Formant Bandwidth and Resilience of Speech to Noise

Master Thesis

Leny Vinceslas

August 5, 2011

Internship for the ATIAM Master's degree

ENS - Laboratoire Psychologie de la Perception - Hearing Group

Supervised by: Alain de Cheveigné

Abstract

This work deals with formant bandwidth and its role in the perception of speech. In particular it presents a study of the relationship between formants bandwidth and speech intelligibility.

In a first time we investigate on methods for formant features extraction by means of evaluation tests based on synthetic vowels. Three different tools were evaluated: PRAAT, STRAIGHT and a script we developed for the occasion. The first two tools use linear predictive coding (LPC) and the third one is built on an implementation of pitch synchronous envelope (PSE) coupled with a simple peaks picking algorithm. Results were not promising, we observed a lack of reliability on the three techniques but we used our implementation of PSE for rest of the study. The first between-subject analysis did not any serious relationship between intelligibility rank and bandwidth. The analysis of a second within-subject shows a small effect of the Lombard speech over the formants bandwidth. A last between-subject analysis was carried out by deducing glottis open quotient from electroglottogam, but no tendencies could be observed.

Finally we verified the initial assumption by completing a psychoacoustical experiment. Double synthetic vowels were presented to subjects who were asked to report the perceived stimuli. Vowels were synthesized by varying relative loudness, pitch and formant bandwidth. The observed effect was not as strong as expected but it gives the confirmation that formant bandwidth has a real role on intelligibility of vowels.

Keywords: psychoacoustic, hearing, speech, voice, forced speech, Lombard effect, formant bandwidth, noise, electroglottogram,vowel synthesis.

Acknowledgments

I warmly thank all my subjects for participating in the testing cabin at the right time, and for pressing the right button.

I would also like to thank all my classmates from the Master ATIAM and in addition a general thanks goes to the whole teaching staff for their kindness and their availability.

I greatly appreciated to meet the hearing team who were very interesting and remarkable people.

I would like to sincerely thank Daniel Pressnitzer for having introduced me to Alain.

I am extremely grateful to Trevor Agus, young postdoc, for his valuable advices and friendly help on ANOVA.

I especially want to thank my supervisor, Alain de Cheveigné, for his guidance, his insightful tips and his continuous support during my stay at the hearing lab. His perpetual energy for research kept me motivated throughout my internship and furthermore, he was always available and enthusiastic.

Preface

The present work results of an internship for the Master's degree in Acoustic, Signal processing and Computer science applied to Music at IRCAM and in partnership with the Université Paris VI and TelecomParis.

This internship took place at the Hearing Group of the Laboratoire of Psychology of Perception which is part of the Ecole Normale Superieure, CNRS and University Paris Descartes.

The aim of the Hearing team is to better understand the functional principles involved in the perception of speech, music, and complex sounds with a particular focus on time: the perception of temporal structure within sound, and the time-domain models of auditory processing. Research topics include speech perception in normal-hearing and hearing impaired people, pitch, sound localization, auditory scene analysis, memory for complex sounds, etc. The methodological approach of the group combines behavioural experiments, neuro-computational models, physiology, brain imaging, signal processing, and clinical applications.

Contents

List of Figures

Chapter 1

Introduction

Voices differ according to how well they survive masking by noise or other speakers. Some people's voice can be heard at a distance despite background babble, while others are barely intelligible even from close by. Acoustic power and a clear elocution are obvious factors, but some studies suggest that formant bandwidth could also be implicated in speech quality.

1.1 Motivation

Speech intelligibility is crucial to ensure reliable communication between persons but in noisy environments such as restaurants or factories, intelligibility can be severely deteriorated by concurrent signals such as other speakers, reverberation, or noise. Understanding speech in noise is a challenge for the elderly and the hearing-impaired.

Intelligibility depends on the acoustic environment (level of noise and reverberation), but also on the ability of the listener (hearing acuity, linguistic skills), and characteristics of the speech itself. The latter characteristics are the focus of this study. Speech quality is likely to depend on multiple factors, including factors intrinsic to the speaker, as well as variations of these characteristics due to voluntary or involuntary control. Deliberately clear articulation, and the Lombard effect (increase in vocal effort in the presence of noise), are examples of changes in voice characteristics that can affect intelligibility. The acoustical nature of such characteristics is yet unclear: our aim in this project is to contribute to their understanding.

From a practical point of view, the finding of reliable acoustic correlates of intelligibility would benefit applications that aim at enhancing perception and voice transmissions, such as hearing aids, cochlear implants or mobile phones. The definition of a set of perceptual attributes involved in speech quality could lead to the prediction of speakers' intelligibility which would allow to operate a selection on talkers that are employed in speech database recordings. Likewise, new knowledge on speech perception would impact speech synthesis and more generally all speech technologies.

In this study we focus on one particular acoustic correlate: formant bandwidth. There are several reasons for this focus. One is that formant bandwidth is in part determined by the open/closed quotient of the vocal folds, which is known to covary with vocal effort, and is also highly variable between speakers. Another is that there is evidence from psychophysical experiments with mixtures of synthetic vowels that formant bandwidth may affect mutual masking and intelligibility. Finally, formant bandwidth has not yet been explored thoroughly, compared to other acoustic factors such as intensity, rate of speech, spectral tilt, fundamental frequency, etc

1.2 Objectives

The goal of the present field of research is to provide a better understanding of factors involved in speech intelligibility in noise in general and the role of formant bandwidth in particular. On the way to this goal, we will also investigate the relations between vocal effort and formant bandwidth as occurs in Lombard speech.

1.3 Means and methodology

The relation between formant bandwidth and intelligibility in noise can be investigated using several approaches. Given that intelligibility of speech (in particular speech in noise) is known to vary between speakers, we can focus on the acoustic characteristics of the speech of a population of speakers, and relate it to measures of their intelligibility. Alternatively, given that the intelligibility of a given speaker varies according to speaking style, we can look at the acoustic correlates of variations in phonation that are known to affect intelligibility. We can also look at characteristics of the production process that determine these acoustic characteristics, such as revealed by analysis of the EGG signal. Finally, we can use psychoacoustical techniques to study the effect of direct manipulations of acoustic parameters of synthetic speech on intelligibility.

Each approach has advantages and disadvantages:

- Cross-speaker studies take advantage of the large inter-speaker differences in intelligibility in noise that have been revealed experimentally. However there are other differences between speakers that might also affect intelligibility (rate, elocution, etc.), and these may mask the formant-related factors.

- Within-speaker studies (for example comparing normal and Lombard speech of the same speaker) have the potential advantage that they neutralize some of the unwanted between-speaker variability. However, again, other acoustic factors may covary with the factors of interest, complicating the analysis.

- Production-based measures (EGG) provide insight to the main physical determinant of bandwidth (glottis open quotient), but they depend on a robust analysis of the EGG signal, which is a challenge (see below).

- Behavioral measures allow a direct assessment of the perceptual effects of acoustic parameters, but one must rule out spurious effects due to analysis and/or synthesis artifacts.

