
	characters	4-gram (char.) + 3-gram (word)	**7.6**	14.8
CTC	characters		15.5	42.3
		RNNLM (char.)	13.3	31.0
		RNNLM (word)	12.8	27.3
	bpe		*20.1*	28.4
		RNNLM (bpe)	*17.5*	21.2
Attention	characters		13.2	24.4
		RNNLM (char.)	12.8	23.6
		RNNLM (words)	12.7	23.0
	bpe		*19.5*	22.7
		RNNLM (bpe)	*18.4*	21.8
RNN Transducer	characters		8.5	19.7
		RNNLM (char.)	8.2	18.8
		RNNLM (words)	8.0	18.1
	bpe		*15.5*	18.5
		RNNLM (bpe)	*14.7*	**17.4**
CTC-Attention	characters		10.4	22.1
		RNNLM (char)	10.1	20.6
		RNNLM (words)	9.6	18.6
	bpe		*15.3*	18.7
		RNNLM (bpe)	***14.5***	17.8
RNN Transducer + attention	characters		8.2	19.1
		RNNLM (char.)	8.0	18.3
		RNNLM (words)	**7.8**	17.6
	bpe		*15.6*	18.4
		RNNLM (bpe)	*14.9*	17.5

TABLE III.4.1: End-to-end ASR systems and their performances on the test set of ESTER2

The results of our experiments in terms of Character Error Rate (CER) and Word Error Rate (WER) on the test set are gathered in Table III.4.1. For CER we also report errors in the metric: correct, substituted, inserted and deleted characters.
It should be noted that the default CER computation in all frameworks does not use a special character for space during scoring. As important information relative to this character, denoting word-boundary errors, can be observed through the WER variation

during comparison, we kept the initial computation for CER. Thus, for low CER variations, bigger WER differences are expected notably between traditional and end-to-end systems.

III.4.1 Baseline system and its end-to-end variant

The phone-based chain model trained with lattice-free MMI criterion has a WER of 14.2 on the test set. Compared to the best reported system during the ESTER campaign (WER 12.1% [Galliano et al., 2009]), the performance show a relative degradation of 14.8%. Although the compared system rely on a HMM-GMM architecture, it should be noted that a triple-pass rescoring (+ post-processing) is applied, a consequent number of parameters is used, and a substantial amount of data is used for training the language model (more than 11 times our volume). Adding i-vectors features the performance of our model is further improved, leading to a WER of 13.7.

For the *end-to-end* phone-based system we denote a small WER degradation of 0.2% compared to the original system without i-vectors, which is a good trade-off considering the removal of the initial HMM-GMM training. Switching to characters as acoustic units we obtain a WER of 14.8, corresponding to a CER of 7.6. The detailed report show that all types of errors are quite balanced, with however a higher number of deletions. The system remains competitive even with orthographic units, despite the low correspondence between phonemes and letters in French. On the same note, a plain conversion of phonetic lexicon to a grapheme-based one does not negatively impact the performances. This was not excepted considering the use of alternative phonetic representation in French to denotes possible *liaisons* (the pronunciation of the final consonant of a word immediately before a following vowel sound in preceding word).

III.4.2 End-to-end systems

Character-based models While, without language model, the attention-based model outperforms CTC model as expected, RNN-transducer performances exceed our initial estimations, surpassing previous models in terms of CER and WER. RNN-transducer even outperforms these models coupled with language model, regardless of the level of knowledge included (character and word-level). The CER obtained with this model is 8.5 while the WER is 19.7. This represent a relative decrease of almost 40% for the CER and 17% for the WER against the attention-based model with word LM, the second best system for *classic* end-to-end. Compared to the end-to-end chain model system modeling characters, we observe a small CER difference of 0.9 which corresponds to a WER difference of 4.9. While the CER is competitive, errors at word-level seem to indicate difficulties to model word boundaries compared to baseline systems.

Extending the comparison to hybrid models, only the RNN-transducer with attention mechanism could achieve similar or better results than its vanilla version. Although the joint CTC-attention procedure is beneficial to correct some limitations from individual approaches, the system can only reach a CER of 10.4 equivalent to a WER of 22.1. However, by adding word LM and using multi-level decoding, the system can achieve closer WER performance (18.6) despite the significant difference in terms of CER (9.6).
For the hybrid transducer relying on additional attention module, performances in all experiments are further improved compared to standard, reaching 8.2% CER and 19.1% WER without language model.

Concerning the best systems, it should be noted that the RNN-transducer performance is further improved with the use of language model, obtaining a CER of 8.0, close to our baseline score (7.6), with a word LM. In terms of WER it represents a relative improvement of 8.5% against previous results, which is however still far from the performance denoted with the baseline system for this metric (14.8%). For the RNN-transducer with an attention decoder, we achieve even better performance with a CER of 7.8 equivalent to a WER of 17.6. This is our best model with characters as *acoustic* units.

Subword-based models Replacing characters with subword units improves the overall performance of all end-to-end methods. The gain is particularly important for CTC lowering the WER from 42.3 to 28.4 without language model. The gain observed when adding the language model to CTC is impressive with a relative improvement of almost 28% on WER (from 28.4 to 21.2). For the system relying only on attention, the WER is further improved without and with language model but, unlike when we used characters, the model is outperformed on both CER and WER by the model relying on CTC. Although we observe a similar CER for both methods we also note a significant difference in terms of correct characters and WER (almost 6%). The attention making mostly consecutive mistakes on the same words or groups of words (particularly at the beginning and end of utterances) while the CTC tends to recognize part of words as independent, thus incorrectly recognizing word boundaries. Adding RNN-transducer to the comparison, both previous methods are surpassed, on CER (20.1 for CTC, 17.5 for attention and 15.2 for transducer) and on WER (21.1 for CTC, 21.8 for attention and 18.4 for transducer). Decoding with an external language model, the CER and WER are further improved by about 5.5% and 6.0%. It should be noted that the transducer model without language model exceed CTC and attention coupled to the subword LM.

Adding the hybrid systems to the comparison, we denote some differences compared to character-based systems. The RNN-transducer is not improved with attention mechanism and even slightly degraded for both CER and WER. The same observations can be done with and without LM addition. It seems the attention mechanism has more difficulty to model intra-subwords relations than intra-characters relations. However further work should be allocated to extend the comparison with different attention mechanisms, such as multi-head attention, and estimate the influence of architecture depending on output dimensions and representations.
Concerning the last hybrid system, joint CTC-attention is better suited to subword than characters, reaching comparable performances to transducer even without language model: 18.7% against 18.4 for RNN-transducer and 18.5. Although transducer are reported as our best system, it should be noted that joint CTC-attention reach equal or better performance on subword errors. Talking only about conventional ASR metric, we consider the two hybrid systems and vanilla transducer equivalent for subword units.

III.4.3 Conclusion

We showed that end-to-end approaches and different orthographic units were rather suitable to model the French language. RNN-transducer was found specially competitive with character units compared to other end-to-end approaches. Among the two orthographic units, subword was found beneficial for most methods to address the problems described in section III.4.2 and retain information on ambiguous patterns in French. Extending with language models, we could obtain promising results compared to traditional phone-based systems. The best performing systems being for character unit the RNN-transducer with additional attention module, achieving 7.8% in terms of

CER and 17.6% on WER. For subword units, classic RNN-transducer, RNN-transducer with attention and joint CTC-attention show comparable performance on subword error rate and WER, with the first one being slightly better on WER (17.4%) and the last one having a lower error rate on subword (14.5%).

III.5 Overview

Apart from the numerous experiments carried out by Florian, he also has contributed greatly to the open source ESPNET project [Watanabe et al., 2018] with more than 62000 lines of code. His main contribution include:

- Finetuning techniques: knowledge transfer and parameter freezing
- the Conformer architecture [Guo et al., 2021]
- contributions to the Transducer model: Beam search algorithms (notably the N-Step Constrained beam search (NSC) and the modified Adaptive Expansion Search (mAES)) and auxiliary loss functions [Boyer et al., 2021])

The most recent results for French are displayed on Table III.5.1.

Year	Unit	AM	LM	CER	WER	RTF
2009	phones	HMM-GMM (LIMSI) [Galliano et al., 2009]	?	X	12.1	> 5.0[1]
2021	phones	HMM-TDNN + *SelfAttn* [Heba, 2021]	4+5-gram	X	**11.7**	**0.4**
2018	phones	chain LF-MMI	3-gram (word)	X	13.7	0.31
2019	bpe	RNN-T Att	RNNLM (word)	7.8	17.6	N.A
2019	bpe	RNN-T	RNNLM (bpe)	*14.7*	17.4	N.A
2021	characters	RNN-T	X	6.9	15.6	0.128
		+ Aux. task.		6.7	15.1	**0.126**
2021	characters	Conformer/RNN-T	X	5.1	11.6	0.134
		+ Aux. task.		**5.0**	**11.3**	0.130

TABLE III.5.1: Recent results on the ESTER2 test set

Part IV

New speech biomarkers for sleepiness detection (PhD thesis Vincent Martin 2019-2022)

IV.1 Introduction

One of the major challenges for diagnosing and treating neuro-psychiatric pathologies is symptom quantification and follow-up of chronic patients in order to adapt treatment and measure early relapses. Such an ecological monitoring is possible thanks to connected medical devices (measuring for instance weight, blood pressure or physical activities) but crucial information about how the patients report clinical symptoms like fatigue or sleepiness are difficult to measure. Regular in-person appointments between doctors and patients are useful but miss a large part of variability of symptoms at home in response to treatment. Furthermore, the growing number of patients increases the queuing time and often results in episodic follow-ups with unevenly spaced interviews.

Apart from the clinical interviews, it is nonetheless possible to measure some symptoms (*e.g.* sadness or sleepiness) with a range of behavioural analysis techniques: looking at eye movements and examining verbal expressions or body movements [Poursadeghiyan et al., 2018, Xu et al., 2018b]. Thanks to recent advances in speech processing, it seems now possible to detect precise cues in voice allowing to characterise the state of a speaker. This could potentially allow to measure the level of sleepiness, fatigue or sadness [Cummins et al., 2018]. This method has multiple advantages as recording voice data is not invasive and it neither requires specific sensors nor complex calibration processes. It can thus be set up in various environments, outside laboratories, and allows regular and non-restrictive monitoring of patients.

The aim of Vincent Martin thesis is to focus on vocal manifestations of Excessive Daytime Sleepiness (EDS) in order to find vocal biomarkers of these troubles. These biomarkers will then be integrated to clinical measurements collected during interviews by virtual doctors developed at SANPSY. This thesis will take advantage, on one side on the LaBRI skills in speech analysis and machine learning, and on the other side on the expertise of SANPSY on sleepiness troubles for the data collection and the clinical validation of the results.

The proposed schedule of the thesis include the following steps:

- Define vocal biomarkers that can describe the troubles induced by EDS, in close relationship with clinicians at SANPSY.
- Study these biomarkers for their validation in an automatic classification system using data collected at the Bordeaux hospital.
- Implement this approach in the ”virtual doctor” environment developed by SANSPY and carry out clinical validation to validate our method for the diagnosis and follow-up of patients.

IV.2 The MSLT Database

During the course of Vincent thesis, we recorded a new database, the MSLT database. It has been elaborated and is collected at the Bordeaux University Hospital Sleep Clinic, France. All the recorded patients are suspected of excessive daytime sleepiness or nocturnal breathing disorders. A summary of the database is presented in Table IV.2.2.

IV.2.1 Procedure of the MSLT

The procedure of the MSLT is the following. The patients are welcomed the evening prior to the exam for a first night of polysomnography. The day of the exam, they are asked to take a nap every two hours at 9am, 11am, 1pm, 3pm and 5pm. Approximately ten minutes before the beginning of the exam, the voice of the patients is recorded and they fill the Karolinska Sleepiness Scale Questionnaire - KSS [Akerstedt and Gillberg, 1990]. After completing the Cartoon Faces scale lights are switched off and the test begins. The patients have a 20 minutes period to fall asleep: if they stay awake during this period, the test is terminated. If they fall asleep, the recording is extended for a 15 minutes period. After that the lights are turned off, one epoch of any sleep stage is required to define sleep onset [Littner et al., 2005]. The maximum length of the sleep onset period being 20 minutes, all the MSLT values are under or equal to 20 minutes. This procedure is summarised in the Figure IV.2.1.

IV.2.2 Recording procedure

Each patient reads six different texts that are the same at constant session. These texts are presented in Section IV.2.3 The first text is read during the reference recording (Session 0, see Figure IV.2.1), carried out the day before the exam around 6pm, time at which the circadian cycle is at its apex [Sedgwick, 1998]. This recording allows the patient to familiarise with the procedure and the material. The recording procedure is the following. First, the patients are asked to quietly read the text, to acquaint themselves with the content of the text. Second, the patients fill the KSS questionnaire. Third, they are asked to read the text aloud and their voice is recorded. This procedure is the same for each iteration of the test. All the recordings are made in the room in which the patient takes the test, with an omni-directional Audio-technica AT4022 microphone connected to a Tascam DR-100 MKIII audio recorder. To ensure minimum alteration of the recordings due to environment and position of the vocal apparatus, the patients are either in their bed or installed at their desk, the positions of the patient and the microphone being the same for all the iterations. This procedure is summarised in Figure IV.2.2

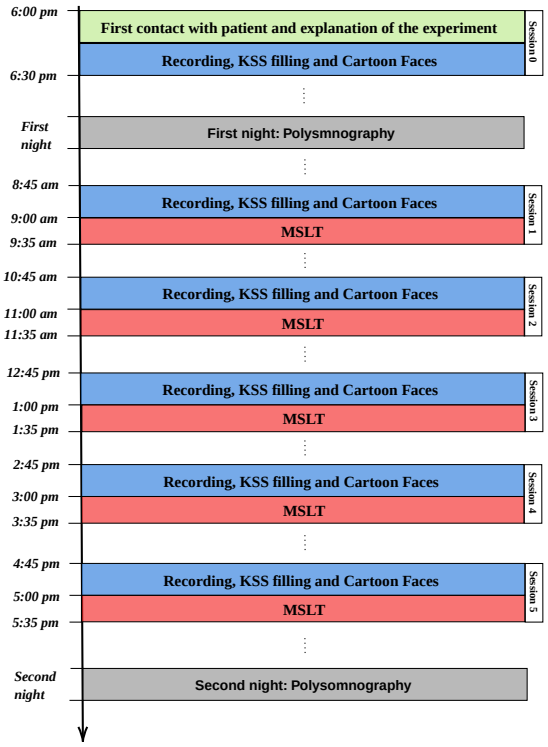


FIGURE IV.2.1: Typical time table of a patient during the recording of the MSLT database.

IV.2.3 Read texts

This section presents the choice of the tasks chosen to record voice of the subjects and justify the choice of the texts.

IV.2.3.1 Reading tasks

As our subjects are patients, they could have untreated hypersomnia. This could lead to difficulties to carry out tasks involving a high cognitive load. As reading has a lower cognitive load than spontaneous speech [Christodoulides, 2016], we choose to focus on reading tasks. Furthermore, such a task assures valid comparison between patients since

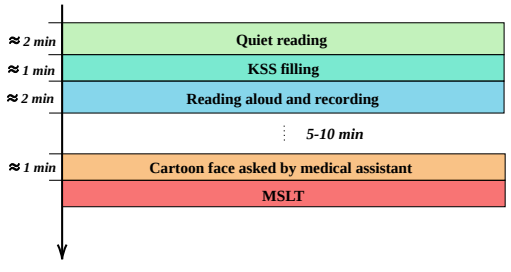


FIGURE IV.2.2: Detail of the procedure to record the voice of patients before a MSLT iteration

all the patients are asked to read the same texts. The recordings should also be less contaminated by emotions compared with spontaneous speech.

IV.2.3.2 Choice of the texts

The texts have to be as neutral as possible regarding the emotional state of the patients (neither boring nor too exiting) to avoid an alteration of their sleepiness state. This constraint is completed by the need of simple grammar and vocabulary, allowing readers with different reading skills. As it is already widely used in phonetic studies [Goldman et al., 2016, Raake, 2002], we choose extracts from *Le Petit Prince* by Antoine de Saint-Exupéry. The first chapters of the French version are truncated so as to be approximately 200 words long while keeping the coherence of the meaning of the text. We thus extracted six texts corresponding to the six iterations presented in Section IV.2.1 The texts are printed with *Times New Roman* font with a 15pt size to ensure a good readability by all patients. Using this protocol, the length of the recordings varies between 50 seconds and 2 minutes (mean duration: 80.8 seconds, std: 21.5 seconds).

IV.2.3.3 Reading level of the patients in the MSLT database

To take into account the reading skills of the patients and the difficulties of the texts, the ELFE score (*Évaluation de la Lecture en FluencE* - Evaluation of the reading with fluency) developed in [Cogniscience, 2008] is measured for each reading. It consists on subtracting the number of mistakes not handled by the patient to the number of words correctly read in one minute. Since the variations of the ELFE score may be influenced by the session, the sex, the subjective sleepiness state (KSS) but also by the texts, which may not have the same difficulty level. To study these different influences, we conduct a multivariate ANOVA taking into account the KSS, the difficulty of each text and the sex of the subjects to explain the variations of the ELFE score. The influence of the session is dominant ($F = 49.0$, $p < 10^{-16}$), but the KSS ($F = 2.4$, $p < 10^{-2}$) and the sex (cross-interaction between sex and KSS, $F = 2.4$, $p < 5 \times 10^{-2}$) also have an influence on the ELFE score. We assume that the major differences between the ELFE distributions across the experience are only due to variations of subjective sleepiness and sex, the minor differences in difficulty level having a negligible effect.

Furthermore, a Spearman's ρ led to the conclusion that the mean ELFE score is directly correlated to the social level of the reader ($\rho = 0.34$, $p = 0.0008$). As women recorded in this database have a higher social level than men (see last line of Table IV.2.2), the influence of the sex factor over the ELFE score may be explained by a difference of social level.

IV.2.4 Medical data

Sleepiness estimation faces two main challenges. On the one hand, as most our subjects are patients with Excessive Daytime Sleepiness, their objective sleepiness measured by EEG does not necessarily correlates with their perceptual sleepiness. On the other hand, the temporal granularity of the sleepiness estimation varies from one questionnaire to another. To study the different parameters influencing voice production, the database includes both subjective and objective measures, at two different time levels: few minutes before the MSLT iteration (designated as *MSLT iteration scale*) and the habits of the patients on several days or weeks before the test (designated as *Patients scale*). As this

Category of **questionnaire**	**Questionnaires**	**Reference**	**Description**
Objective sleepiness measure			
Sleepiness	MSLT value	[Littner et al., 2005]	Time (in min) between beginning of the test and sleeping onset (0-20 min)
Subjective sleepiness measures (MSLT iteration scale)			
Sleepiness	KSS	[Akerstedt and Gillberg, 1990]	9 items about sleepiness (1-9)
	Cartoon Faces	[Maldonado et al., 2004]	5 graphical items about sleepiness (0-4)
Subjective sleepiness and co-morbidity factors measures (Patient scale)			
Sleepiness Fatigue	Epworth Sleepiness Scale (ESS)	[Shahid, 2012, p.149]	8 items about daytime sleepiness (0-24)
	Insomnia Severity Index (ISI)	[Shahid, 2012, p.191]	7 items about insomnia (0-28)
	Functional Outcomes of Sleep Questionnaire-10 (FOSQ-10)	[Shahid, 2012, p.179]	10 items about the impact of daytime sleepiness on activities of daily living (10-40)
	Fatigue Severity Scale (FSS)	[Shahid, 2012, p.167]	9 items about fatigue (9-63)
	Toronto & Hospital Alertness Test	[Shahid, 2012, p.391]	10 items to measure alertness (0-50)
	Part A of ADHD Self-Report Scale (ASRS)	[Schweitzer et al., 2001]	6 items about attention-deficit/hyperactivity disorder (0-24)
	Barcelona Scale	[Guaita et al., 2015]	2 items about sleepiness (0-6)
	Hobson Scale	[Hobson et al., 0030]	4 items about excessive daytime sleepiness (0-16)
Anxiety and Depression	Hospital Anxiety and Depression scale	[Zigmond and Snaith, 1983]	7 items about depression 7 items about anxiety (0-21)
Alcohol	Cut-down, Annoyed, Guilty, Eye-opener Questionnaire (CAGE)	[Shahid, 2012, p.415]	4 items about alcohol consumption (0-4)
Cigarettes	Cigarette Dependence Scale, short version (CDS-5)	[Courvoisier and Etter, 2008]	5 items about cigarettes dependence (5-25)
Social level measures			
Reading Level	Évaluation de la Lecture en FluencE (ELFE)	[Cogniscience, 2008]	Number of words read in one minute minus the number of errors
Education level	-	-	Years of study after the French Certificate of general education

TABLE IV.2.1: Medical information (patient scale and MSLT iteration scale) collected for the database.

database has been elaborated in France, all the questionnaires mentioned in this article are in French. A summary of the database is proposed in Table IV.2.2.

IV.2.4.1 MSLT and subjective sleepiness scales

During the MSLT test, three sleepiness measures are collected, then averaged over the five iterations (Session 1-5) of the protocol explained in Section IV.2.1

Subjective sleepiness scales

Two perceptual questionnaires are filled by the patient during the interview before each MSLT iteration: the KSS and the Cartoon Faces. The KSS is a nine items questionnaire going from 1: 'extremely alert' to 9: 'Very sleepy, great effort to keep awake, fighting sleep' (resp. 'Très éveillé' and 'Très somnolent, avec de grands efforts pour rester éveillé, luttant contre le sommeil' in French). It is the most used sleepiness questionnaire in studies about influence of sleepiness on speech [Schuller et al., 2011, Schuller et al., 2019] and has already be proved a confident measure of subjective sleepiness [Akerstedt et al., 2014].

The Cartoon Faces Sleepiness Scale consists on five cartoon faces reflecting five different states of sleepiness. It has the advantage of not necessitating the comprehension of any language and is easier and more intuitive to answer when, for example, dealing with patient having severe sleep disease.

Clinical Data (MSLT scale)

These two subjective measures are completed by the objective measure of EEG during the iteration of the MSLT, providing the time needed by the patient to fall asleep after the beginning of the test. In the following, this measure will be denominated 'MSLT iteration value'.

IV.2.4.2 Medical questionnaires (Patient scale)

Multiple questionnaires and medical measures are collected about the patient to take into account two aspects of the challenge. On the one hand, all the physiological parameters that could affect the vocal production are measured and integrated to the database. On the other hand, medical questionnaires that allow the estimation of the different components of sleepiness, fatigue and depression are collected. These clinical measures are completed with a Polysomnography the night preceding the exam, the collect of the pathologies and the treatments that can affect voice (psychostimulants, myorelaxants, ...) and diverse physiological measures such as height, weight, age, neck size, ... The clinical data collected anonymously in this study are presented in Table IV.2.1.

Questionnaires	Mean MSLT $\leq$ 8 (SL)			Mean MSLT $>$ 8 (NSL)			All		
	M	F	Both	M	F	Both	M	F	Both
Objective sleepiness measure									
mean MSLT (0-20)	5.0 (1.9) ****	4.5 (2.2) ****	4.8 (2.0) ****	12.9 (3.7) ****	13.6 (3.2) ****	13.4 (3.4) ****	9.8 (4.9) **	12.3 (4.4) **	11.3 (4.8)
Physiological measures									
Number	15	8	23	23	48	71	38	56	94
Age	37.8 (18.2)	29.9 (6.7)	35.0 (15.7)	37.8 (15.1)	36.1 (12.1)	36.6 (13.2)	37.8 (16.4)	35.2 (11.7)	36.3 (13.9)
Body Mass Index	25.7 (4.1)	25.6 (5.7)	25.6 (4.7) *	26.6 (4.8) ***	23.3 (6.0) ***	24.4 (5.8) *	26.2 (4.5) ***	23.6 (6.0) ***	24.7 (5.6)
Height (m)	1.76 (0.05) ***	1.64 (0.06) ***	1.72 (0.08)	1.78 (0.05) ****	1.64 (0.06) ****	1.69 (0.09)	1.77 (0.05) ****	1.64 (0.06) ****	1.69 (0.09)
Weight (kg)	79.2 (12.7)	69.3 (17.9)	75.8 (15.4) *	84.3 (14.7) ****	62.4 (13.5) ****	69.5 (17.3) *	82.3 (14.2) ****	63.4 (14.4) ****	71.0 (17.1)
Neck size (cm)	41.8 (3.2) **	36.1 (3.1) **	39.8 (4.2) *	41.8 (3.2) ****	35.6 (3.4) ****	37.6 (4.5) *	41.8 (3.2) ****	35.6 (3.4) ****	38.1 (4.5)
Cigarettes/day	1.5 (3.4)	2.8 (7.3)	2.0 (5.1)	1.5 (4.0)	2.3 (6.3)	2.0 (5.7)	1.5 (3.8)	2.4 (6.5)	2.0 (5.6)
Alcohol glasses/day	0.2 (0.5) *	0.0 (0.1)	0.1 (0.4)	0.6 (1.1) *	0.1 (0.3)	0.3 (0.7)	0.4 (0.9) *	0.1 (0.3) *	0.2 (0.6)
Subjective sleepiness measures (MSLT iteration scale)									
Mean KSS (1-9)	4.0 (1.1) *	5.2 (1.1) *	4.4 (1.2)	4.6 (1.5)	4.5 (1.3)	4.5 (1.3)	4.4 (1.4)	4.6 (1.3)	4.5 (1.3)
Mean Cartoon Faces (0-5)	1.5 (0.6)	1.8 (0.5)	1.6 (0.6)	1.6 (0.6)	1.7 (0.6)	1.6 (0.6)	1.6 (0.6)	1.7 (0.6)	1.6 (0.6)
Subjective sleepiness and co-morbidity factors measures (Patient scale)									
Fatigue	11	7	18	19	45	64	30	52	82
Snoring	3	3	6	7	10	17	10	13	23
Hypertension	3	2	5	5	2	7	8	4	12
Observed Sleepiness Apnea	4	2	6	6	7	13	10	9	19
ESS (0-24)	14.9 (5.1)	17.9 (4.6)	16.0 (5.1)	12.8 (6.0)	14.8 (4.4)	14.1 (5.0)	13.6 (5.7)	15.2 (4.5)	14.6 (5.1)
ISI (0-28)	13.3 (5.3)	14.4 (5.5)	13.7 (5.4)	16.0 (5.3)	15.1 (5.3)	15.4 (5.3)	14.9 (5.5)	15.0 (5.3)	14.9 (5.4)
FOSQ-10 (10-40)	25.1 (7.5)	20.0 (8.3)	23.3 (8.1)	21.7 (5.3)	21.3 (7.5)	21.4 (6.8)	23.0 (6.5)	21.1 (7.6)	21.9 (7.2)
FSS (9-63)	35.0 (10.7) ***	49.0 (10.6)	39.9 (12.6) ***	49.7 (10.4) ***	49.3 (11.2)	49.4 (11.0) ***	43.9 (12.8) *	49.3 (11.1) *	47.1 (12.1)
Toronto (0-50)	28.2 (8.8) **	24.7 (7.5)	27.0 (8.5) **	21.7 (6.5) **	22.9 (8.2)	22.5 (7.7) **	24.3 (8.1)	23.1 (8.1)	23.6 (8.2)
ASRS (0-24)	10.9 (5.7)	12.1 (4.5)	11.3 (5.3)	13.9 (5.2)	11.8 (4.9)	12.5 (5.1)	12.7 (5.6)	11.8 (4.8)	12.2 (5.2)
Barcelona (0-6)	2.4 (1.3)	2.4 (1.3)	2.4 (1.3)	2.2 (0.9)	2.3 (0.9)	2.3 (0.9)	2.3 (1.1)	2.3 (1.0)	2.3 (1.0)
Hobson (0-12)	4.3 (2.4)	5.9 (3.2)	4.8 (2.8)	3.8 (2.5)	4.1 (2.3)	4.0 (2.4)	4.0 (2.4)	4.3 (2.6)	4.2 (2.5)
HAD Depression (0-21)	4.5 (3.2) **	4.8 (4.4)	4.6 (3.7) **	7.1 (2.7) **	7.0 (4.6)	7.0 (4.1) **	6.1 (3.2)	6.7 (4.6)	6.4 (4.1)
HAD Anxiety (0-21)	6.3 (3.0) *	8.0 (3.4)	6.9 (3.2)	9.0 (4.6) *	8.3 (4.0)	8.5 (4.2)	8.0 (4.3)	8.2 (3.9)	8.1 (4.1)
CAGE 0-4	0.3 (1.0) *	0.2 (0.4)	0.3 (0.8)	0.7 (0.9) *	0.2 (0.6)	0.4 (0.7)	0.5 (0.9)	0.2 (0.6)	0.4 (0.8)
CDS-5 5-25	6.5 (3.4)	7.1 (5.6)	6.7 (4.3)	6.6 (3.5)	7.3 (4.9)	7.0 (4.5)	6.5 (3.5)	7.2 (5.0)	7.0 (4.4)
Social level measures									
Mean ELFE	176.2 (31.0)	176.2 (41.5)	176.2 (35.0)	167.9 (32.8) **	188.6 (31.8) **	181.9 (33.6)	171.2 (32.4) **	186.8 (33.6) **	180.5 (34.0)
Education level	3.9 (2.2)	4.8 (1.4)	4.2 (2.0) *	4.5 (2.3) *	5.7 (2.6) *	5.3 (2.6) *	4.3 (2.2) **	5.5 (2.5) **	5.0 (2.5)

TABLE IV.2.2: Summary of the data collected for the database. The different colors represent the result of Mann-Whitney tests. Green: Sig. Difference between sex ind. from the sleepiness level. Red: Sig. Difference between sleepiness group ind. from the sex. Blue: (resp. Orange) Difference between Sleepy and Non-Sleepy men (resp. women).
(*: $p < 5 \times 10^{-2}$, **: $p < 10^{-2}$, ***: $p < 10^{-3}$, ****: $p < 10^{-4}$)

IV.3 Features

During the duration of the thesis, we investigated three different kinds of features. First, we used acoustic or low-level features, extracted directly from the speech signal. Then, in accordance with speech therapists, we took advantage of the reading task we used to study the impact of the sleepiness level on reading errors. Finally, in order to automatise the same process, we derive features from the end-to-end automatic speech recognition system errors.

IV.3.1 Acoustic features

We wish to use features that can be understood and used by physicians. As a consequence, several features are extracted on two time scales. On one hand, excerpt level features are computed directly on each recording, using either an automatic vocalic segments detection algorithm [Pellegrino and André-Obrecht, 2000] or voiced segments detected using a fundamental frequency extraction algorithm [Sjölander, 2004]. On the other hand, other features are computed on each voiced segment to characterise the regularity of production of harmonic sounds. These features are averaged for each recording.

IV.3.1.1 Excerpt level features

The statistics on the duration/proportion of voiced segments or automatically detected vowels should reflect the global behaviour of the speaker. An example of excerpt level feature extraction is given on Figure IV.3.1.

The features extracted using this time-frame paradigm are:

- durvoiced: the total duration of voiced parts (in s.);
- pervoiced: the percentage in duration of voiced parts;
- durvowel: the total duration of vocalic segments (in s.);
- pervowel: the percentage in duration of vocalic segments.

This feature set provides 4 features per recording.