In the project we explored several of these approaches, in an attempt to overcome these difficulties and arrive at a reliable answer to our question. A major problem that we discovered was the poor reliability of existing methods to measure formant bandwidths in natural speech. Thus, a portion of the project was devoted to the development and evaluation of tools to measure speech formant bandwidths.

1.4 Outline

This report is structured in 7 chapters. The first chapter introduces the motivation of this work, the scope, and the main goals. The second chapter provides the general background behind the hypothesis that motivates this project. In particular it gives an introduction on voice production and sound perception, physiological factors that affect speech intelligibility, and their acoustic correlates. Chapter 3 introduces the methodology followed in the present work. It presents the used methods for analysis as well as the databases and finish by a presentation of the perceptual test. Chapter 4 provides a first idea of the capacities of the analysis methods by means of an evaluation over synthetic vowels. Chapter 5 presents the results of the formants bandwidth analysis and gives a short discussion about the problems encountered. Chapter 6 describes the building of the perceptual test and comment the results obtained. Finally chapter 7 gives the conclusion of the project plus a brief view of the future work.

Chapter 2

Background

2.1 Factors modifying intelligibility

Speakers differ in their intelligibility. Some are clear and easy to follow, while others appear to mumble and require effort to understand. Intelligibility varies also with noise and reverberation, but here again speakers in the resilience of their speech to such interference. This may, in part, reflect between-speaker differences in anatomy or physiology. For example a study that measured intelligibility of a set of words and sentences pronounced by a population of speakers found that women tend to be more intelligible than men or children [22].

For a given speaker, the intelligibility of the speech produced by that speaker may vary according to the context. In particular the speaker may adapt his or her speaking style, voluntarily or involuntarily, according to the perceived difficulty in transmission or understanding by a listener.

For a given speaker, the intelligibility of the speech produced by that speaker may vary according to the context. In particular the speaker may adapt his or her speaking style, voluntarily or involuntarily, according to the perceived difficulty in transmission or understanding by a listener. For instance when exposed to competing talkers, babble or stationary noise, talkers increase their intelligibility by means of the so called Lombard effect [20] by which the presence of noise induces an increase in vocal effort [21] [14] or other modifications to produce a more highly articulated speech [1]. The Lombard effect has a different impact on male and female speakers, and its acoustic correlates are highly variable from speaker to speaker [14]..

It is also interesting to note that speech intelligibility depends also on the linguistic aptitudes of listeners. For instance children or non-native adults need a clearer speech stream than adult natives in order to fully understand a message [22], as illustrated in figure 2.1. While such factors are not of direct interest in this study, they are worth noting, in particular because the relative important of acoustic factors might differ according to the target listener population.

2.2 Acoustic correlates

This section gives an overview of the acoustic and thus spectral features of speech that are associated to the previously described factors.

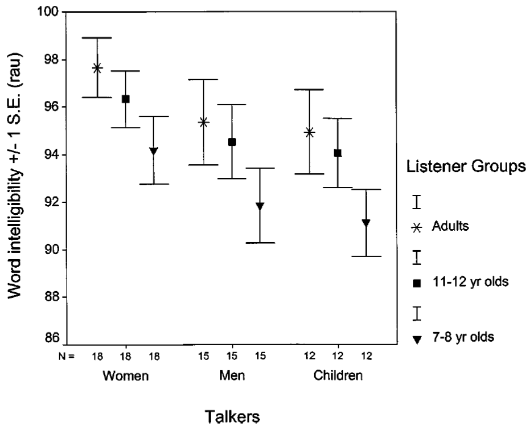


Figure 2.1: Word intelligibility rates for group of men, women, and 12-13 year old talkers, listened by groups of adults, 11-12 years old and 7-8 years old children.[12]

2.2.1 Fundamental frequency and spectral center of gravity

Variations of intelligibility can be related to both spectral and prosodic [1] changes in the speech signal. The high intelligibility of female voices is in fact associated to an increase of fundamental frequency as well as a growth of breath[16] which together contribute to increase the spectral center of gravity, and let us suppose less masking of mid-to-high frequency information than for male voices [9]. Consequently, speech seems to be more robust to noise when energy in the 1000 to 3000 Hertz range is increased [12]. In Lombard effect speakers attempt to compensate for the energy masking effect of the noise on their own speech by boosting the general level of energy and also by increasing the fundamental frequency and the formant energy[21] [6] [3]. Moreover vocal effort tends to increase the formant central frequency F1 while F2 has been reported to increase or decrease depending on the voice [21].

2.2.2 Rhythm, stress and intonation

As in clear speech, Lombard effect also plays a role on prosodic features by modifying phonemes, words and sentences durations as well as amplitude modulations [12] [7]. On the contrary pause durations and frequency which determine the speech rate are not necessary components affecting the speech intelligibility [17].

2.2.3 Roles of formant features

Since loudness, fundamental frequency and spectral center of gravity play an important role in speech intelligibility, it appears that other factors like formant

[1] In linguistics, prosody is referring to the rhythm, stress, and intonation of speech.

central frequency and formant bandwidth have also a small impact on the voice quality. Formants are peaks in an acoustic spectrum which results from resonant frequencies of the acoustic system. These formants describe the spectral structure of voiced speech and are the characteristic that allows us to identify the type of vowels produced as well as to recognize the identity of a known speaker.

Regarding a speech signal, the voice identity is mainly related to the formants central frequency. It is reported that a formants central frequency shift of 5% (in any direction) does not allow voice recognition. On the other hand a modification of formants bandwidth does not imply any alteration of the voice personality [18].

However formants bandwidth has an important effect on vowel identification. In noisy environment a decrease of the formant bandwidth induce an improvement of the vowel identification rate. Conservatively when the bandwidth is widened a significant reduction of the vowel identification rate can be observed [8] [26]. This increase or decrease of the vowel identification accuracy is largely effective in the variation of speech intelligibility [25].