IV.3.1.2 Voiced segments features

The voiced segment feature extraction is illustrated on Figure IV.3.2. These features include measurements on the fundamental frequency and intensity curves:

- F0MEAN: mean of fundamental frequency over a voiced segment;
- F0VAR: variance of fundamental frequency over a voiced segment;
- F0SLOPE: slope of the linear approximation of the fundamental frequency over a voiced segment;

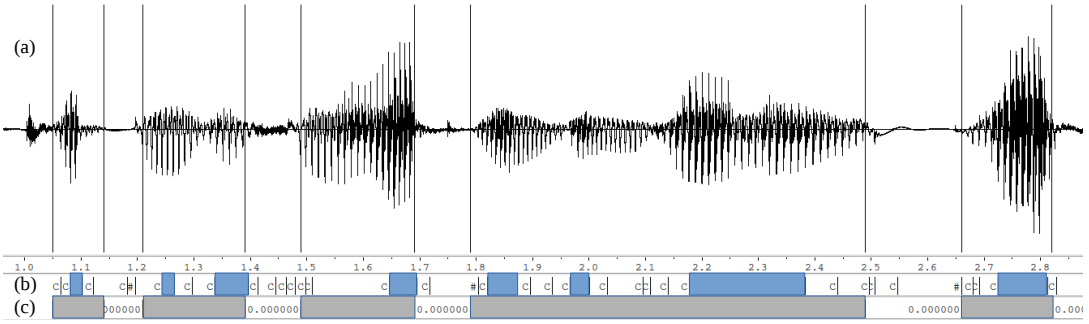


FIGURE IV.3.1: Illustration of the result of the pre-processing steps for excerpt level features on the sentence "... sich Nordwind und Sonne, wer...": (a) signal, (b) vocalic segments (c) voiced segments.

- F0MAX: maximum of fundamental frequency over a voiced segment;
- F0MIN: minimum of fundamental frequency over a voiced segment;
- F0EXTEND: extend of fundamental frequency values over a voiced segment.

The same features are computed on the intensity curve (NRJMEAN, NRJVAR, NRJMAX, NRJMIN, NRJEXTEND). This results in 12 more features (6 on F0, 6 on intensity). We also computed the F0MEAN, F0VAR, NRJMEAN and NRJVAR features on vocalic segments, adding 4 features to the set.

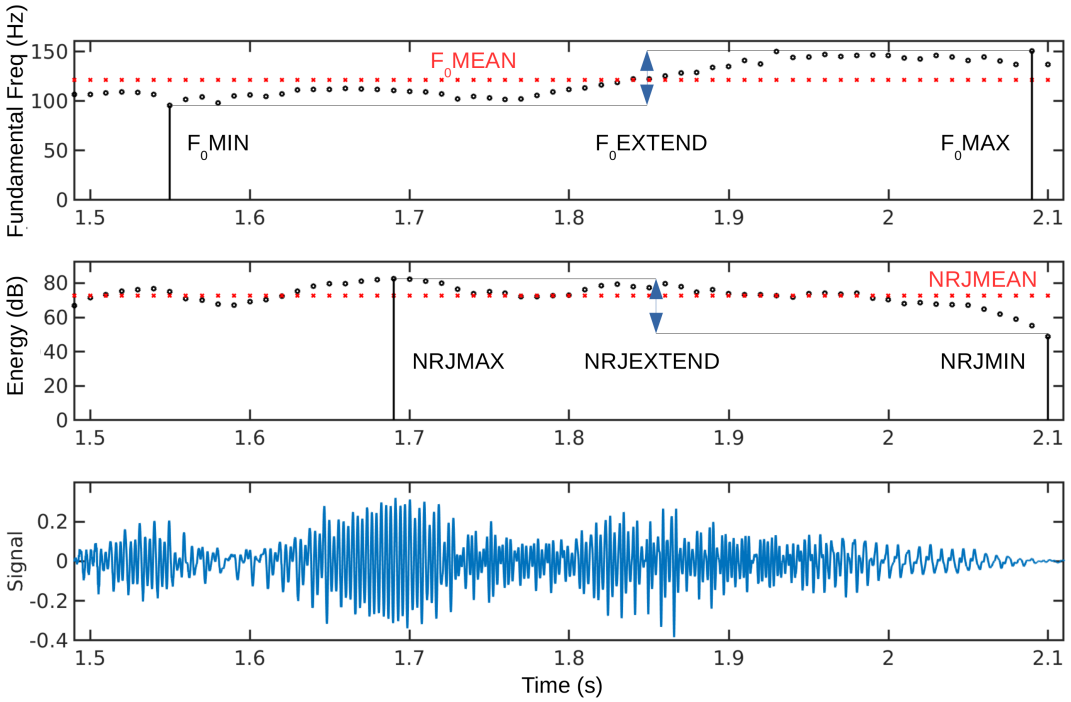


FIGURE IV.3.2: Illustration of the extraction of features on a voiced segment. Upper pane: fundamental frequency in Hz. Middle pane: intensity in dB. Bottom pane: signal.

Furthermore, additional features are computed using the COVAREP matlab toolkit [Degottex et al., 2014] which we modified to add some features and compute them

only on voiced segments. These features have already been used for characterising singing styles [Rouas and Ioannidis, 2016] and for spoken social attitudes classification [Rouas et al., 2019]. They are the following ones: harmonics amplitude (H1, H2, H4), formants amplitude (A1, A2, A3), frequencies (F1, F2, F3, F4) and bandwidth (B1, B2, B3, B4), differences between harmonics amplitude (H1-H2, H2-H4), differences between amplitude of harmonics and formants (H1-A1, H1-A2, H1-A3), cepstral peak prominence (CPP), harmonics to noise ratios on different frequency bands (HNR05, HNR15, HNR25, HNR35). All these features are averaged over each recording, yielding an additional set of 24 features per recording.

IV.3.2 Reading mistakes

In recent year we had the chance to welcome speech therapist intern students for a small period of time. We submitted them our pathological sleepiness characterisation problem and had them listen to the database we recorded. The main idea that emerged from our common reflexion was to take advantage our the nature of the reading task we use for collecting the data. Their observation is that reading errors may be more common for subjects having sleepiness troubles. In accordance with that observation, we have selected five types of errors that may be represented well enough in our data:

- Stumbling errors (”Achoppements” in French): ”hesitations, breaks in the speech rhythm” [Brin-Henry et al., 2018].

These errors mainly measure the *assembling* capacities of the reader. *Assembling* is the fact to put together independent syllables so as to form a word: when a subject begins to read a word, stops then continue, the process of assembling the word has been interrupted, leading to a stumbling. We have chosen to not take into account hesitations between words (breaks of the speech flow), but only breaks that occur inside words and unnatural vowel lengthening testifying hesitation.

- Paralexia (”Paralexies” in French) : ”identification error of written words consisting in the production of a word instead of another” [Brin-Henry et al., 2018].

Contrary to stumbling errors, paralexia reflect the *addressing* capacities of the reader. Contrary to *assembling*, *addressing* can be defined as the fact to read a word wholly, without deciphering it or slicing it into syllables. Paralexia are symptomatic errors involving this type of reading. We have generalised this category to the pronunciation of any other word, existing or not, that is read instead of the correct one. For example, collapsing errors (the deletion of one syllable in a word) are counted as paralexia in this study. The pronounced word has however to be similar to the expected one, to differentiate this error from additions and deletions of words.

- Deletions of words: this error occurs when the speaker forgets to pronounce a word and goes directly to the next one. Even if self-correction occurs afterwards, the deletion error is counted.

- Additions of words: this error occurs when the speaker adds a word that is not present in the text. Even if self-correction occurs afterwards, the addition error is taken into account.

- Syntactic reversals: this error occurs when words in a sentence are inverted.

If a paralexia, deletion, addition or syntactic reversal error has already been counted, self-correction results in not taking into account an additional stumbling error, except if the patient mistakes during its resumption.

IV.3.3 Automatic Speech Recognition errors

Annotating the previous errors is time-consuming and requires training to differentiate errors. In an attempt to automatize the labeling of reading errors, we measured the errors made by ASR systems. Indeed, when subjects are sleepy, their articulation and prosody are impaired [Krajewski et al., 2012] while the number of hesitations and repeats increases. This alteration of speech due to sleepiness may induce errors in ASR systems that could be used as biomarkers of sleepiness. Thanks to recent advances on end-to-end ASR systems allowing intermediate transcription units such as characters or tokens, it is now possible to transcribe not only words but also portions of words (Byte Pair Encoding – BPE).

In this study, we use an end-to-end system using RNN transducters with attention, based either on words, BPE, or characters, to transcribe words or BPE. The language model is trained on a word, BPE, or character version of the ESTER corpus [Galliano et al., 2009]. A complete review of such systems and their performances is proposed in [Boyer and Rouas, 2019]. The end-to-end system achieving the best performances is the character-based one with a word-based RNN language model achieving 17.6% of Word Error Rate on the ESTER corpus.

We consider two types of errors in this study: insertions and substitutions. Each type of error is computed on tokens and on words, and we consider both the raw number of errors and their proportion over the total number of transcription units, leading to 8 features per system.

IV.4 Experiments

IV.4.1 Speaker selection (exclusion criteria)

Labelling the database with the errors described in section IV.3.2, some reading profiles have been drawn to exclude subjects from this study. Admittedly, excluding speakers can reduce the size of the corpus but this however ensures that the computed vocal features and reading mistakes are mostly linked to sleepiness, excluding the influence of pathologies and reading disorders on these markers.

First, we kept away the patients that have medical history of stroke or transient ischaemic attacks: the errors made by these patients could possibly due to sequelae of these events (alexia or visuo-attentional disorder are common sequelae of these pathologies). Based on this criteria, we have excluded three patients whom had a very slow reading flow and produced a lot of errors. In the same vein, patients with current neuromuscular pathologies (e.g. disphonia, myotonia, Huntington's chorea, epilepsy) are also excluded from the database: involuntary muscular contraction or lack of muscular control are likely to lead to the observation of numerous stumbling errors. Based on this criteria, three additional patients (respectively suffering from epilepsy, spasmodic disphonia and Steinert myotony) have been excluded.

Other criteria concerning more specifically the reading abilities include the Attention deficit disorder (associated or not to hyperactivity). As a matter of fact, as attention plays an important role when reading, these patient may skip words or lines. It has to be noted that sleepiness may also be the cause of such behaviour, but differentiating the origin of these mistakes may prove difficult. As a precaution, we therefore choose to exclude patients that produce such reading errors. Two patients having skipped a row and another being diagnosed with Attention deficit disorder, the three presenting incoherent readings, their recordings were removed from the study. Concerning fluency, we took into account disorders such as stuttering or cluttering as they are difficult to differentiate from the errors induced by sleepiness One patient presenting characteristics of cluttering has therefore been excluded from our corpus. Finally, we also excluded three patients suffering from dyslexia or disorder implying the deciphering of the text despite having read it silently a few minutes before.

One supplementary patient suffering from different serious inflammatory diseases (Crohn, Basedown and Ankylosing Spondylitise diseases) and not presenting a satisfactory reading level has been excluded.

TABLE IV.4.1: Distribution of the speakers across Sex and Sleepiness class.
SL: Sleepy (MSLT $\leq$ 8 min.), NSL: Non-Sleepy (MSLT $>$ 8 min.)

Sex	SL	NSL	TOTAL
Women	10	48	58
Men	11	24	35
TOTAL	**21**	**72**	**93**

Excluding these recordings, we have kept a total of 93 speakers out of the 105 original ones.

After having excluded the previously mentioned patients of the database, concise statistics concerning the studied subjects are presented in Table IV.4.1. We clustered patients between Sleepy (SL) and Non-Sleepy (NSL) using the 8 minutes medical limit on the mean MSLT value used to assess narcolepsy [Aldrich et al., 1997, Littner et al., 2005].

IV.4.2 Classification pipeline

IV.4.2.1 Feature selection

For each speaker, each previously pre sented set of features is computed on each of the 5

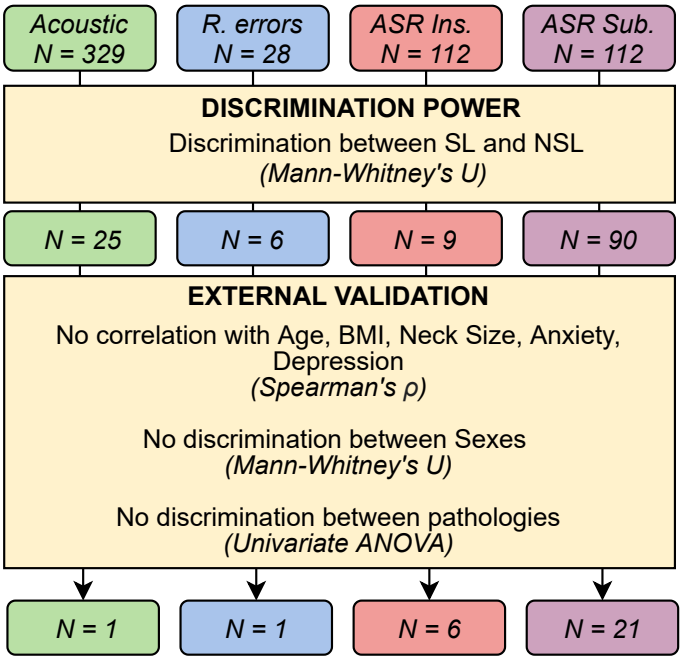


FIGURE IV.4.1: Feature selection pipeline. R. errors: reading errors; ASR: Automatic Speech Recognition errors, Ins: insertions; Sub: substitutions. SL: Sleepy (MSLT $\leq$ 8 min.), NSL: Non-Sleepy (MSLT $>$ 8 min.). BMI: Body Mass Index

naps of the MSLT, to which we aggregate the mean and the standard deviation across the naps, resulting in 7 measures for each feature and speaker. The proposed feature selection pipeline is represented in Figure IV.4.1. The selected features have to comply with two constraints:

- Their distribution for each sleepiness class, measured by a Mann-Whitney's U ($p < 0.05$) has to be statistically different (discrimination power of the features);
- They should not correlate (Spearman's ρ, $p > 0.05$) with age, Body Mass Index (BMI), neck size, anxiety or depression (measured by the Hospital Anxiety and Depression scale [Zigmond and Snaith, 1983]); they should not discriminate sex (Mann Whitney's U, $p > 0.05$) or pathologies (Univariate ANOVA, $p > 0.05$).

This pipeline, inspired by the external validation of psychometric questionnaires, ensures that the selected features classify only sleepiness, independently from the other measured

speaker traits. Indeed, even if some of these factors could correlate with sleepiness, our aim is to train the classifier to learn the concept of sleepiness, not to learn a confounding factor correlating with sleepiness, that could still give good classification performances but make the interpretation of such results impossible.

Moreover, compared with performance-driven feature selection pipelines, this one works with few samples – statistical tests do not require large amounts of data – and is independent of the chosen metric: however the performances of the system are measured, the selected features remain the same.

IV.4.2.2 Classification

To ensure generalization and avoid overlearning, the classification is carried out under Leave One Speaker Out Cross-Validation: each speaker is turn-by-turn isolated as a test speaker, while the classification system is trained on the others. Estimated and ground-truth sleepiness classes of the test speaker are stacked and the classification metrics are computed on this aggregation.

As the goal of this study is not to optimize the best possible classifier but to validate the use of new features for objective sleepiness estimation through voice, the previously selected features are aggregated (early-fusion), scaled, orthogonalized by a Principal Components Analysis (PCA) and classified by a logistic regression using the Python module sci-kit learn [Pedregosa et al., 2011] with a *newton-cg* solver and a balanced class-weighting.

IV.4.3 Results

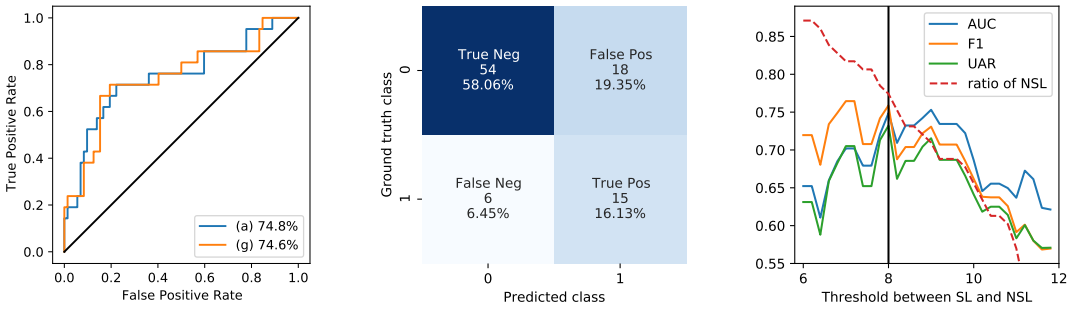
IV.4.3.1 Classification performances

TABLE IV.4.2: Classification performances of the proposed pipeline. UAR: Unweighted Average Recall, F1-score: class-weighted average F1-score, AUC: Area Under the ROC Curve.
R. errors: Reading errors, ASR: Automatic Speech Recognition system errors.

	Features	UAR	F1	AUC
(a)	ASR	**73.2%**	**75.8%**	**74.8%**
(b)	R. errors	57.7%	73.4%	22.1%
(c)	Acoustic	59.5%	66.0%	60.1%
(d)	ASR + R. errors	71.8%	74.0%	74.5%
(e)	ASR + Acoustic	73.2%	75.9%	74.4%
(f)	R. errors + Acoustic	61.3%	70.2%	67.7%
(g)	All	73.9%	76.8%	74.6%

The obtained Unweighted Average Recall (UAR), weighed F1-score, and Area Under the ROC Curve (AUC) for the different feature combinations are presented in Table IV.4.2.

The best performances are obtained by the system (g), aggregating the three sets of features: this system achieves 73.9% of UAR, 76.8% of weighted F1-score, and 74.6% of AUC. In this system, the selected reading error is the number of additions on the fourth nap, and the selected acoustic features are the bandwidth of the first Formant on the first nap. The selected ASR errors are detailed below.



(a) (b) (c)

FIGURE IV.4.2: **(a)** ROC of the system (a) and (g) and their corresponding AUC. **(b)** Confusion matrix of the system (a) **(c)** Performances of the system (a) depending on the threshold between Sleepy – SL – and Non-Sleepy – NSL – classes

However, the ASR features alone (system (a)) achieve classification performances that are only slightly below (73.2% of UAR, 75.8% of F1-score, and 74.8% of AUC): the acoustic and reading errors seem to have little importance on the classification. Regarding the cost-benefices balance of the hand-labeled reading mistakes in comparison with the small performance enhancement they are the cause of when combined with ASR and acoustic features (0.7% of absolute improvement regarding UAR, 1% regarding the weighted F1-score), we choose to discard these features. As the combination between ASR and acoustic features (system (c)) achieves poorer results than ASR features alone, we choose to focus on system (a), based only on ASR features.

IV.4.3.2 Performances of the selected system

The ROC curve of the systems (a) and (g) and the confusion matrix of the system (a) are respectively represented in Figures IV.4.2a) and IV.4.2b). The ROC curve confirms the close similarity of performances of systems (a) and (g) and consolidates the choice to focus on system (a).

Moreover, inspired by a previous study [Martin et al., 2019], we represented in Figure IV.4.2c) the performances of the system (a) depending on the threshold to distinguish SL from NSL. This graph represents the specificity of the selected features to the phenomena they aim at measuring. As intended, the best performances are obtained for a threshold of 8 minutes. Moreover, excepting a crook for 7.5 minutes, these features achieve performances higher than 70% for thresholds between 7.0 minutes and 9.5 minutes, allowing physicians to select the severity of objective EDS they want to detect.

IV.4.4 Discussions

IV.4.4.1 PCA Analysis

Along the cross-validation process, the parameters of the PCA and the weights of the logistic regression are averaged. Figure IV.4.3 represents the four different PCA dimensions and their averaged corresponding weights in the classification. The most important PCA component in the decision of the classifier is the fourth dimension, relying on the

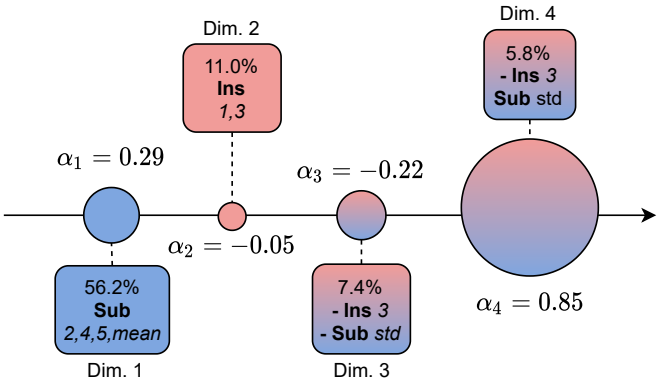


FIGURE IV.4.3: PCA components and their associated weight (α_i) in the logistic regression for the system (a). For each box, from top to bottom: mean ratio of explained variance in the PCA, feature (in bold), nap during which it is measured (in italic).
Red background: Insertions, Blue background: Substitutions. - : negative PCA weight; Sub: substitutions; Ins: insertions. Dim: Dimension

difference between the number of insertions during the third nap and the standard deviation of the substitutions (mean coefficient of the logistic regression: $\alpha_4 = 0.85$). The sum of the same features directs the third dimension of the PCA, which has a weight of $\alpha_3 = -0.22$ in the classifier decision. The insertions errors measured at the third nap are made by a character-based ASR system with a word-based language model, whereas the involved substitutions are made by a BPE-based ASR system without a language model. The second most important dimension is directed by the substitutions measured on second, fourth, and fifth naps, and the mean value across the naps ($\alpha_1 = 0.29$). These errors come from 7 ASR systems based on different units, with and without language models. Finally, the least important PCA dimension relies on the number of insertions on the first and the third naps ($\alpha_2 = -0.05$). Contrary to the insertions of the third nap, the selected insertions of the first nap are made by a BPE-based ASR system without a language model.

IV.4.4.2 Measures of the features

Regarding the selected features, the insertions seem to be relevant precisely on the first and third naps, excluding the others. When studying the selected features after the first step of the validation process, the insertions of the first, third but also fourth nap and their standard deviation across the naps discriminate objective EDS. However, insertions during the fourth nap correlate significantly with the BMI (Spearman's ρ, $p < 0.05$) and their standard deviation across naps discriminate pathologies (Univariate ANOVA, $p < 0.05$). This could be explained by two phenomena. First, the texts are different for each iteration of the MSLT procedure. The ASR systems having an inherent error rate depending on the content of the processed text reading, it may be possible that the third text could be the only one on which the link between insertions and objective EDS is distinguished independently from the other speaker traits. Second, the speakers are recorded five times during the day, in different emotional, fatigue, or circadian states, filtering the expression of the speakers' traits. Indeed, the first recording is made at 9 am after breakfast and patients lunch few minutes before the third nap: the state induced by taking a meal could

favor voice phenomena inducing the ASR system to produce insertions errors linked with sleepiness, but discriminating it from other traits.

Part V

Conclusions and Perspectives

Several aspects of speech analysis have been described in this document. While some low-level acoustic and prosodic descriptors have been used in the first two parts (singing styles using phonation modes identification and social affects characterisation), the possibilities offered by the recent development in speech recognition described in the third section have found some application in the last part dedicated to pathological sleepiness detection. In the future, I wish to continue to bring together different aspects of speech analysis with the incorporation of social affects in the speech recognition/translation domain, to find new applications for speech analysis such as the analysis of caregivers professional speech, while tackling the important problem of dealing with spontaneous speech. These points constitue my future research project and are detailed below:

- Since we started our research on sleepiness detection through speech analysis, the use of voice technologies for medical application became very trendy. But even if the results I presented on sleepiness detection are very encouraging, there is an important drawback for many possible applications: we are still not considering spontaneous speech or "in-the-wild" data collection. In order to fulfill that goal that may also voice technologies to be used in at-home personalised medecine or precision medecine, we need to gather the data in the most ecological way possible. This is the subject of the PRIME interdisciplinary CNRS team we wish to build together with the SANPSY laboratory.

- Take into account the speakers state in order to constrain automatic transcription or translation systems. In the framework of the FVLLMONTI European project, I am collaborating with hardware specialists (dealing with nanowire transistors) in order to create small mobile devices (such as earplugs) able to execute efficiently, in terms of computing as well as energy consumption, deep neural networks. The target application is automatic transcription and translation. One of the novelty of the project will reside in the addition of information about the speakers state, built on my research on social affects characterisation, that could modify the behaviour of the translation system.

- On another topic, I am beginning a project on the study of the speech of caregivers in dependent elderly people facilities. This project is the outcome of a meeting with the creators of "humanitude", a method consisting in formalising the interactions at different levels, touch, gazing and speech. This method, which was created in France, have now an international dimension and is thus used in several countries, including Japan which have an great number of elderly people. In that framework, the first aim is to evaluate the impact of the "humanitude" training of the vocal postures of caregivers by measuring the differences in parameters extracted from speech between experts or novice staff. Moreover this will be a cross-cultural study since we will gather data not only in France but also in Japan thanks to a collaboration with Kyoto University. The results we will obtain on speech aspects will be completed by the outcome of studies on other modalities (gaze and touch) carried out by our Japanese colleagues. Finally, the target application is to build self-training tools that will complement the training from the instructors in order to better answer the demand and be available as self-evaluation tools for already trained staff.

Bibliography

[Airas and Alku, 2007] Airas, M. and Alku, P. (2007). Comparison of multiple voice source parameters in different phonation types. In *Proceedings of Interspeech 2007*. Citeseer.

[Akerstedt et al., 2014] Akerstedt, T., Anund, A., Axelsson, J., and Kecklund, G. (2014). Subjective sleepiness is a sensitive indicator of insufficient sleep and impaired waking function. *Journal of Sleep Research*, 23(3):240–252.

[Akerstedt and Gillberg, 1990] Akerstedt, T. and Gillberg, M. (1990). Subjective and objective sleepiness in the active individual. *The International Journal of Neuroscience*, 52(1-2):29–37.

[Aldrich et al., 1997] Aldrich, M. S., Chervin, R. D., and Malow, B. A. (1997). Value of the multiple sleep latency test (MSLT) for the diagnosis of narcolepsy. *Sleep*, 20(8):620–629.

[Alku, 1992] Alku, P. (1992). Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. *Speech communication*, 11(2-3):109–118.

[Alku, 2011] Alku, P. (2011). Glottal inverse filtering analysis of human voice production ? A review of estimation and parameterization methods of the glottal excitation and their applications. *Sadhana. Academy Proceedings in Engineering Sciences*, 36(October):623–650.

[Alku et al., 2002] Alku, P., Bäckström, T., and Vilkman, E. (2002). Normalized amplitude quotient for parametrization of the glottal flow. *the Journal of the Acoustical Society of America*, 112(2):701–710.

[Alku et al., 1997] Alku, P., Strik, H., and Vilkman, E. (1997). Parabolic spectral parameter - a new method for quantification of the glottal flow. *Speech Communication*, 22(1):67–79.

[Bahdanau et al., 2016] Bahdanau, D., Chorowski, J., Serdyuk, D., Brakel, P., and Bengio, Y. (2016). End-to-end attention-based large vocabulary speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4945–4949.

[Battenberg et al., 2017] Battenberg, E., Chen, J., Child, R., Coates, A., Li, Y. G. Y., Liu, H., Satheesh, S., Sriram, A., and Zhu, Z. (2017). Exploring neural transducers for end-to-end speech recognition. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 206–213.

[Boersma and Weenink, 2012] Boersma, P. and Weenink, D. (2012). Praat: Doing phonetics by computer [Computer program].

[Boyer and Rouas, 2019] Boyer, F. and Rouas, J.-L. (2019). End-to-End Speech Recognition: A review for the French Language. Technical report.

[Boyer et al., 2021] Boyer, F., Shinohara, Y., Ishii, T., Inaguma, H., and Watanabe, S. (2021). A Study of Transducer Based End-to-End ASR with ESPnet: Architecture, Auxiliary Loss and Decoding Strategies. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 16–23.

[Brin-Henry et al., 2018] Brin-Henry, F., Courrier, C., Lederle, E., and Masy, V. (2018). *Dictionnaire d'Orthophonie*. Ortho-Edition.

[Chang and Lin, 2011] Chang, C.-C. and Lin, C.-J. (2011). LIBSVM : A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3).

[Childers and Lee, 1991] Childers, D. G. and Lee, CK. (1991). Vocal quality factors: Analysis, synthesis, and perception. *the Journal of the Acoustical Society of America*, 90(5):2394–2410.

[Chorowski et al., 2015] Chorowski, J., Bahdanau, D., Serdyuk, D., Cho, K., and Bengio, Y. (2015). Attention-based models for speech recognition: 29th Annual Conference on Neural Information Processing Systems, NIPS 2015. *Advances in Neural Information Processing Systems*, 2015-January:577–585.

[Christodoulides, 2016] Christodoulides, G. (2016). *Effects of Cognitive Load on Speech Production and Perception*. PhD thesis, Université Catholique de Louvain.

[Cogniscience, 2008] Cogniscience (2008). E.L.FE - Evaluation de la Lecture en FluencE. Technical report.

[Courvoisier and Etter, 2008] Courvoisier, D. and Etter, J.-F. (2008). Using item response theory to study the convergent and discriminant validity of three questionnaires measuring cigarette dependence. *Psychology of Addictive Behaviors: Journal of the Society of Psychologists in Addictive Behaviors*, 22(3):391–401.

[Cummins et al., 2018] Cummins, N., Baird, A., and Schuller, B. W. (2018). Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning. *Methods*, 151:41–54.

[de Moraes, 2008] de Moraes, J. A. (2008). The Pitch Accents in brazilian portuguese: Analysis by synthesis. In *Proceedings of Speech Prosody*, pages 389–397.

[Degottex et al., 2014] Degottex, G., Kane, J., Drugman, T., Raitio, T., and Scherer, S. (2014). COVAREP - A collaborative voice analysis repository for speech technologies. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference On*, pages 960–964. IEEE.

[Drugman and Dutoit, 2009] Drugman, T. and Dutoit, T. (2009). Glottal closure and opening instant detection from speech signals. In *Interspeech*, pages 2891–2894.

[Drugman et al., 2012] Drugman, T., Thomas, M., Gudnason, J., Naylor, P., and Dutoit, T. (2012). Detection of glottal closure instants from speech signals: A quantitative review. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(3):994–1006.

[Estève et al., 2010] Estève, Y., Bazillon, T., Antoine, J.-Y., Béchet, F., and Farinas, J. (2010). The EPAC Corpus: Manual and Automatic Annotations of Conversational Speech in French Broadcast News. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

[Feugère et al., 2015] Feugère, L., Doval, B., and Mifune, M.-F. (2015). Using pitch features for the characterization of intermediate vocal productions. In *Proc of 5th International Workshop on Folk Music Analysis (FMA),*.

[Fujisaki and Hirose, 1993] Fujisaki, H. and Hirose, K. (1993). Analysis and perception of intonation expressing paralinguistic information in spoken Japanese. In *ESCA Workshop on Prosody*.

[Galliano et al., 2005] Galliano, S., Geoffrois, E., Mostefa, D., Choukri, K., Bonastre, J.-f., and Gravier, G. (2005). The ESTER phase II evaluation campaign for the rich transcription of French broadcast news. In *In Proceedings of the 9th European Conference on Speech Communication and Technology (INTERSPEECH '05*, pages 1149–1152.

[Galliano et al., 2009] Galliano, S., Gravier, G., and Chaubard, L. (2009). The ester 2 evaluation campaign for the rich transcription of french radio broadcasts. In *In In: Proceddings of Interspeech, Brighton (United Kingdom*.

[Garnier et al., 2007] Garnier, M., Henrich, N., Castellengo, M., Sotiropoulos, D., and Dubois, D. (2007). Characterisation of voice quality in western lyrical singing: From teachers' judgements to acoustic descriptions. *Journal of interdisciplinary music studies*, 1(2):62–91.