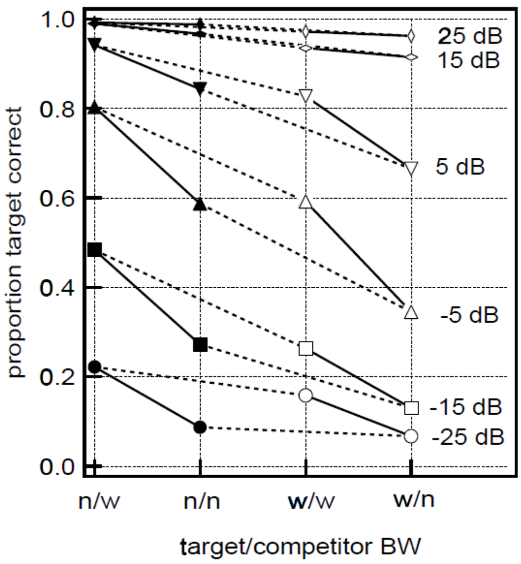
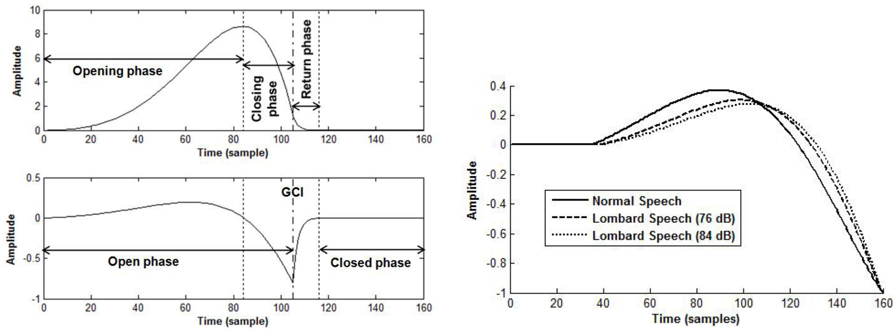


Figure 2.2: Identification rate as a function of formant bandwidth with n=narrow formants bandwidth and w=wide formants bandwidth [4].

Assuming an acoustic environment composed of several speakers; voices are not equally robust to the competitive speeches. Here again formants bandwidth affect the intelligibility of concurrent synthetic vowels. A vowel with narrow formants is a more potent masker, and also more resistant to masking, than a vowel with wide formants. A two-fold change in formant bandwidth would have an effect similar to a 10 dB change in relative amplitude between vowels within a mixture [4]. Figure 2.2 shows the proportion of correctly detected vowel during the vowel detection test described in [4]. It can be seen that target vowels are easily perceived when presented with a wide formant bandwidth competitive vowel.

2.2.4 Acoustic impact of the open quotient

In a physiological point of view Formant bandwidth reflects the damping of vocal tract resonances [10] [16]. This damping appears to be dependent to the open quotient (proportion of time that the glottis remains open). Figure 2.3 shows on its left side one period of a typical waveform of the glottal flow and its derivative. A more open glottal configuration results in a glottal waveform with greater low-frequency and weaker high-frequency components than a waveform produced with a more adducted glottal configuration. The more open glottal configuration generally leads to a louder source of aspiration noise and produces larger formants bandwidths [11]. Open quotient is also affected by speaking style and stress, so that formant bandwidth may be narrower in pressed speech, or speech spoken in a noisy environment [13] and [6]. Figure 2.3 presents on its right side the time and amplitude differences of the open glottal phase for normal and two level of Lombard speech. It is noticeable that the main variation occurs in the rapidity of the open phase time response.



(a) **Left**: Typical waveforms of one cycle of: (top) the glottal flow, (bottom) the glottal flow derivative[6]

(b) **Right**: Differences in the glottal open phase for normal and Lombard speech[6]

Figure 2.3: Typical opening phase waveforms

Chapter 3

Methodology

This chapter aims at providing an overview of the methods and tools used and developed throughout the present study. Since this work focus on vowels formant bandwidth and its implication on speech perception it is essential to describe in a first time the possible tools for formant features extraction. Then a review of the different analyzed databases is given. Each database allows the measure of a particular factor such as intelligibility among several talkers, variation of formants bandwidth between normal and Lombard speech or correlation of the open quotient with the formant bandwidth. The next section describes the speech analysis that is processed and presents the expected results. Finally, the last section of this chapter argument on the necessity of a perceptual test and provides a short description.

3.1 Methods for formant features extraction

Since formant location is a very important cue for human speech recognition, formant tracking is a substantial problem in the speech analysis framework. In this study three different tools for formants tracking are used. The two first used the popular formant tracking method by mean of linear predictive coding (LPC). A third tool using pitch synchronous envelope (PSE) method coupled with a simple peaks tracking is also implemented as an alternative method. In chapter 4 an evaluation of these three different tools is carried out before processing the formant bandwidth analysis on the database. The results provided by this first test give us an indication of the accuracy of our measurements. Nevertheless it is known that formants tracking methods are not very robust yet and still introduce errors such as bandwidth estimation [24].

3.1.1 Formant tracking by linear predictive coding

Linear predictive coding is currently the most widely used technique in phonetic for finding formants in the speech spectrum. In this method speech signal is first divided into frames by a sliding window, then linear predictive coefficient analysis model the signal by polynomial coefficients which best predict the signal. From those coefficients a root-solving algorithm is able of estimating the frequency position and width of the formant which composed the speech signal.

However, as mentioned above, the root extraction algorithm introduces several issues. The main problem is due to the use of an incorrect number of LP coefficients. Indeed, each pair of coefficient represents a formant, so that the use of too many or too few filter coefficients will directly impact the number of detected formant [27] [19]. Our two first analysis tools, PRAAT and STRAIGHT are based on the LPC method.

3.1.2 Formant tracking by spectral peaks picking

An alternative technique to the LPC formant tracking is implemented to track and measure formants. A pitch synchronous envelope (PSE) coupled with a simple peak tracking provide us with formants central frequency and formants bandwidth. This technique is implemented as it follows: In a first time the fundamental frequency of the speech signal is estimated by mean of a YIN implementation [5]. In a second time a preemphasis is applied to the signal and then a pitch synchronous interpolation is processed. At this point a simple Fast Fourier Transform is computed and a matrix of the spectral envelope estimation of each frame is returned. Here the previous fundamental frequency analysis allows us to discard non-voiced parts of the signal which are not interesting for our study. The next step is to detect the peaks of the envelopes which correspond to the formants and to measures their central frequency and bandwidth.

Many other techniques are existing [24][19][28], but since they are branches of LPC or peaks picking methods they will not be used in this study.

3.2 Speech databases

In order to observe formant bandwidth variations, formant features extraction is applied on three different databases. According to their difference of content the results of each database analysis will allow different assumptions.

3.2.1 The UCL speaker database

The UCL speaker database [2]results from a study on differences of intelligibility across many speakers realized in 2002 by Duncan Markham and Valerie Hazan. It gathers sets of 124 monosyllabic English words (UCL Markham Word test) pronounced by 45 subjects aged of 7 to 12, and adults (see table 3.1). Speech recordings were made an anechoic chamber. Glottal activity was also measured using an electrolaryngograph. Recordings were made at a sampling rate of 44.1 kHz.

Once the material recorded, a perceptual test has been designed with the aim to rank the subjects regarding their intelligibility (the ranking is presented in the appendix A.1). The stimuli used for perceptual testing were the recorded words leveled to a RMS level of -18 dB and then mixed with a 20-speaker multitalker babble levelled to -24 dB to obtain a Signal-to-Noise (SNR) ratio of +6 dB.

The formant feature extractions from this database are allowing to investigate on a possible correlation between perceptual intelligibility and acoustical cues such as formants bandwidth.