[Garnier et al., 2008] Garnier, M., Wolfe, J., Henrich Bernardoni, N., and Smith, J. (2008). Interrelationship between vocal effort and vocal tract acoustics: A pilot study. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 2008, pages 2302–2305.

[Goldman et al., 2016] Goldman, J.-P., Honnet, P.-E., Clark, R., Garner, P. N., Ivanova, M., Lazaridis, A., Liang, H., Macedo, T., Pfister, B., Ribeiro, M. S., Wehrli, E., and Yamagishi, J. (2016). The SIWIS Database: A Multilingual Speech Database with Acted Emphasis. In *Interspeech 2016*, pages 1532–1535. ISCA.

[Graves, 2012] Graves, A. (2012). Sequence Transduction with Recurrent Neural Networks. *arXiv:1211.3711 [cs, stat]*.

[Graves et al., 2006] Graves, A., Fernández, S., Gomez, F., and Schmidhuber, J. (2006). Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 369–376, New York, NY, USA. Association for Computing Machinery.

[Graves et al., 2013] Graves, A., Mohamed, A.-r., and Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6645–6649, Vancouver, BC, Canada. IEEE.

[Grichkovtsova et al., 2012] Grichkovtsova, I., Morel, M., and Lacheret, A. (2012). The role of voice quality and prosodic contour in affective speech perception. *Speech Communication*, 54(3):414–429.

[Grillo and Verdolini, 2008] Grillo, E. U. and Verdolini, K. (2008). Evidence for distinguishing pressed, normal, resonant, and breathy voice qualities by laryngeal resistance and vocal efficiency in vocally trained subjects. *Journal of voice : official journal of the Voice Foundation*, 22(5):546–552.

[Gu et al., 2011] Gu, W., Zhang, T., and Fujisaki, H. (2011). Prosodic analysis and perception of mandarin utterances conveying attitudes. In *INTERSPEECH*, pages 1069–1072. ISCA.

[Guaita et al., 2015] Guaita, M., Salamero, M., Vilaseca, I., Iranzo, A., Montserrat, J. M., Gaig, C., Embid, C., Romero, M., Serradell, M., León, C., de Pablo, J., and Santamaria, J. (2015). The Barcelona Sleepiness Index: A New Instrument to Assess Excessive Daytime Sleepiness in Sleep Disordered Breathing. *Journal of Clinical Sleep Medicine : JCSM : Official Publication of the American Academy of Sleep Medicine*, 11(11):1289–1298.

[Guerry et al., 2016] Guerry, M., Rilliard, A., Erickson, D., and Shochi, T. (2016). Perception of prosodic social affects in Japanese: A free-labeling study. In *SPEECH PROSODY, 8th*, pages 811–815.

[Guo et al., 2021] Guo, P., Boyer, F., Chang, X., Hayashi, T., Higuchi, Y., Inaguma, H., Kamo, N., Li, C., Garcia-Romero, D., Shi, J., Shi, J., Watanabe, S., Wei, K., Zhang, W., and Zhang, Y. (2021). Recent Developments on Espnet Toolkit Boosted By Conformer. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5874–5878.

[Gupta et al., 2014] Gupta, V., Kenny, P., Ouellet, P., and Stafylakis, T. (2014). I-vector-based speaker adaptation of deep neural networks for French broadcast audio transcription. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*.

[Hadian et al., 2018] Hadian, H., Sameti, H., Povey, D., and Khudanpur, S. (2018). End-to-end Speech Recognition Using Lattice-free MMI. In *Interspeech 2018*, pages 12–16. ISCA.

[Hall et al., 2009] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA data mining software: An update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

[Heba, 2021] Heba, A. (2021). *Reconnaissance Automatique de La Parole à Large Vocabulaire : Des Approches Hybrides Aux Approches End-to-End*. Theses, Université toulouse 3 Paul Sabatier.

[Henrich and Popeil, 2003] Henrich, N. and Popeil, L. (2003). Acoustical description of 8 common singing styles produced by a single female singer: Preliminary results. In *Care of the Professional Voice Symposium*, Philadelphia.

[Henrich Bernardoni, 2006] Henrich Bernardoni, N. (2006). Mirroring the voice from Garcia to the present day: Some insights into singing voice registers. *Logopedics, phoniatrics, vocology*, 31:3–14.

[Hillenbrand et al., 1994] Hillenbrand, J., Cleveland, R. A., and Erickson, R. L. (1994). Acoustic correlates of breathy vocal quality. *Journal of Speech, Language, and Hearing Research*, 37(4):769–778.

[Hobson et al., 0030] Hobson, D. E., Lang, A. E., Martin, W. R. W., Razmy, A., Rivest, J., and Fleming, J. (2002-01-23/0030). Excessive daytime sleepiness and sudden-onset sleep in Parkinson disease: A survey by the Canadian Movement Disorders Group. *JAMA*, 287(4):455–463.

[Hori et al., 2017a] Hori, T., Watanabe, S., and Hershey, J. R. (2017a). Multi-level language modeling and decoding for open vocabulary end-to-end speech recognition. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 287–293.

[Hori et al., 2017b] Hori, T., Watanabe, S., Zhang, Y., and Chan, W. (2017b). Advances in joint CTC-attention based end-to-end speech recognition with a deep CNN encoder and RNN-LM: 18th Annual Conference of the International Speech Communication Association, INTERSPEECH 2017. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2017-August:949–953.

[Ioannidis and Rouas, 2012] Ioannidis, L. and Rouas, J.-L. (2012). Exploiting semantic content for singing voice detection. In *Sixth IEEE International Conference on Semantic Computing (IEEE ICSC2012)*.

[Ioannidis and Rouas, 2014] Ioannidis, L. and Rouas, J.-L. (2014). Caractérisation et classification automatique des modes phonatoires en voix chantée. In *XXXèmes Journées d'études Sur La Parole*.

[Kane and Gobl, 2013] Kane, J. and Gobl, C. (2013). Wavelet maxima dispersion for breathy to tense voice discrimination. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(6):1170–1179.

[Kannan et al., 2018] Kannan, A., Wu, Y., Nguyen, P., Sainath, T., Chen, Z., and Prabhavalkar, R. (2018). An Analysis of Incorporating an External Language Model into a Sequence-to-Sequence Model. pages 1–5828.

[Kim et al., 2017] Kim, S., Hori, T., and Watanabe, S. (2017). Joint CTC-attention based end-to-end speech recognition using multi-task learning: 2017 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2017. *2017 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2017 - Proceedings*, pages 4835–4839.

[Ko et al., 2015] Ko, T., Peddinti, V., Povey, D., and Khudanpur, S. (2015). Audio augmentation for speech recognition. In *Interspeech 2015*, pages 3586–3589. ISCA.

[Koehn, 2005] Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

[Krajewski et al., 2012] Krajewski, J., Schnieder, S., Sommer, D., Batliner, A., and Schuller, B. (2012). Applying multiple classifiers and non-linear dynamics features for detecting sleepiness from speech. *Neurocomputing*, 84:65–75.

[Kudo, 2018] Kudo, T. (2018). Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

[Kudo and Richardson, 2018] Kudo, T. and Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

[Lamel et al., 1991] Lamel, L., Gauvain, J. L., and Eskenazi, M. (1991). BREF, a large vocabulary spoken corpus for French. In *Eurospeech*.

[Laver, 1980] Laver, J. (1980). *The Phonetic Description of Voice Quality*. Cambridge Studies in Linguistics.

[Littner et al., 2005] Littner, M. R., Kushida, C., Wise, M., Davila, D. G., Morgenthaler, T., Lee-Chiong, T., Hirshkowitz, M., Daniel, L. L., Bailey, D., Berry, R. B., Kapen, S., Kramer, M., and Standards of Practice Committee of the American Academy of Sleep Medicine (2005). Practice parameters for clinical use of the multiple sleep latency test and the maintenance of wakefulness test. *Sleep*, 28(1):113–121.

[Maldonado et al., 2004] Maldonado, C. C., Bentley, A. J., and Mitchell, D. (2004). A pictorial sleepiness scale based on cartoon faces. *Sleep*, 27(3):541–548.

[Marine Guerry, 2019] Marine Guerry (2019). *Perception Interculturelle Des Affects Sociaux Multimodaux Entre Langues Premières et Secondes : Cas de l'anglais, Du Français et Du Japonais*. PhD thesis, Bordeaux Montaigne.

[Martin et al., 2020a] Martin, V. P., Chapouthier, G., Rieant, M., Rouas, J.-L., and Philip, P. (2020a). Using reading mistakes as features for sleepiness detection in speech. In *10th International Conference on Speech Prosody 2020*, Proc. 10th International Conference on Speech Prosody 2020, pages 985–989, Tokyo, Japan.

[Martin et al., 2021] Martin, V. P., Rouas, J.-L., Boyer, F., and Philip, P. (2021). Automatic Speech Recognition Systems Errors for Objective Sleepiness Detection Through Voice. In *Interspeech 2021*, pages 2476–2480. ISCA.

[Martin et al., 2020b] Martin, V. P., Rouas, J.-L., Micoulaud-Franchi, J.-A., and Philip, P. (2020b). The objective and subjective sleepiness voice corpora. In *12th Edition of Its Language Resources and Evaluation Conference.*, pages 6525–6533, Marseille, France.

[Martin et al., 2020c] Martin, V. P., Rouas, J.-L., and Philip, P. (2020c). Détection de la somnolence objective dans la voix. In Benzitoun, C., Braud, C., Huber, L., Langlois, D., Ouni, S., Pogodalla, S., and Schneider, S., editors, *6e Conférence Conjointe Journées d'Études Sur La Parole (JEP, 33e Édition)*, pages 406–414, Nancy, France. ATALA.

[Martin et al., 2019] Martin, V. P., Rouas, J.-L., Thivel, P., and Krajewski, J. (2019). Sleepiness detection on read speech using simple features. In *10th Conference on Speech Technology and Human-Computer Dialogue - SpeD 2019*, Timisoara, Romania.

[Mixdorff et al., 2017] Mixdorff, H., Hönemann, A., Rilliard, A., Lee, T., and Ma, M. K. (2017). Audio-visual expressions of attitude: How many different attitudes can perceivers decode? *Speech Communication*, 95:114–126.

[Morlec et al., 2001] Morlec, Y., Bailly, G., and Aubergé, V. (2001). Generating prosodic attitudes in French: Data, model and evaluation. *Speech Communication*, 33(4):357–371.

[Pedregosa et al., 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830.

[Peeters, 2003] Peeters, G. (2003). Automatic classification of large musical instrument databases using hierarchical classifiers with intertia ratio maximization. In *115th Convention of AES*, New York, USA.

[Pellegrino and André-Obrecht, 2000] Pellegrino, F. and André-Obrecht, R. (2000). Automatic language identification: An alternative approach to phonetic modeling. *Signal Processing*, 80(7):1231–1244.

[Poursadeghiyan et al., 2018] Poursadeghiyan, M., Mazloumi, A., Nasl Saraji, G., Baneshi, M. M., Khammar, A., and Ebrahimi, M. H. (2018). Using Image Processing in the Proposed Drowsiness Detection System Design. *Iranian Journal of Public Health*, 47(9):1371–1378.

[Povey et al., 2011] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., and Vesely, K. (2011). The kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, Hilton Waikoloa Village, Big Island, Hawaii, US. IEEE Signal Processing Society.

[Povey et al., 2016] Povey, D., Peddinti, V., Galvez, D., Ghahremani, P., Manohar, V., Na, X., Wang, Y., and Khudanpur, S. (2016). Purely Sequence-Trained Neural Networks for ASR Based on Lattice-Free MMI. In *Interspeech 2016*, pages 2751–2755. ISCA.

[Prabhavalkar et al., 2017] Prabhavalkar, R., Rao, K., Sainath, T. N., Li, B., Johnson, L., and Jaitly, N. (2017). A Comparison of Sequence-to-Sequence Models for Speech Recognition. In *Interspeech 2017*, pages 939–943. ISCA.

[Proutskova et al., 2013] Proutskova, P., Rhodes, C., Crawford, T., and Wiggins, G. (2013). Breathy, resonant, pressed - automatic detection of phonation mode from audio recordings of singing. *Journal of New Music Research*, 42(2):171–186.

[Raake, 2002] Raake, A. (2002). Does the Content of Speech Influence its Perceived Sound Quality? In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain. European Language Resources Association (ELRA).

[Rilliard et al., 2013] Rilliard, A., Erickson, D., Shochi, T., and De Moraes, J. A. (2013). Social face to face communication - American English attitudinal prosody. In *Proc. of Interspeech 2013*.

[Rilliard et al., 2009] Rilliard, A., Shochi, T., Martin, J.-C., Erickson, D., and Aubergé, V. (2009). Multimodal indices to Japanese and French prosodically expressed social affects. *Language and speech*, 52(2-3):223–243.

[Rouas and Ioannidis, 2016] Rouas, J.-L. and Ioannidis, L. (2016). Automatic classification of phonation modes in singing voice: Towards singing style characterisation and application to ethnomusicological recordings. In *Interspeech*, volume 2016, pages 150–154, San francisco, United States.

[Rouas et al., 2019] Rouas, J.-L., Shochi, T., Guerry, M., and Rilliard, A. (2019). Categorisation of spoken social affects in Japanese: Human vs. machine. In *International Congress of Phonetic Sciences ICPhS*, Melbourne, Australia.

[Sadanobu, 2004] Sadanobu, T. (2004). A natural history of Japanese pressed voice. *Journal of the Phonetic Society of Japan*, 8(1):29–44.

[Schuller et al., 2019] Schuller, B., Batliner, A., Bergler, C., Pokorny, F., Krajewski, J., Cychosz, M., Vollmann, R., Roelen, S.-D., Schnieder, S., Bergelson, E., Cristia, A., Seidl, A., Warlaumont, A., Yankowitz, L., Noeth, E., Amiriparian, S., Hantke, S., and Schmitt, M. (2019). The INTERSPEECH 2019 Computational Paralinguistics Challenge: Styrian Dialects, Continuous Sleepiness, Baby Sounds & Orca Activity. pages 2378–2382.

[Schuller et al., 2011] Schuller, B., Steidl, S., Batliner, A., Schiel, F., and Krajewski, J. (2011). The INTERSPEECH 2011 speaker state challenge. In *Twelfth Annual Conference of the International Speech Communication Association*.

[Schweitzer et al., 2001] Schweitzer, J. B., Cummins, T. K., and Kant, C. A. (2001). ATTENTION-DEFICIT/HYPERACTIVITY DISORDER. *Medical Clinics of North America*, 85(3):757–777.

[Sedgwick, 1998] Sedgwick, P. M. (1998). Disorders of the sleep-wake cycle in adults. *Postgraduate Medical Journal*, 74(869):134–138.

[Shahid, 2012] Shahid, A. (2012). *Stop, That and One Hundred Other Sleep Scales*. Springer, New York.

[Shochi et al., 2020] Shochi, T., Guerry, M., Rilliard, A., Donna, E., and Rouas, J.-L. (2020). The combined perception of socio-affective prosody: Cultural differences in pattern matching. *The Journal of the Phonetic Society of Japan*, 24:84–96.

[Shochi et al., 2009] Shochi, T., Rilliard, A., Aubergé, V., and Erickson, D. (2009). The role of prosody in affective speech. In Hancil, S., editor, *The Role of Prosody in Affective Speech*, volume Linguistic Insights 97, chapter Intercultural perception of English, French and Japanese social affective prosody, pages 31–59. Peter Lang.

[Shochi et al., 2018] Shochi, T., Rouas, J.-L., Guerry, M., and Erickson, D. (2018). Cultural differences in pattern matching: Multisensory recognition of socio-affective prosody. In *Interspeech 2018*, Hyderabad, India.

[Sjölander, 2004] Sjölander, K. (2004). The snack sound toolkit.

[Smith et al., 2005] Smith, C. G., Finnegan, E. M., and Karnell, M. P. (2005). Resonant voice: Spectral and nasendoscopic analysis. *Journal of voice : official journal of the Voice Foundation*, 19(4):607–22.

[Stolcke, 2002] Stolcke, A. (2002). SRILM - An extensible language modeling toolkit. In *ICSLP*, pages 901–904.

[Sundberg, 1987] Sundberg, J. (1987). *The Science of the Singing Voice*. Northern Illinois University Press.

[Sundberg, 1995] Sundberg, J. (1995). Vocal fold vibration patterns and phonatory modes. *Folia Phoniatrica Logopedica*, 47:218–228.

[Titze, 2002] Titze, I. R. (2002). Regulating glottal airflow in phonation: Application of the maximum power transfer theorem to a low dimensional phonation model. *The Journal of the Acoustical Society of America*, 111(1):367–376.

[Urban, 1988] Urban, G. (1988). Ritual wailing in amerindian Brazil. *American Anthropologist*, 90(2):385–400.

[Valstar et al., 2016] Valstar, M., Gratch, J., Schuller, B., Ringeval, F., Lalanne, D., Torres Torres, M., Scherer, S., Stratou, G., Cowie, R., and Pantic, M. (2016). Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pages 3–10. ACM.

[Watanabe et al., 2018] Watanabe, S., Hori, T., Karita, S., Hayashi, T., Nishitoba, J., Unno, Y., Soplin, N. E. Y., Heymann, J., Wiesner, M., Chen, N., Renduchintala, A., and Ochiai, T. (2018). ESPNet: End-to-end speech processing toolkit. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 2018-September, pages 2207–2211.

[Wu et al., 2016] Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J. R., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., and Dean, J. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *ArXiv*.

[Xu et al., 2018a] Xu, H., Chen, T., Gao, D., Wang, Y., Li, K., Goel, N., Carmiel, Y., Povey, D., and Khudanpur, S. (2018a). A Pruned Rnnlm Lattice-Rescoring Algorithm for Automatic Speech Recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5929–5933, Calgary, AB. IEEE.

[Xu et al., 2018b] Xu, J., Min, J., and Hu, J. (2018b). Real-time eye tracking for the assessment of driver fatigue. *Healthcare Technology Letters*, 5(2):54–58.

[Yanushevskaya et al., 2015] Yanushevskaya, I., Ní Chasaide, A., and Gobl, C. (2015). The relationship between voice source parameters and the Maxima Dispsersion Quotient (MDQ). *Interspeech 2015*.

[Young et al., 2006] Young, S. J., Evermann, G., Gales, M. J. F., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P. C. (2006). *The HTK Book, Version 3.4*. Cambridge University Engineering Department, Cambridge, UK.

[Zigmond and Snaith, 1983] Zigmond, A. S. and Snaith, R. P. (1983). The hospital anxiety and depression scale. *Acta Psychiatrica Scandinavica*, 67(6):361–370.

Appendices

A Detailed CV

A.1 Curriculum vitae

A.2 Situation Actuelle (depuis 09/2010) :

CHARGÉ DE RECHERCHE CNRS

au LABORATOIRE BORDELAIS DE RECHERCHE EN INFORMATIQUE (LaBRI - UMR 5800) à Talence, dans le Département "Image et Son".

A.3 Responsabilités administratives :

- Responsable adjoint de l'équipe "Image et Son" du LaBRI (environ 30 permanents) de 2013 à 2017.
- Responsable des relations internationales du LaBRI de 2015 à 2017.

A.4 Expérience Professionnelle :

10/2008-09/2010 : Chargé de Recherche CNRS 2ème classe

au LABORATOIRE D'INFORMATIQUE POUR LA MÉCANIQUE ET LES SCIENCES DE L'INGÉNIEUR (LIMSI - UPR 3251) à Orsay, dans le groupe Traitement du Langage Parlé.

02/2007-10/2008 : Chercheur contractuel

à l'INSTITUT NATIONAL DE RECHERCHE SUR LES TRANSPORTS ET LEUR SÉCURITÉ (INRETS - désormais IFSTTAR) à Villeneuve d'Ascq.
Travail sur l'analyse automatique de la conversation dans l'enceinte d'un véhicule de transport, dans le cadre du projet BOSS (On Board Wireless Secured Video Surveillance).

09/2006-02/2007 : Post-doctorant

au LABORATÓRIO DE SISTEMAS DE LÍNGUA FALADA (L2F) à Lisbonne, Portugal, dirigé par Isabel Trancoso. Bourse de post-doctorat attribuée par la *Fundaçáo para a Ciência e a Tecnologia (Ministério da Ciência, Tecnologia e Ensino Superior)*. Travail sur l'identification automatique des langues dans des documents audio-vidéo de journaux télévisuels en préalable à l'utilisation de modules de reconnaissance de la parole.

2005-2006 : Chercheur contractuel

à l'INSTITUT NATIONAL DE RECHERCHE SUR LES TRANSPORTS ET LEUR SÉCURITÉ (INRETS) à Villeneuve d'Ascq dans le cadre du projet EVAS (Etude de système de Vidéo et AudioSurveillance Sans fil) : Comment une analyse audio peut compléter une vidéo surveillance intelligente ? Reconnaissance d'événements audio liés à la sécurité des passagers dans les transports en communs.

2001-2005 : Doctorant

à l'INSTITUT DE RECHERCHE EN INFORMATIQUE DE TOULOUSE (IRIT - UMR 5505)
Thème de recherche : Identification automatique des langues basée sur des paramètres rythmiques et mélodiques. Contributions aux recherches de l'équipe sur l'indexation automatique de documents audio-visuels et en fusion de données.

A.5 Encadrement & Enseignement :

Encadrement de thèses

1. Vincent Martin (Bourse MESRI 10/2019 -) Co-direction avec Pierre Philip (PU-PH Sanspy USR 3413) Sujet de thèse : Intelligence Artificielle et Santé : mise au point de nouveaux biomarqueurs vocaux pour optimiser le suivi de la fatigue, de la somnolence et des troubles de l'humeur chez des patients souffrant de maladies chroniques
2. Florian Boyer (bourse CIFRE 05/2017 - 10/2021). Collaboration avec la start-up Airudit. Sujet de thèse : Reconnaissance de parole pour le français et intégration dans un système de compréhension du langage parlé
3. Leonidas Ioannidis (Bourse Université Bordeaux 1) 10/2011 - 01/2016 (abandon). co-encadrement avec Myriam Desainte-Catherine (30%) Sujet de thèse : Analyse automatique de la voix chantée.

Post-doctorants & Ingénieurs

1. Vincent Martin (Octobre 2018 - Octobre 2019) : Détection de la somnolence dans la voix de patients atteints de troubles du sommeil. Financement Région Nouvelle Aquitaine (projet IS-OSA)
2. Arnaud Dessein (Avril 2015 - Juin 2016) : Voice morphing through optimal transport. Financement Cluster CPU de l'IDEX Bordeaux. Co-encadrement avec Nicolas Papadakis (Institut de Mathématiques de Bordeaux).
3. Zuheng Ming (Septembre 2014 - Décembre 2015) : Visual analysis and synthesis of affects. Financement : projet IdEx interdisciplinaire " Seduction ". Co-encadrement avec Aurélie Bugeau (LaBRI) et Takaaki Shochi (CLLE-ERRSàB et LaBRI).
4. Dominique Fourer (février 2014 - février 2015) : Classification d'instruments dans les données ethnomusicologiques. Financement : projet ANR " DIADEMS ".

Encadrement de stagiaires

1. 2021 : Aymeric Ferron (2A ENSEIRB : implémentation d'une application web pour test cognitif dans un contexte de détection de la somnolence dans la voix
2. 2019 : Pierre Thivel (3A ENSEA) : détection de la somnolence dans la voix par apprentissage machine
3. 2017 : Raphaël Ollivier (Master2) : caractérisation de la voix pour l'aide au diagnostic de la dépression (collab. CHU Bordeaux)
4. 2017 : Julien Boissy (Master1) : reconnaissance de locuteur en temps réel
5. 2016 : Florian Boyer (Master2) : Transcription automatique temps réel en Français avec des réseaux de neurones profonds
6. 2016 : Raphaël Ollivier (Master1) : étude de paramètres pour la caractérisation des modes phonatoires en voix chantée
7. 2014 : Yohan Boclé (Master2) : Reconnaissance automatique du timbre des instruments de musique
8. 2014 : Gaëlle Pamphile (Master2) : Caractérisation de la voix chantée et parlée
9. 2013 : Bernardo Cohen (Master2) : Synthèse sonore dans le cadre de la description des unités minimales pour la musique électroacoustique
10. 2013 : Alexis Gay (Master2) : étude psychophysique de la perception de l'inharmonicité dans la musique électroacoustique
11. 2013 : Alban de Martin (Master1) : Création d'une interface graphique pour le contrôle du logiciel "fluxus" de reconnaissance vocale en temps réel appliqué au spectacle vivant.
12. 2013 : Céline Manenti (Master1) : Analyse, synthèse de sons percussifs (notamment de sons cannelés) à partir du critère de masse
13. 2013 : Soumia Kramdi (Master1) : Modélisation automatique de la prosodie pour la détection d'attitudes chez les apprenants d'une langue seconde
14. 2012 : Bernardo Cohen (Master1) : Caractérisation des unités minimales pour la musique électroacoustique

Participation à l'enseignement

- 2018-2019 : 70h eTD : M2 Informatique (Université de Bordeaux) : 6h eTD, M1 Informatique : 24h eTD - Analyse Classification Indexation de Données; L3 : 40h eTD - Projets de Communication Transdisciplinaires
- 2017-2018 : 30h eTD : M2 Informatique (Université de Bordeaux) : 6h eTD - traitement automatique de la parole, 3h eTD speech processing (master international), M1 Informatique : 24h eTD - Analyse Classification Indexation de Données
- 2016-2017 : 52h eTD dont : M2 Informatique (Université de Bordeaux) : 6h eTD - traitement automatique de la parole, 3h eTD speech processing (master international), M1 Informatique : 24h eTD - Analyse Classification Indexation de Données, Licence Informatique : 16h eTD Initiation à la programmation en C.
- 2015-2016 : 11h eTD dont M2 Informatique (Université de Bordeaux) : 7.5h eTD - Traitement automatique de la parole

- 2014-2015 : 20h eTD dont M2 Informatique (Université de Bordeaux) : 10h eTD - Traitement automatique de la parole
- 2013-2014 : 11h eTD dont : M2 Sciences du Langage (Université de Bordeaux Montaigne) : 6h eTD - Initiation au traitement automatique de la parole, applications pour les recherches en linguistique

Organisation de conférences et de workshops

Membre du comité d'organisation des conférences suivantes :

- International workshop on audio-visual affective prosody in social interaction & second language learning (AVAP 2015) - 5 et 6 mars 2015, Bordeaux
- International Conference: Context-based Spoken Japanese Language - 4 et 5 avril 2014, Bordeaux
- International Workshop of Cross Cultural research on Speech Communication & Second Language learning processing - 15 mars 2013, Bordeaux

A.6 Formation :

2005 : Doctorat en Informatique

École doctorale Informatique et Télécommunication - Université Paul Sabatier - Toulouse.
Titre : *Caractérisation et Identification Automatique des Langues*.
Directeur de Recherche (co-directeur) : Régine André-Obrecht (François Pellegrino).
Membres du jury :

Kamel Smaïli	Professeur, LORIA, Univ. Nancy 2	rapporteur
Louis-Jean Boé	IR, ICP, Univ. Stendhal, Grenoble	rapporteur
Véronique Aubergé	CR1, ICP, Univ. Stendhal, Grenoble	co-rapporteur
Daniel Dours	Professeur, IRIT, Univ. Paul Sabatier, Toulouse	président
Martine Adda-Decker	CR1, CNRS, LIMSI, Orsay	examinatrice

Thèse soutenue le 11 Mars 2005 - Mention Très Honorable.

2001 : D.E.A. Signal Image et Acoustique - option Acoustique

Université Paul Sabatier - Toulouse.
Titre du mémoire : *Intérêt de la prosodie en Identification automatique des langues.*
Responsable : Régine André-Obrecht.

2000 : Maitrise Traitement de l'Information

Université Paul Sabatier - Toulouse.

1999 : Licence d'Electronique Electrotechnique et Automatique

Université Paul Sabatier - Toulouse.

1998 : D.U.T. de Mesures Physiques

Université Paul Sabatier - Toulouse.

A.7 Transfert technologique, relations industrielles et valorisation

A.7.1 Participation à des contrats de recherche

1. PADE (Prosodie : Accents, Dialectes, Expressivité) - 2011-2015
   - Financement : Programme Jeunes Chercheurs de l'ANR - 160k€
   - Participants : LIMSI, GIPSA-Lab, Showa Music University (Japon), Universidade Federal do Rio de Janeiro (Brésil), LaBRI, CLLE-ERSSàB
   - Responsable de la tâche 2 : Distances prosodiques
2. Projet MexCulture (Indexation de collections multimédia pour la préservation et la dissémination de la culture mexicaine, ANR Blanc International, 2012-2015)
   - Financement : ANR - 623k€
   - Partenaires : CEDRIC, INA, LaBRI, IPN - Mexique, UNAM - Mexique.
   - Participation à la tâche 2 : Description du contenu audio/parole.
3. Projet DIADEMS (Description, Indexation, Accès aux Documents Ethnomusicologiques et Sonores, ANR CONTINT, 2012-2015)
   - Financement : ANR - 785k€
   - Partenaires : IRIT, LESC, Parisson, LaBRI, MNHN, LAM-IJLRA
   - Implication :
     - tâche 2 : Détection de segments sonores homogènes
     - tâche 3 : Structuration (analyse du contenu) - Responsable de la tâche
4. Projet MAVOIX (Manipulation acoustique de la voix pour la prosodie des affects sociaux, PEPS-IDEX CNRS, 2013-2014)
   - Financement : IDEX de Bordeaux - 20 k€
   - Partenaires : LaBRI, ERSSàB, IRCAM
5. Projet SEDUCTION (Social affEcts Discrimination Using Combined acousTic and vIsual informatiON) (Labex CPU, Univ. Bordeaux, 2014-2015)
   - Financement : Labex CPU de l'Université de Bordeaux - 100 k€
   - Partenaires : LaBRI
6. Projet AIME (Affective Interaction in Multimodal Environment), BIS-Japan (IDEX Univ. Bordeaux, 2014-2015)
   - Financement : Labex CPU de l'Université de Bordeaux - 17 k€
   - Partenaires : LaBRI - Université de Kyoto
7. Projet SoAPS (Social Affective Prosody Synthesis – 2015) - Programme "Visiting Scholars" de l'IdEx de Bordeaux
   - Invitation du Pr. Sagisaka (Université de Waseda, Tokyo, Japon) pendant 1 mois à Bordeaux (Février 2015)
8. Projet ParOral (2016)

- Projet Programme "Arts et Sciences" IDEX Bordeaux
- Thème : Reconnaissance vocale en temps réel pour le spectacle vivant
- Partenaires : metteur en scène Georges Gagneré
- Montant 12k€
- Rôle : PI

9. Projet DADOT (Data Analysis with Discrete Optimal Transport) (2015-2016)
    - Projet BIS-Japan, IDEX Univ. Bordeaux
    - Thème : Collaboration avec l'Université de Kyoto (Yamamoto Cuturi Lab.) sur le Transport Optimal
    - Partenaires Institut de Mathématiques de Bordeaux et l'Université de Kyoto, Japon (Yamamoto Cuturi Lab.)
    - Rôle : partenaire
10. Projet Virtual Laughter (Social role of laughter in the virtual immersive environment) 2016-2017
    - Projet PEPS IDEX Univ. Bordeaux
    - Thème : Etude de la perception et de la réalisation du rire social en Français et en Japonais
    - Partenaires CLLE-ERSSàB (Bordeaux, SHS) et Université de Kyoto (Nishida Lab.).
    - Montant 15k€
    - Rôle : PI
11. Projet "Cultural difference of Social laughter" (2017-2018)
    - Projet BIS-Japan IDEX Univ. Bordeaux
    - Thème : Etude des aspects cognitifs et culturels du rire social
    - Partenaires CLLE-ERSSàB (Bordeaux, SHS) et Université de Tsukuba (Département de Psychologie).
    - Montant 9k€
    - Rôle : partenaire
12. Projet "IS-OSA" (Solution numérique Innovante pour un traitement personnalisé de l'apnée obstructive du sommeil) (Région Nouvelle Aquitaine, 2018-2021)
    - Projet Labex BRAIN,
    - Partenaires : AGFA Healthcare (désormais DEDALUS), AVAD, SEFAM, CHU de Bordeaux, SANPSY USR3413
    - Thème : développer une solution innovante dans la prise en charge du Syndrome d'Apnées du Sommeil, validée à l'échelle régionale qui sera à terme exploitable à l'échelle nationale et internationale.
    - Aide de la région : 900 k€
    - Rôle : partenaire
13. Projet "SOMVOICE" (Voice Biomarkers to Predict Excessive Daytime Sleepiness) (2018-2020)
    - Projet Labex BRAIN (IdeX Bordeaux)
    - Thème : Etude du comportement des biomarqueurs vocaux sur une population de sujets sains
    - Partenaires : SANPSY USR3413
    - montant 110k€
    - Rôle : partenaire

14. Projet "HUMAVOX" (Analyse acoustique et cognitive de la prosodie affective dans le soin gériatrique) (2019-2021)
    - Projet interMSHS
    - Thème : Etude de l'empathie dans la parole de personnels soignants en EPHAD
    - Partenaires CLLE-ERSSàB, IMS, IBISC EA4526
    - montant 18k€
    - Rôle : partenaire
15. Projet "FVLLMONTI" (Ferroelectric Vertical Low energy Low latency low volume Modules fOr Neural network Transformers In 3D)
    - Projet européen FETPROACT 2020
    - Thème : implémentation de traduction vocale automatique (reconnaissance de parole + traduction) sur un appareil miniature autonome non connecté (oreillette)
    - Partenaires : LAAS CNRS UPR8001, INL CNRS URM5270, EPFL (Suisse), Global TCAD Solutions GmbH (Allemagne), NaMLab GmbH (Allemagne)
    - montant 4,5 M€
    - Rôle : partenaire

A.7.2 Relations industrielles

- société AIRUDIT (précédemment EA4T) : Thèse CIFRE de Florian BOYER 2017/2021.