Speaker group	N	Age range	Mean	St.dev.
Adult females (AF)	18	22-58	33;11	10;9
Adult males (AM)	15	20-51	30;7	10;5
Child females (CF)	6	13-14	13;2	0;5
Child males (CM)	6	12-14	13;2	0;9

Figure 3.1: Summary of speaker's profile

3.2.2 The Lombard effect database

In their Lombard effect database[21] Martin Cooke and Youyi Lu gather a corpus of 400 sentences pronounced by height talkers. Talkers were exposed to different level (82 dB SPL and 96 dB SPL) of multi-talker babble while recording. As a result, the sentences of a same speaker are impacted by the recording conditions. Depending on the level of background noise a more or less strong vocal effort is induced. The comparison of the formants bandwidth of normal and Lombard speech allows to examine a potential correlation between vocal efforts and decreases of formants bandwidth.

3.2.3 The Electroglottogram database

The Electroglottogram database is in fact part of the material recorded by Duncan MARKHAM and Valerie HAZAN for the UCL speaker database. Here we assume it as an independent set of data since it introduces non-audio signal that requires a special analysis of the open quotient as described in the subsection 5.2.

The cross analysis of electroglottograms and corresponding speech signals allows to investigate on the relation between formant bandwidth variations and open quotient of the glottis.

3.3 Databases analysis

Given the potential estimation error that occurs in the measurement of formant bandwidth such as it is presented in the articles [23] and [27], it is important to evaluate our measurement methods before proceeding to the analysis on human speech. For that purpose we designed an evaluation test consisting of analyzing synthetic vowels. The use of synthetic vowels allow us to apply a cross validation by knowing the expected results. Moreover synthetic vowels are stationary and do not contain non-voiced speech.

Once strengths and weaknesses of the features extraction tools are identified, we can carry out the analysis on real speech with a more critical point of view. The formant features estimation will mainly take place in the analysis of the UCL speaker database and of the Lombard effect databases. Results are then compared with the intelligibility ranking provided with the database A.1 in order to find a possible relation between speaker intelligibility and average of formant bandwidth.

3.4 Perceptual test

The last step of the present study is the confirmation of the previous measurements. In order to validate the observed formants bandwidth variations a perceptual experiment is designed. This experiment mainly focuses on effect of the variation of formants bandwidth on competing vowels identification. This experiment provides the proof that formants bandwidth are playing a role in the reliance of speech to noise.

Chapter 4

Evaluation of the the formant features extraction

4.1 Synthetic vowels

A synthetic set of vowel is generated by a MATLAB implementation of a cascade formant synthesizer [15] which proceeds as it follows: In a first time the vowels parameters (pitch, duration, formants central frequency, formants bandwidth, jitter) are loaded. In a second time a first order glottal filter generate the vowel envelope from the givens parameters. Then real and imaginary parts are deduced from spectral envelope and used for additive synthesis. A random jitter can be applied before the synthesis in order to increase the naturalness of the vowel. The resulting samples are final normalized.

The synthetic database is made of the five fundamental vowels (/a/, /e/, /i/, /o/, /u/) of 500 milliseconds sampled at 44.1 kHz. Each vowel was generated with 8 different pitches (from 52 Hz to 400 Hz) and 3 bandwidths (half normal bandwidth, normal bandwidth and twice normal bandwidth) which give us a total of 5*8*3=120 samples. Figure 4.1 displays the five vowel envelopes. Solid lines represent normal formant bandwidths, dotted lines represent narrowed formant bandwidths and dashed lines represent widen formant bandwidths. As it can be observed, vowels with narrowed bandwidth present deeper valleys and higher peaks. In that case energy is more concentrated around formants. Along the same line, energy of widen formants vowels become more spread over the spectrum. All vowels parameters can be found in the appendix B.1.

4.1.1 Praat

A first evaluation is carried out using PRAAT, one of the most popular tools for phonetic analysis. It offers a set of standard tools for speech analysis and synthesis. In PRAAT formant features extraction is completed by a standard LPC implementation. Praat offers the possibility of using scripts instead of standard user interface so that we can automatize the analyses.

The results of the evaluation are presented in the figure 4.2. The three firsts graphs show median and the percentile at 10% and 90% of the estimated formants bandwidth as a function of the synthesis parameters while the fourth

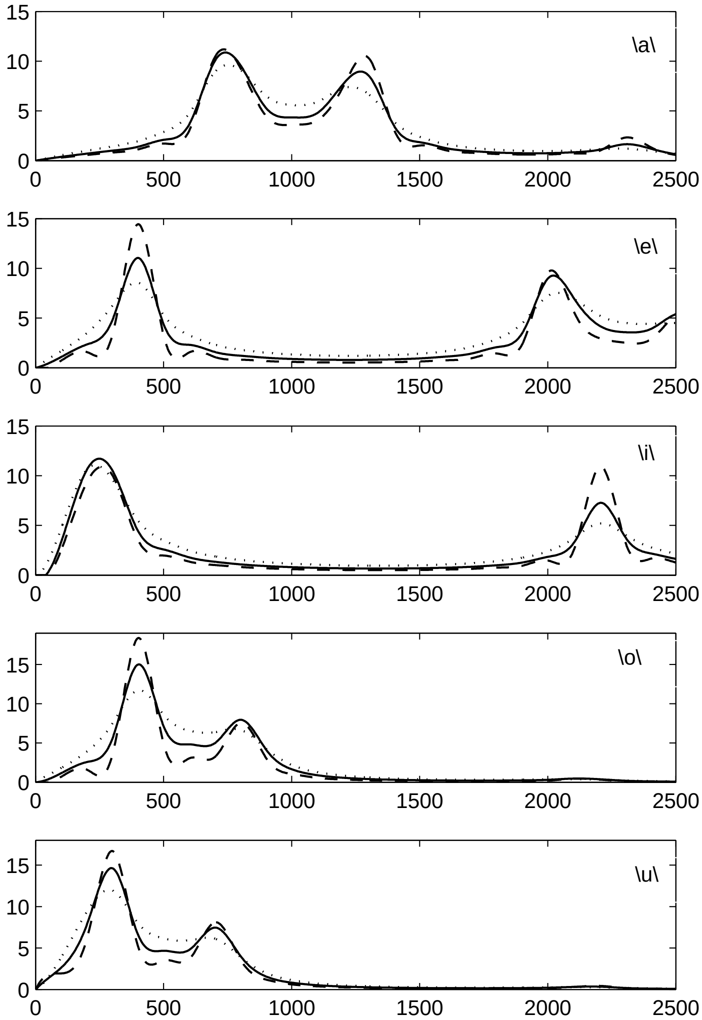


Figure 4.1: Spectral envelope of French vowels. Dotted lines: wide formants (twice normal bandwidth. continuous lines: normal formants. dashed lines: narrow formants (half normal bandwidth).

graph shows the estimated formants central frequency as a function of the synthesis parameters. It is noticeable that the variability of the bandwidth estimation is very large and gets even larger when the bandwidth increases. Nevertheless, the central frequency estimation provides much fair results.