  Airudit (ex EA4T) est une entreprise spécialisée dans la conception des interfaces Humains / Systèmes pilotées par la voix. Dans ce cadre, Airudit a créé et développé en interne une plateforme disruptive de services de compréhension du langage naturel humain associant un moteur sémantique dont le squelette repose sur des ontologies métiers augmentées, des algorithmes de traitement automatique du langage naturel (SLU/ NLP/ NLU), et un moteur de Reconnaissance Automatique de la Parole (R.A.P) utilisant des réseaux de neurones profonds.

- société ADN.AI : Contrat de prestation de service sur l'étude de faisabilité du logiciel MERLIN pour la transformation de voix (2018, 2 mois).

  Adn.ai est le premier Voice and Creative Service Provider (VCSP) publicitaire au service des marques et des consommateurs. Nous diffusons des publicités intelligentes avec lesquelles les consommateurs peuvent parler. Notre technologie de voice ads rend interactifs tous les médias. Nos voice ads permettent d'engager un dialogue one to one, sans cookie, et de collecter de la data en toute transparence avec le consentement explicite de vos prospects.

A.8 Publication list

Publications

Journal Articles

[A1] Vincent P. Martin, Jean-Luc Rouas, Jean-Arthur Micoulaud-Franchi, Pierre Philip, Jarek Krajewski. "How to Design a Relevant Corpus for Sleepiness Detection Through Voice?" In: *Frontiers in Digital Health* 3 (Sept. 2021). DOI: 10.3389/fdgth.2021.686068. URL: https://hal.archives-ouvertes.fr/hal-03351753.

[A2] Vincent P. Martin, Jean-Luc Rouas, Pierre Philip. "Détection de la somnolence dans la voix : nouveaux marqueurs et nouvelles stratégies". In: *Revue TAL* (2020). URL: https://hal.archives-ouvertes.fr/hal-03157410.

[A3] Takaaki Shochi, Marine Guerry, Albert Rilliard, Erickson Donna, Jean-Luc Rouas. "The combined Perception of Socio-affective Prosody: Cultural Differences in Pattern Matching". In: *The Journal of the Phonetic Society of Japan* 24 (Dec. 2020), pp. 84–96. DOI: 10.24467/onseikenkyu.24.0_84. URL: https://hal.archives-ouvertes.fr/hal-03098638.

[A4] Takaaki Shochi, Marine Guerry, Hanako Suzuki, Mami Kanzaki, Jean-Luc Rouas, Toyoaki Nishida, Yoshimasa Ohmoto. "Vocal aspect of social laughter during virtual interaction". In: *Working Papers em Linguística* 19.2 (Mar. 2019), pp. 54–77. DOI: 10.5007/1984-8420.2018v19n2p54. URL: https://hal.archives-ouvertes.fr/hal-02407158.

[A5] Arnaud Dessein, Nicolas Papadakis, Jean-Luc Rouas. "Regularized Optimal Transport and the ROT Mover's Distance". In: *Journal of Machine Learning Research* (2018). URL: https://hal.archives-ouvertes.fr/hal-01540866.

[A6] Myriam Desainte-Catherine, Pierre Hanna, Matthias Robine, Jean-Luc Rouas. "LaBRI. Modélisation du son, de la parole et de la musique". In: *Revue des Sciences et Technologies de l'Information* 32.3-4 (2013), pp. 515–518. URL: https://hal.archives-ouvertes.fr/hal-00866056.

[A7] Jean-Luc Rouas, Isabel Trancoso, Céu Viana, Mónica Abreu. "Language and Variety Verification on Broadcast News for Portuguese". In: *Speech Communication* 50.11-12 (Oct. 2008), p. 965. DOI: 10.1016/j.specom.2008.05.006. URL: https://hal.archives-ouvertes.fr/hal-00499218.

[A8] Jean-Luc Rouas, Isabel Trancoso, Céu Viana, Mónica Abreu. "Language and Variety Verification on Broadcast News for Portuguese". In: *Speech Communication, special issue on Iberian languages* 50.11-12 (2008), pp. 965–979. URL: https://hal.archives-ouvertes.fr/hal-00664601.

[A9] S. Ambellouis, L. Khoudour, Jean-Luc Rouas, A. Flancquart. "Système d'Aide à la Vidéo et a l'Audio Surveillance des Systèmes de Transport". In: *Génie logiciel* 82 (2007), pp. 56–64. URL: https://hal.archives-ouvertes.fr/hal-01695722.

[A10] Jean-Luc Rouas. "Automatic prosodic variations modelling for language and dialect discrimination". In: *IEEE Transactions on Audio, Speech and Language Processing* 15.6 (Aug. 2007). DOI: 10.1109/TASL.2007.900094. URL: https://hal.inria.fr/hal-00657977.

[A11] Jérôme Farinas, Jean-Luc Rouas, François Pellegrino, Régine André-Obrecht. "Extraction automatique de paramètres prosodiques pour l'identification automatique des langues". In: *Traitement du Signal* 22.2 (2005), pp. 81–97. URL: https://hal.archives-ouvertes.fr/hal-00664990.

[A12] Jean-Luc Rouas, Jérôme Farinas, François Pellegrino, Régine André-Obrecht. "Rhythmic unit extraction and modelling for automatic language identification". In: *Speech Communication* 47.4 (2005), pp. 436–456. URL: https://hal.archives-ouvertes.fr/hal-00664988.

[A13] Julien Pinquier, Jean-Luc Rouas, Régine André-Obrecht. "Fusion de paramètres pour une classification automatique parole/musique robuste". In: *Technique et science informatiques (TSI) : Fusion numérique/symbolique* 22 (2003), pp. 831–852. URL: https://hal.archives-ouvertes.fr/hal-01695732.

Conferences

[C1] Vincent P Martin, Jean-Luc Rouas, Florian Boyer, Pierre Philip. "Automatic Speech Recognition systems errors for objective sleepiness detection through voice". In: *Interspeech 2021*. Brno (virtual), Czech Republic: ISCA, Aug. 2021, pp. 2476–2480. DOI: 10.21437/Interspeech.2021-291. URL: https://hal.archives-ouvertes.fr/hal-03328827.

[C2] Vincent P. Martin, Jean-Luc Rouas, Florian Boyer, Pierre Philip. "Automatic Speech Recognition systems errors for accident-prone sleepiness detection through voice". In: *EUSIPCO 2021*. Dublin (en ligne), Ireland, Aug. 2021. URL: https://hal.archives-ouvertes.fr/hal-03324033.

[C3] Vincent P. Martin, Gabrielle Chapouthier, Mathilde Rieant, Jean-Luc Rouas, Pierre Philip. "Détection de la somnolence par estimation d'erreurs de lecture". In: *6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 1 : Journées d'Études sur la Parole*. Ed. by Christophe Benzitoun, Chloé Braud, Laurine Huber, David Langlois, Slim Ouni, Sylvain Pogodalla, and Stéphane Schneider. Nancy, France: ATALA, 2020, pp. 397–405. URL: https://hal.archives-ouvertes.fr/hal-02798563.

[C4] Vincent P. Martin, Gabrielle Chapouthier, Mathilde Rieant, Jean-Luc Rouas, Pierre Philip. "Using reading mistakes as features for sleepiness detection in speech". In: *10th International Conference on Speech Prosody 2020*. Proc. 10th International Conference on Speech Prosody 2020. Tokyo, Japan, May 2020, pp. 985–989. URL: https://hal.archives-ouvertes.fr/hal-02495149.

[C5] Vincent P. Martin, Jean-Luc Rouas, Jean-Arthur Micoulaud-Franchi, Pierre Philip. "The Objective and Subjective Sleepiness Voice Corpora". In: *12th Edition of its Language Resources and Evaluation Conference.* Poster Session. Marseille, France, May 2020, pp. 6525–6533. URL: https://hal.archives-ouvertes.fr/hal-02489433.

[C6] Vincent P. Martin, Jean-Luc Rouas, Pierre Philip. "Détection de la somnolence objective dans la voix". In: *6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 1 : Journées d'Études sur la Parole*. Ed. by Christophe Benzitoun, Chloé Braud, Laurine Huber, David Langlois, Slim Ouni, Sylvain Pogodalla, and Stéphane Schneider. Nancy, France: ATALA, 2020, pp. 406–414. URL: https://hal.archives-ouvertes.fr/hal-02798565.

[C7] Vincent P. Martin, Jean-Luc Rouas, Pierre Thivel, Jarek Krajewski. "Sleepiness detection on read speech using simple features". In: *10th Conference on Speech Technology and Human-Computer Dialogue - SpeD 2019*. Timisoara, Romania, Oct. 2019. DOI: 10.1109/SPED.2019.8906577. URL: https://hal.archives-ouvertes.fr/hal-02132438.

[C8] Jean-Luc Rouas, Takaaki Shochi, Marine Guerry, Albert Rilliard. "Categorisation of spoken social affects in Japanese: human vs. machine". In: *International Congress of Phonetic Sciences ICPhS*. Melbourne, Australia, Aug. 2019. URL: https://hal.archives-ouvertes.fr/hal-02317743.

[C9] Takaaki Shochi, Jean-Luc Rouas, Marine Guerry, Donna Erickson. "Cultural Differences in Pattern Matching: Multisensory Recognition of Socio-affective Prosody". In: *Interspeech 2018*. Hyderabad, India, Sept. 2018. DOI: 10.21437/interspeech.2018-1795. URL: https://hal.archives-ouvertes.fr/hal-01913705.

[C10] Takaaki Shochi, Marine Guerry, Jean-Luc Rouas, Marie Chaumont, Toyoaki Nishida, Yoshimasa Ohmoto. "Perceptual and acoustic correlates of spontaneous vs. social laughter". In: *Proc. 1st International Workshop on Vocal Interactivity in-and-between Humans, Animals and Robots*. Skovde, Sweden, Aug. 2017. URL: https://hal.archives-ouvertes.fr/hal-01695717.

[C11] Dominique Fourer, Takaaki Shochi, Jean-Luc Rouas, Albert Rilliard. "Perception of prosodic transformation for Japanese social affects". In: *Speech Prosody*. Vol. 2016. Boston, United States, May 2016, pp. 989–993. DOI: 10.21437/SpeechProsody.2016-203. URL: https://hal.archives-ouvertes.fr/hal-01392309.

[C12] Jean-Luc Rouas, Léonidas Ioannidis. "Automatic Classification of Phonation Modes in Singing Voice: Towards Singing Style Characterisation and Application to Ethnomusicological Recordings". In: *interspeech*. Vol. 2016. San francisco, United States, Sept. 2016, pp. 150–154. DOI: 10.21437/Interspeech.2016-1135. URL: https://hal.archives-ouvertes.fr/hal-01392305.

[C13] Takaaki Shochi, Jean-Luc Rouas, Ming Zuheng, Marine Guerry, Aurélie Bugeau, Erickson Donna. "Cultural differences in pattern matching: multisensory recognition of socio-affective prosody". In: *International Congress of Psychology (ICP)*. Yokohama, Japan, July 2016. URL: https://hal.archives-ouvertes.fr/hal-01314830.

[C14] Zuheng Ming, Aurélie Bugeau, Jean-Luc Rouas, Takaaki Shochi. "Facial Action Units Intensity Estimation by the Fusion of Features with Multi-kernel Support Vector Machine". In: *11th IEEE International Conference on Automatic Face and Gesture Recognition Conference and Workshops*. Michel Valstar, Jeff Cohn, Lijun Jin, Gary McKeown, Marc Méhu, and Maja Pantic. Ljubljana, Slovenia, May 2015. URL: https://hal.archives-ouvertes.fr/hal-01126775.

[C15] Zuheng Ming, Aurélie Bugeau, Jean-Luc Rouas, Takaaki Shochi. "Facial Action Units Intensity Estimation by the Fusion of Features with Multi-kernel Support Vector Machine". In: *11th IEEE International Conference on Automatic Face and Gesture Recognition*. Michel Valstar, Jeff Cohn, Lijun Jin, Gary McKeown, Marc Méhu, and Maja Pantic. Ljubljana, Slovenia, May 2015, pp. 1–6. DOI: 10.1109/FG.2015.7284870. URL: https://hal.archives-ouvertes.fr/hal-02489819.

[C16] Pauline Mouawad, Myriam Desainte-Catherine, Jean-Luc Rouas. "Multilabel Classification of Non-Verbal Communication of Emotions". In: *8th International Workshop on Machine Learning and Music*. Vancouver, Canada, 2015. URL: https://hal.archives-ouvertes.fr/hal-01230566.

[C17] Toyoaki Nishida, Masakazu Abe, Takashi Ookaki, Divesh Lala, Sutasinee Thovuttikul, Hengjie Song, Yasser Mohammad, Christian Nitschke, Yoshimasa Ohmoto, Atsushi Nakazawa, Takaaki Shochi, Jean-Luc Rouas, Aurélie Bugeau, Fabien Lotte, Zuheng Ming, Geoffrey Letournel, Marine Guerry, Dominique Fourer. "Synthetic Evidential Study as Augmented Collective Thought Process – Preliminary Report". In: *Asian Conference on Intelligent Information and Database Systems (ACIIDS)*. Ed. by Ngoc Thanh Nguyen, Bogdan Trawiński, and Raymond Kosala. Vol. 9011. Intelligent Information and Database Systems. Bali, Indonesia: Springer, Mar. 2015, pp. 13–22. DOI: 10.1007/978-3-319-15702-3_2. URL: https://hal.archives-ouvertes.fr/hal-01695718.

[C18] Alejandro Ramírez, Jenny Benois-Pineau, Mireya Saraí García Vázquez, Andrei Stoian, Michel Crucianu, Mariko Nakano, Francisco Garcia-Ugalde, Jean-Luc Rouas, Henri Nicolas, Jean Carrive. "The Mex-Culture Multimedia platform: Preservation and dissemination of the Mexican Culture". In: *2015 13th International Workshop on Content-Based Multimedia Indexing (CBMI)*. Content-Based Multimedia Indexing (CBMI), 2015 13th International Workshop on. Prague, Czech Republic: IEEE, June 2015, pp. 1–6. DOI: 10.1109/CBMI.2015.7153624. URL: https://hal.archives-ouvertes.fr/hal-01436983.

[C19] Jean-Luc Rouas, Dominique Fourer. "Classifying instruments with timbral features: application to ethnomusicological recordings". In: *5th International Workshop on Folk Music Analysis*. Paris, Unknown Region, 2015. URL: https://hal.archives-ouvertes.fr/hal-01695719.

[C20] Takaaki Shochi, Dominique Fourer, Jean-Luc Rouas, Marine Guerry, Albert Rilliard. "Perceptual Evaluation of Spoken Japanese Attitudes". In: *18th International Congress of Phonetic Science*. Glasgow, Scotland UK, Unknown Region, Aug. 2015. URL: https://hal.archives-ouvertes.fr/hal-01230562.

[C21] Dominique Fourer, Marine Guerry, Takaaki Shochi, Jean-Luc Rouas, Jean-Julien Aucouturier, Albert Rilliard. "Analyse prosodique des affects sociaux dans l'interaction face à face en japonais". In: *XXXèmes Journées d'études sur la parole*. Le Mans, France, June 2014. URL: https://hal.archives-ouvertes.fr/hal-00992083.

[C22] Dominique Fourer, Jean-Luc Rouas, Pierre Hanna, Matthias Robine. "AUTOMATIC TIMBRE CLASSIFICATION OF ETHNOMUSICOLOGICAL AUDIO RECORDINGS". In: *International Society for Music Information Retrieval Conference (ISMIR 2014)*. Taipei, Taiwan, Oct. 2014. URL: https://hal.archives-ouvertes.fr/hal-01095153.

[C23] Dominique Fourer, Takaaki Shochi, Jean-Luc Rouas, Jean-Julien Aucouturier, Marine Guerry. "Going ba-na-nas: Prosodic analysis of spoken Japanese attitudes". In: *Speech Prosody 2014*. Dublin, Ireland, May 2014, p. 4. URL: https://hal.archives-ouvertes.fr/hal-00981263.

[C24] Léonidas Ioannidis, Jean-Luc Rouas, Myriam Desainte-Catherine. "Caractérisation et classification automatique des modes phonatoires en voix chantée". In: *XXXèmes Journées d'études sur la parole*. Le Mans, France, June 2014. URL: https://hal.archives-ouvertes.fr/hal-00992084.

[C25] Raphael Marczak, Pierre Hanna, Jean-Luc Rouas, Jasper Van Vught, Gareth Schott. "From Automatic Sound Analysis of Gameplay Footage [Echos] to the Understanding of Player Experience [Ethos]: an Interdisciplinary Approach to Feedback- Based Gameplay Metrics". In: *40th International Computer Music Conference (ICMC) joint with the 11th Sound amp; Music Computing conference (SMC)*. athens, Greece, Sept. 2014. URL: https://hal.archives-ouvertes.fr/hal-01080041.

[C26] Raphael Marczak, Gareth Schott, Pierre Hanna, Jean-Luc Rouas. "Feedback-based gameplay metrics". In: *FDG Foundations of Digital Games*. Greece, May 2013, pp. 71–78. URL: https://hal.archives-ouvertes.fr/hal-00868293.

[C27] Jean-Luc Rouas, Boris Mansencal, Joseph Larralde. "Tale following: Real-time speech recognition applied to live performance". In: *SMC Sound and Music Computing*. Stockholm, Sweden, July 2013, pp. 389–394. URL: https://hal.archives-ouvertes.fr/hal-00868248.

[C28] Leonidas Ioannidis, Jean-Luc Rouas. "Exploiting Semantic Content for Singing Voice Detection". In: *Sixth IEEE International Conference on Semantic Computing (IEEE ICSC2012)*. Parlemo, Italy, Sept. 2012. URL: https://hal.inria.fr/hal-00759923.

[C29] Philippe Boula De Mareüil, Jean-Luc Rouas, Manuela Yapomo. "In search of cues discriminating West-african accents in French". In: *Interspeech*. Florence, Italy, Aug. 2011, pp. 725–728. URL: https://hal.inria.fr/hal-00664512.

[C30] Jean-Luc Rouas, Mayumi Beppu, Martine Adda-Decker. "Comparaison des propriétés acoustiques de la parole lue, préparée et conversationnelle en français". In: *JEP*. Mons, Belgium, May 2010. URL: https://hal.inria.fr/hal-00664502.

[C31] Jean-Luc Rouas, Mayumi Beppu, Martine Adda-Decker. "Comparison of Spectral Properties of Read, Prepared and Casual Speech in French". In: *LREC*. Valetta, Malta, May 2010. URL: https://hal.inria.fr/hal-00664499.

[C32] A Nasser, D Hamad, Jean-Luc Rouas, S Ambellouis. "The use of kernel methods for audio events detection". In: *ICTTA'08, 3rd International Conference on Information and Communication Technologies: From Theory to Applications*. Unknown, Unknown Region, 2008. URL: https://hal.archives-ouvertes.fr/hal-01695721.

[C33] Jean-Luc Rouas, Isabel Trancoso, Céu Viana, Mónica Abreu. "Portuguese Variety Identification on Broadcast News". In: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2008)*. Las Vegas, United States, Mar. 2008, pp. 4229–4232. URL: https://hal.archives-ouvertes.fr/hal-00664984.

[C34] Jean-Luc Rouas, Melissa Barkat-Defradas, François Pellegrino, Rym Hamdi. "Identification automatique des parlers arabes par la prosodie". In: *Journées d'Etude sur la Parole*. Dinard, France, June 2006. URL: https://hal.archives-ouvertes.fr/hal-00664987.

[C35] Jean-Luc Rouas, Jérôme Louradour, Sébastien Ambellouis. "Audio Events Detection in Public Transport Vehicle". In: *9th International IEEE Conference on Intelligent Transportation Systems (ITSC'2006)*. Toronto, Canada, 2006. URL: https://hal.archives-ouvertes.fr/hal-00664991.

[C36] Van-Thinh Vu, François Brémond, Gabriele Davini, Monique Thonnat, Quoc-Cuong Pham, Nicolas Allezard, Patrick Sayd, Jean-Luc Rouas, Sébastien Ambellouis, Amaury Flancquart. "Audio-Video Event Recognition System For Public Transport Security". In: *International conference on Crime Detection and Prevention (ICDP'2006)*. London, United Kingdom, June 2006, pp. 414–419. URL: https://hal.archives-ouvertes.fr/hal-00664986.

[C37] Jean-Luc Rouas. "Modeling Long and Short-term prosody for language identification". In: *9th European Conf. on Speech Communication and Technology (INTERSPEECH'2005 - EUROSPEECH), Lisboa, Portugal, 04/09/05-08/09/05*. Ed. by International Speech Communication Association (ISCA. Lisboa, Portugal, 2005. URL: https://hal.archives-ouvertes.fr/hal-00664989.

[C38] Jorge Gutiérrez, Jean-Luc Rouas, Régine André-Obrecht. "Application of uncertainty-based methods to fuse language identification expert decisions". In: *International Processing and Management of Uncertainty in Knowledge-based Systems (IPMU 2004)*. Perogia, Italia, Unknown Region, 2004. URL: https://hal.archives-ouvertes.fr/hal-01695725.

[C39] Jorge Gutiérrez, Jean-Luc Rouas, Régine André-Obrecht. "Fusing language identification systems using performance confidence indexes". In: *International Conference on Acoustics, Speech and Signal Processing (ICASSP 2004)*. Montréal, Canada, 2004. URL: https://hal.archives-ouvertes.fr/hal-01695728.

[C40] Jorge Gutiérrez, Jean-Luc Rouas, Régine André-Obrecht. "Weight loss functions to make risk-based language language identification fused decisions". In: *International Conference on Pattern Recognition (ICPR 2004)*. Cambridge, United Kingdom, 2004. URL: https://hal.archives-ouvertes.fr/hal-01695724.

[C41] François Pellegrino, Jérôme Farinas, Jean-Luc Rouas. "Automatic Estimation of Speaking Rate in Multilingual Spontaneous Speech". In: *Speech Prosody*. Nara, Japan, 2004, pp. 517–520. URL: https://hal.archives-ouvertes.fr/hal-01695727.

[C42] Jean-Luc Rouas, Jérôme Farinas, François Pellegrino. "Evaluation automatique du débit de la parole sur des données multilingues spontanées". In: *Journées d'Etude sur la Parole*. Fès, Morocco: LPL-ENS/Fes-AFCP-ISCA, 2004, pp. 437–440. URL: https://hal.archives-ouvertes.fr/hal-01695726.

[C43] Jorge Gutiérrez, Jean-Luc Rouas, Régine André-Obrecht. "Application de l'analyse factorielle discriminante à l'identification des langues". In: *Rencontres Jeunes Chercheurs en Parole (RJC Parole 2003)*. Grenoble, France, 2003. URL: https://hal.archives-ouvertes.fr/hal-01695729.

[C44] Nathalie Parlangeau-Vallès, Jérôme Farinas, Dominique Fohr, Irina Illina, Ivan Magrin-Chagnolleau, Odile Mella, Julien Pinquier, Jean-Luc Rouas, Christine Sénac. "Audio Indexing on the Web: a Preliminary Study of Some Audio Descriptors". In: *7th World Multiconference on Systematics, Cybernetics and Informatics - SCI'2003*. Colloque avec actes et comité de lecture. internationale. Orlando, Florida, United States, July 2003, 4 p. URL: https://hal.inria.fr/inria-00107706.

[C45] Julien Pinquier, Julie Mauclair, Jean-Luc Rouas, Régine André-Obrecht. "Détection de la parole et de la musique : fusion de deux approches". In: *19e Colloque GRETSI sur le traitement du signal et des images (GRETSI'2003), Paris, France*. Unknown, Unknown Region: Télécom-Paris, 2003, 78–81, Vol. III. URL: https://hal.archives-ouvertes.fr/hal-01695733.

[C46] Julien Pinquier, Jean-Luc Rouas, Régine André-Obrecht. "A Fusion Study in Speech/Music Classification". In: *International Conference on Acoustics, Speech and Signal Processing (ICASSP'2003) , Hong Kong, China*. ISBN 0-7803-7664-1, Unknown Region: IEEE, 2003, 17–20, Vol. II. URL: https://hal.archives-ouvertes.fr/hal-01695730.

[C47] Jean-Luc Rouas, Jérôme Farinas, François Pellegrino. "Automatic Modelling of Rhythm and Intonation for Language Identification". In: *15th International Congress of Phonetic Sciences (15th ICPhS) , Barcelona, Spain*. Unknown, Unknown Region: Causal Productions Pty Ltd, 2003, pp. 567–570. URL: https://hal.archives-ouvertes.fr/hal-01695735.

[C48] Jean-Luc Rouas, Jérôme Farinas, François Pellegrino, Régine André-Obrecht. "Modeling Prosody for Language Identification on Read and Spontaneous Speech". In: *International Conference on Acoustics, Speech and Signal Processing (ICASSP'2003) , Hong Kong, China*. ISBN 0-7803-7664-1, Unknown Region: IEEE, 2003, 40–43, Vol. I. URL: https://hal.archives-ouvertes.fr/hal-01695731.

[C49] Jérôme Farinas, François Pellegrino, Jean-Luc Rouas, Régine André-Obrecht. "Merging Segmental And Rhythmic Features For Automatic Language Identification". In: *International Conference on Acoutics, Speech and Signal Processing*. Vol. 1. Orlando, Florida, United States, 2002, pp. 753–756. URL: https://hal.archives-ouvertes.fr/hal-00702443.

[C50] Julien Pinquier, Jean-Luc Rouas, Régine André-Obrecht. "Fusion de paramètres pour une classification automatique Parole/Musique robuste". In: *Analyse et Indexation Multimédia - AS STIC "Indexation Multimédia : Transmodalité et Gestion des connaissances", Université Bordeaux 1*. Talence, Unknown Region: LaBRI, 2002. URL: https://hal.archives-ouvertes.fr/hal-01695737.

[C51] Julien Pinquier, Jean-Luc Rouas, Régine André-Obrecht. "Robust speech / music classification in audio documents". In: *International Conference on Spoken Language Processing (ICSLP'2002), Denver, Etats-Unis*. www.causal.on.net, Unknown Region: Causal Productions Pty Ltd, 2002, 2005–2008, Vol. 3. URL: https://hal.archives-ouvertes.fr/hal-01695739.

[C52] Jean-Luc Rouas, Jérôme Farinas, François Pellegrino. "Merging segmental, rhythmic and fundamental frequency features for automatic language identification". In: *Eusipco*. Toulouse, France: Eurasip, 2002, 591–594, vol. III. URL: https://hal.archives-ouvertes.fr/hal-01695738.

[C53] Jean-Luc Rouas, Jérôme Farinas, François Pellegrino, Régine André-Obrecht. "Fusion de paramètres rythmiques et segmentaux pour l'identification automatique des langues". In: *Journées d'Etude sur la Parole*. Nancy, France: LORIA-GFCP-SFA-ISCA, 2002, pp. 105–108. URL: https://hal.archives-ouvertes.fr/hal-01695736.

[C54] Jean-Luc Rouas. "Vers un système complet d'identification automatique des langues". In: *Rencontres Jeunes Chercheurs en Parole(RJC Parole'2001) , Mons, Belgique*. Unknown, Unknown Region: GFCP, 2001, pp. 126–129. URL: https://hal.archives-ouvertes.fr/hal-01695740.

Book chapters

[BC1] S. Ambellouis, A. Flancquart, L. Khoudour, Jean-Luc Rouas, T. Yahiaoui. "Innovations dans les transports guidés urbains et régionaux (Traité systèmes automatisés, IC2)". In: *Innovations dans les transports guidés urbains et régionaux (Traité systèmes automatisés, IC2)*. Ed. by Soulas, C, Wahl, and M. Vol. La vidéo et l'audio surveillance dans les transports guidés. Hermès Science Publications, 2010, pp. 184–196. URL: https://hal.archives-ouvertes.fr/hal-01695720.

[BC2] Jorge Gutierrez, Jean-Luc Rouas, Régine André-Obrecht. "Modern Information Processing: From Theory to Application (Selected Papers of the IPMU'2004)". In: *Modern Information Processing: From Theory to Application (Selected Papers of the IPMU'2004)*. Ed. by B. Bouchon-Meunier, G. Coletti, and R.R. Yager. Vol. Application of uncertainty-based methods to fuse language identification expert decision. Elsevier, 2006, pp. 255–268. URL: https://hal.archives-ouvertes.fr/hal-01695723.

Unpublished

[U1] Florian Boyer, Jean-Luc Rouas. "End-to-End Speech Recognition: A review for the French Language". 10 pages, 2 column-style. Feb. 2020. URL: https://hal.archives-ouvertes.fr/hal-02489009.

[U2] Nathalie Parlangeau-Vallès, Ivan Magrin-Chagnolleau, Dominique Fohr, Irina Illina, Odile Mella, Kamel Smaïli, Christine Sénac, Jérôme Farinas, Julien Pinquier, Jean-Luc Rouas, Régine André-Obrecht, François Pellegrino, David Janiszek. *Projet RAIVES (Recherche Automatique d'Informations Verbales Et Sonores) vers l'extraction et la structuration de données radiophoniques sur Internet*. Contract A02-R-553 || parlangeau-valles02a. Rapport de contrat. IRIT - Institut de recherche en informatique de Toulouse ; LORIA (Université de Lorraine, CNRS, INRIA), 2002. URL: https://hal.inria.fr/inria-00107633.