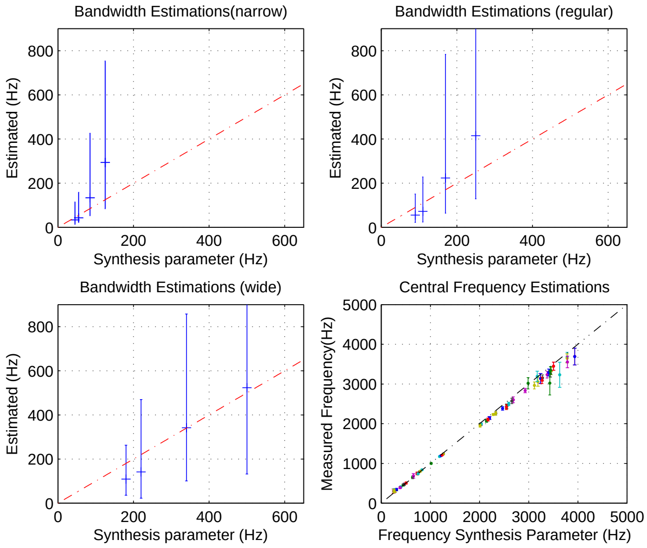


Figure 4.2: Percentile and median of formant bandwidth and central frequency measurements using PRAAT.

4.1.2 STRAIGHT

A second evaluation is accomplished using STRAIGHT, a MATLAB environment developed by Hideki Kawahara for speech analysis, morphing and synthesis. As in Praat, STRAIGHT formant features extraction use a standard implementation of the LPC algorithm.

The results of the evaluation are presented in the figure 4.3 which follow the same organization than the previous one. Here we can observe the same phenomenon than for the Praat evaluation. The variability of the bandwidth estimation stays very large and still grows up when the bandwidth increases. The central frequency estimation still provides correct values with almost no variability.

4.1.3 Pitch Synchronous Envelope plus peaks picking

A third tool for formants features extraction is specially implemented in MATLAB by means of pitch synchronous envelope plus a simple peak picking algorithm. In comparison with the two first methods, this one offers a best computation speed as well as does not requirer to choose a number of poles.

The results of the evaluation are presented in the figure 4.4. It can be seen that if the variability of the bandwidth estimation stays large its median is now closer from the synthesis values. With this method, central frequency estimation provides accurate measurements.

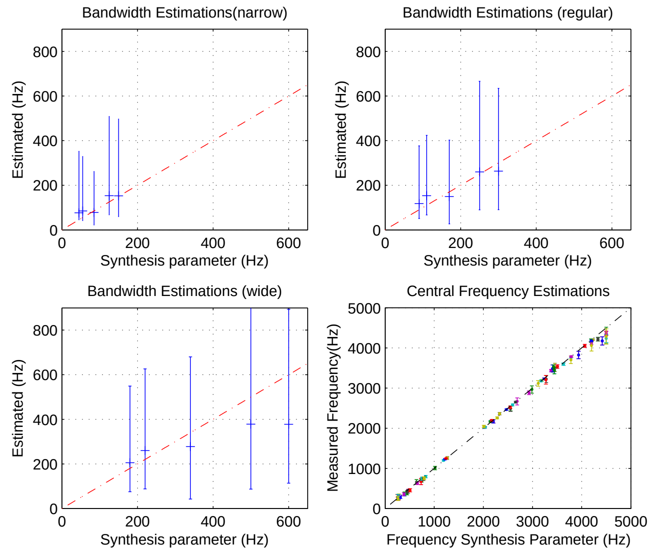


Figure 4.3: Percentile and median of formant bandwidth and central frequency measurements using STRAIGHT.

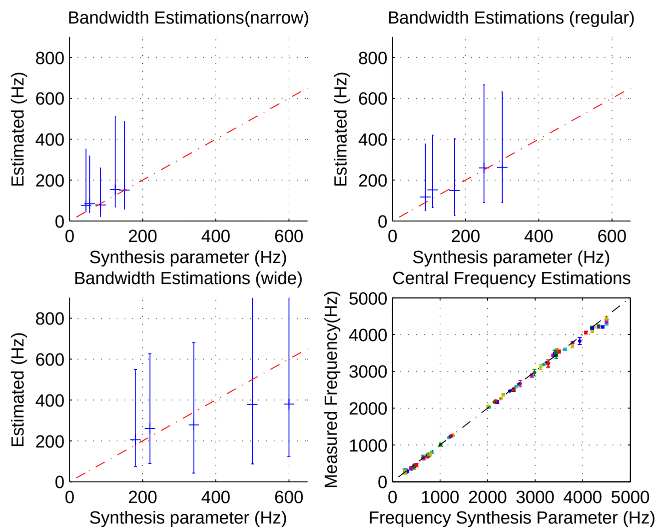


Figure 4.4: Percentile and median of formant bandwidth and central frequency measurements using PSE+peaks picking.

4.1.4 Discussion

The evaluation of three different formant estimation tools leads to the comparison of their performances. Trying several analysis tools has been profitable since it permits to notice that each tool behave differently depending on the situation (narrow or wide formant, high or low pitch). This evaluation highlights the poor accuracy of formants bandwidth estimation and brings us to the fact that none of these previously exposed tools are able to provide measurements with enough accuracy for speech formant bandwidth estimation.

Chapter 5

Speech analysis

In this chapter only the analysis method PSE plus peaks picking is kept. Firstly an analysis of the UCL speaker database is carried out in order to find a correlation between the intelligibility ranking which have been deduced from previous psychoacoustic experiments and the formant bandwidth estimated from audio samples. Secondly the same analysis is applied to the Lombard database with the aim of finding a relationship between normal speech and forced speech. The use of this database allows within-speaker measurements which reduce the large variation of voice features observed while measuring over a high number of speakers. Finally Electroglottograms are used to extract glottis open quotient.

5.1 Measurement of the effect of formants bandwidth

5.1.1 Estimation of the intelligibility of different speakers

Given the ranking of intelligibility provided with the UCL speaker database it becomes possible to map the measures of formants bandwidth with the intelligibily of each speaker. Formants bandwidth of analyzed words are averaged so that it remains only one bandwidth mean per talkers. Figure 5.1displays the intelligibility rank of each speaker as a function of its bandwidth mean. The scatter plot format allows to observe a potential interaction between rank and bandwidth mean. As it is noticeable no interaction between our measures and the indicated rank can be perceived.

The result of this first analysis can be explained by the large variability between-speaker of the voice features. Moreover in the present case the intelligibility ranking is not necessary due to the formants configuration but it also depends on pitch or other prosodic parameters.

5.1.2 Estimation of the Lombard effect

Given the previous results it is interesting to try the same measure within a unique subject. The analysis is therefore based on the Lombard effect database. Figure 5.2 presents for each word, its natural speech mean bandwidth as a function of its Lombard speech mean bandwidth. The scatter plot shows a

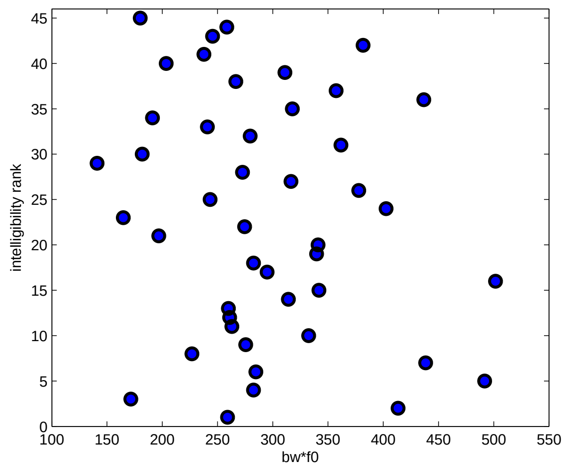


Figure 5.1: intelligibility ranking as a function of the estimated f0*bw

small tendency of narrower formant for the Lombard speech. A set of alternative measures is presented in appendix E.

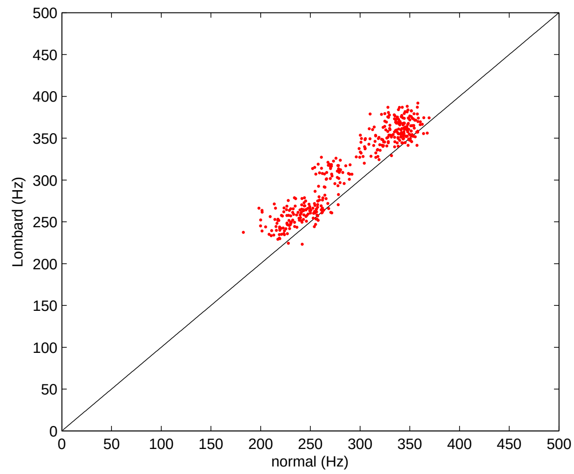


Figure 5.2: natural speech mean bandwidth as a function of its Lombard speech mean bandwidth. p=0.001.

5.2 Electroglottogram and formants bandwidth correlation

Using the Lombard database we could improve a bit our results, but the effect still remains weak. It is knows that glottis open quotient is a physical determinant parameter of the formants bandwidth. Thus we try to infer formant bandwidth from the electroglottogram in order to predict the intelligibility of subjects. A laryngograph is shown in figure 5.3 with its corresponding audio wave form.

The measure of the open quotient was challenging and as you can see in figure 5.4 signal can be very different from one to another so that it becomes difficult to detect the open and close glottis times.

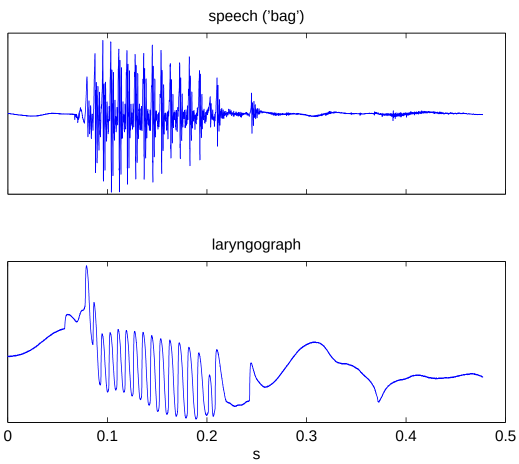


Figure 5.3: Speech wave form and the corresponding EEG for the word "bag"

Figure 5.5 displays the results of the analysis. Graph (a) shows a bargraph of six subject's open quotient. Subjects from the top are more intelligible than subject from the bottom, but this difference does not appear in the bargraph. Graph (b) shows the error percentage (reflecting the intelligibility) as a function of the mean open quotient. Here again no clear tendency can be observed.

5.3 Discussion

Trying to overcome our problem of bandwidth measurement, we complete 3 analysis based on 3 different dataset. Only the Lombard database shows a positive tendency, but as small as we cannot rely on it.

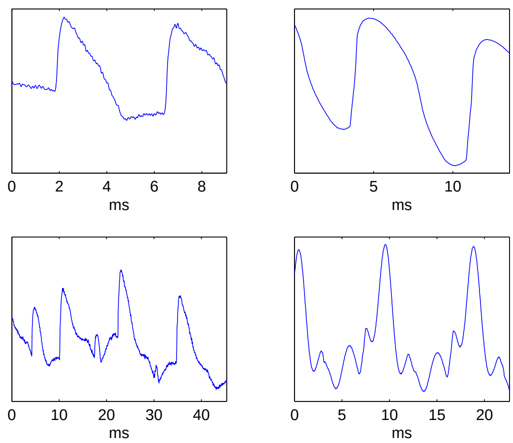
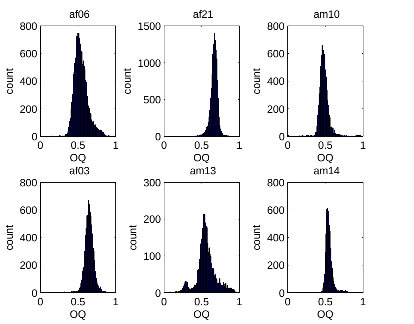
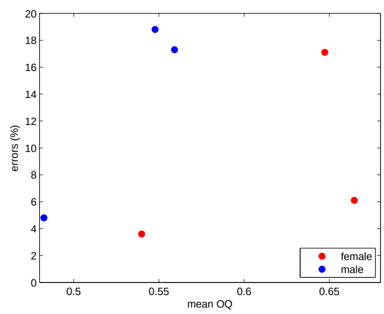


Figure 5.4: Four different EEGs



(a) Bargraph of open quotient measure for 6 subjects: top subjects are more intelligeble than bottom subjects



(b) Intelligibility ranking as a function of mean open quotient

Figure 5.5: Observation of the correlation between glottis open quotient and intelligibility

Chapter 6

Perceptual test

In order to confirm the weak formant bandwidth tendency observed in chapter 5, a psychoacoustic test is designed. Based on the experiment presented in [4], this test aims at measuring the effect of formant bandwidth variations on identification of double vowels.

6.1 Experience description

6.1.1 double vowels Synthesis

By means of the method presented in section 4.1, 5 single vowels are synthesized with varying bandwidths that take the values of one half (narrow), twice (wide) or normal bandwidth with two different fundamental frequencies of 124 Hz and 132 Hz. The double vowels are then created by adding single vowels together with a changing amplitude ratio of either -15dB, -5dB, 5dB or 15dB. Finally double vowels are normalized to a RMS value of 1. Fundamental frequencies of vowel couples can be either equal (Δ F0=0%) or different (Δ F0=6%). Four combinations of bandwidth are possible: narrow/narrow (n/n), narrow/wide (n/w), wide/narrow (w/n) and wide/wide (w/w).

We finally obtained a total of [20 vowel pairs] x [4 amplitude ratios] x [2 delta F0s] x [4 bandwidths]= 640 double vowel conditions. Vowels are about 270 milliseconds of duration with a random starting phase.

6.1.2 Settings and Calibration

The experiment took place in an acoustic isolated cabin with a Fireface 800 from RME as digital to analog converter. The headphones (Sennheiser HD250 Linear II) where calibrated using a Brüel&kjær artificial ear in order to get an average of 76,9 dB SPL. Instructions were displayed to subjects by mean of the MATLAB command window and their feedbacks were inputted with a traditional keyboard.

6.1.3 Subjects

For this experiment, there were a total of 16 subjects, 13 were native French speakers, 1 was francophone for more than 8 years and the last two were French speaker for 3 years. Most of them were Master students or young researcher aged from 22 to 31 with an average age of 25.78 years. One subject has a perfect pitch.

6.1.4 Two-alternative forced choice

In a first version of the experiment, subjects were asked to report either 1 or 2 answers regarding what they could perceive. It turns out that subjects was reporting more often unique vowel than pairs and thus the amount of data was not as profuse as expected. This issue has been be fixed by implementing a Two-alternative forced choice. In other words, subjects were forced to provide 2 answers even when one of the vowels was not perceptible. Using this second method we clearly notice an increase of the correct answers which can be explained by the subjects' will to only report absolutely sure responses.

However it seems that even when a vowel is not clearly perceived the latter influences the choice of the subject.

6.2 Effect of fundamental frequency difference

As a first observation we can look at the effect of ΔF0 over the identification rate of the vowels.

Figure 6.1 right shows the number of vowels as a function of the amplitude mismatch between vowels. It can be see that the number of vowel reported decrease with the amplitude mismatch. This is mainly due to the fact that large amplitudes mismatch make the stimuli similar to a single vowel. Also we see here that when $\Delta F0 \neq 0$ the number of vowel reported is greater.

Figure 6.1 left shows correct identification as a function of the amplitude mismatch. For both ΔF0 the identification rate was better when the acoustical level of the target was higher. Again we observe that the number of correct answers increase when $\Delta F0 \neq 0$.

6.3 Effect of bandwidth variation

The second observation was to look at the effect of the formant bandwidth over the identification rate of vowels. Figure 6.2 left shows the number of correct answer as a function of the background vowel formants bandwidth of the. We clearly see that identification is better when the target formants are narrow. Likewise the identification is also better if the background vowel has wide formants. Thus the case where the identification is optimum if for a double vowel made of a target with narrow format plus a background vowel with wide formants.

This phenomenon can be explained by supposing that background vowels are less masking when their formants are wide. On the other hand targets become more preeminent when their formants are narrowed.

This tendency can be found again in the figure 6.2 right which shows the identification rate as a function of the amplitude mismatch. This graph adds information about the incidence of the effect which seems to be non-present for small amplitudes.

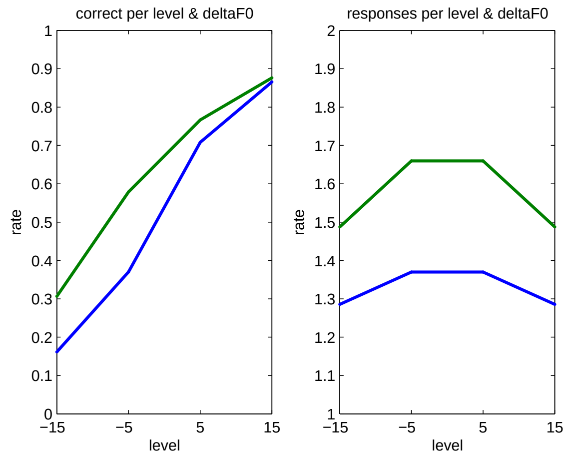


Figure 6.1: Green lines: ΔF0=6%. Blue lines: ΔF0=0%
Left: Correct identification as a function of the amplitude mismatch. **Right**: Number of reported vowel as a function of the amplitude mismatch.

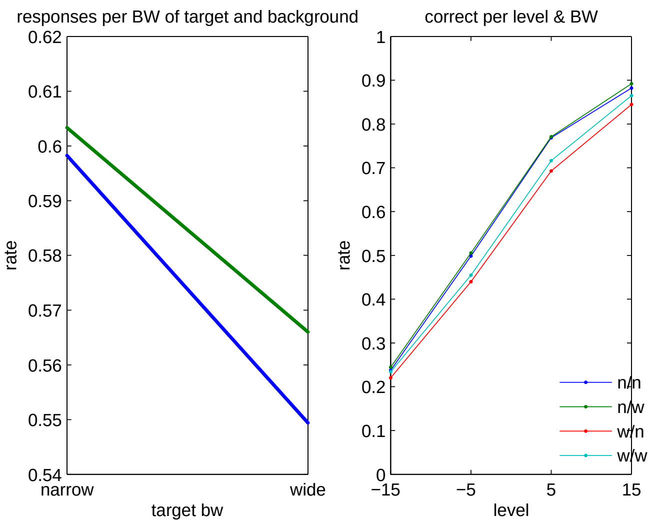


Figure 6.2: **Left**: Correct identification as a function of formant bandwidth of the target. Green line: background vowel with wide formant bandwidth. Blue line: background vowel with narrow formant bandwidth. **Right**: Correct identification as a function of the amplitude mismatch for the 4 combinations of target/background formant bandwidth .

6.4 Data validation by RM ANOVA

Collected data was analyzed by means of repeated measures analysis of variance (RM ANOVA). This procedure allowed us to observe the different sources of variance and to determine whether interactions between conditions where significant or not. Results of the experiment were split into 4 subsets corresponding to the different level of amplitude and the ANOVA of each subset was computed as shown in table 6.3. A selected part of the ANOVA is presented in the appendix D. The ANOVA highlight that conditions were all significant except the bandwidth at -15dB and the ΔF0 at 15dB that were highly dependent to the main effect.

Amplitude ratio	parameter	F	P
-15dB	BW	$F(1, 8) = 0.865$	$p =$ **0.466**
	ΔF0	$F(1, 8) = 49.64$	p =0 .001
-5dB	BW	$F(1, 8) = 7.315$	$p = 0.001$
	ΔF0	$F(1, 8) = 87.20$	$p = 0.001$
5dB	BW	$F(1, 8) = 13.49$	$p = 0.001$
	ΔF0	$F(1, 8) = 48.51$	$p = 0.001$
15dB	BW	$F(1, 8) = 6.400$	$p = 0.001$
	ΔF0	$F(1, 8) = 1.082$	$p =$ **0.315**

Figure 6.3: Summary of the ANOVA

6.5 Discussion

Since results of the formant bandwidth estimation were inconclusive, it was rational to change of investigation methods. The will of using different signal processing tools in order to measure formants bandwidth was inspired by the literature [8], [26] and [25] in which formant bandwidth seemed to be directly correlated with intelligibility. However, the previously presented test has been designed with the goal of verifying the hypothesis from the literature. For this reason the psychoacoustical experience of the this study is a replication of the experience presented in [4], and an achievement of corroborative results was expected.

Statistics from the experiment show the same tendency than in [4], but not as strong as the reference experiment. This difference can be explained by the behavior of the subjects which were not reporting the two vowels as much as possible. So it is possible that a lack of response influence the final result.

Nevertheless, even if the tendency is not as strong as expected, it gives the confirmation that formant bandwidth has a real effect on intelligibility of vowels.

Chapter 7

Conclusion

This report presents a study on formants bandwidth and reliance of speech to noise. In the project several approaches have been explored in order to arrive at a reliable answer.

The first part of the study was to investigate the correlation between formant bandwidths and intelligibility. Since we discovered it was problematic to obtain reliable formant bandwidth estimations from classical speech processing tools, we decided to build an evaluation test gathering 3 different tools of formants feature extraction. Two tools already existed and we implemented a third one based on pitch synchronous envelope and simple peaks picking. The evaluation test shows that none of the three methods were reliable enough to provide reliable measurements.

Despite the poor reliability of the formants bandwidth estimation we performed an analysis of 3 databases. A first database of 45 speakers provided with intelligibility ranking permits a cross-speaker study. As expected the reduced accuracy of the bandwidth estimation did not allow to observe any relationship between formant features and intelligibility. Nevertheless it is also possible that formants bandwidth does not impact the perception as much as pitch or loudness, consequently the implication of the formants bandwidth in the variation of the intelligibility could be masked by more influent parameters.

To overcome this issue a within-speaker analysis was carried out on a database gathering normal and Lombard speech for few talkers. But here again the large variability of the bandwidth measurement was not permitted to obtain satisfactory results.

As the relationship between formant bandwidths and intelligibility could not be proven, essentially due to the formant measurement, it was necessary to find a way of indirect measures. According to the literature glottis open quotient is the main physical determinant parameter of the bandwidth. Therefore we designed an analysis base on EEG which allows the inference of the glottis open quotient. But once again, the inference of the glottis open quotient form EEG was challenging and the results were not significant.

None of the 3 previous analyses were reliable, thus it became necessary to verify the first hypothesis which assume that formant bandwidth play a role in concurrent vowel identification.

With the purpose of checking this hypothesis we replicate the psychoacoustical experience presented in [4]. Thank to this experiments the bandwidth effect could be re-observed. Vowels with narrow formants bandwidth are the more robust to. Likewise, vowels with narrow bandwidths are better maskers than wide formant bandwidth vowels.

7.1 Future work

One of the most important aspects of this project is reliability of the formant bandwidth measures. Thus it is obvious that an improvement of the measurement methods could bring new possibilities in the field of phonetic and speech processing. Moreover it would be worth to have a review of collected comparative evaluations of the latest techniques for formant features extraction.

Regarding the database it would be interesting to merge the advantages of the 3 datasets we used in this project. For example a database which gathers different speaking-style over a large number of subjects and that includes normal and Lombard speech for each speaker as well as EEG recordings of each audio sample.

Concerning the psychoacoustic experiment, the main problem in this study was the lack of data per subject due the non-double forced choice. So the implementation of a double vowel forced choice, would surely accentuate the results.

Bibliography

[1] A. Amano-Kusumoto and J.P. Hosom. A review of research on speech intelligibility and correlations with acoustic features. 2011.

[2] R. Baker and V. Hazan. Lucid: A corpus of spontaneous and read clear speech in british english. In *Proceedings of the DiSS-LPSS Joint Workshop 2010*, 2010.

[3] G. Bapineedu, B. Avinash, S.V. Gangashetty, and B. Yegnanarayana. Analysis of lombard speech using excitation source information. In *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[4] A. de Cheveigné. Formant bandwidth affects the identification of competing vowels. *ICPhS, 2093*, 2096, 1999.

[5] A. De Cheveigné and H. Kawahara. Yin, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111:1917, 2002.

[6] T. Drugman and T. Dutoit. Glottal-based analysis of the lombard effect. In *Eleventh Annual Conference of the International Speech Communication Association*. Citeseer, 2010.

[7] R. Drullman, J.M. Festen, and R. Plomp. Effect of reducing slow temporal modulations on speech reception. *Journal of the Acoustical Society of America*, 95(5):2670–2680, 1994.

[8] J.R. Dubno and M.F. Dorman. Effects of spectral flattening on vowel identification. *The Journal of the Acoustical Society of America*, 82(5):1503, 1987.

[9] M. Fallon, S.E. Trehub, and B.A. Schneider. Children's perception of speech in multitalker babble. *The Journal of the Acoustical Society of America*, 108:3023, 2000.

[10] G. Fant and TV Ananthapadmanabha. Truncation and superposition. *STL-QPSR 2*, 1982.

[11] H.M. Hanson. Glottal characteristics of female speakers: Acoustic correlates. *Journal of the Acoustical Society of America*, 101(1):466–481, 1997.

[12] V. Hazan and D. Markham. Acoustic-phonetic correlates of talker intelligibility for adults and children. *The Journal of the Acoustical Society of America*, 116:3108, 2004.

[13] T. Johnstone, C.M. Van Reekum, T. Bánziger, K. Hird, K. Kirsner, and K.R. Scherer. The effects of difficulty and gain versus loss on vocal physiology and acoustics. *Psychophysiology*, 44(5):827–837, 2007.

[14] J.C. Junqua. The lombard reflex and its role on human listeners and automatic speech recognizers. *Journal of the Acoustical Society of America*, 1993.

[15] D.H. Klatt. Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America*, 67(3):971–995, 1980.

[16] D.H. Klatt and L.C. Klatt. Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America*, 1990.

[17] J.C. Krause and L.D. Braida. Acoustic properties of naturally produced clear speech at normal speaking rates. *The Journal of the Acoustical Society of America*, 115:362, 2004.

[18] H. Kuwabara and K. Ohgushis. Role of formant frequencies and bandwidths in speaker perception. *Electronics and Communications in Japan (Part I: Communications)*, 70(9):11–21, 1987.

[19] S. Kwang-deok and S. Wonyong. A robust formant extraction algorithm combining spectral peak picking and root polishing. *EURASIP Journal on Advances in Signal Processing*, 2006, 2006.

[20] E. Lombard. Le signe de l'élevation de la voix. *Ann. Maladies Oreille, Larynx, Nez, Pharynx*, 37(101-119):25, 1911.

[21] Y. Lu and M. Cooke. Speech production modifications produced by competing talkers, babble, and stationary noise. *The Journal of the Acoustical Society of America*, 124:3261, 2008.

[22] D. Markham and V. Hazan. The effect of talker-and listener-related factors on intelligibility for a real-word, open-set perception test. *Journal of Speech, Language, and Hearing Research*, 47(4):725, 2004.

[23] R.B. Monsen and A.M. Engebretson. The accuracy of formant frequency measurements: A comparison of spectrographic analysis and linear prediction. *Journal of speech and hearing research*, 26(1):89, 1983.

[24] A. Potamianos and P. Maragos. Speech formant frequency and bandwidth tracking using multiband energy demodulation. *Journal of the Acoustical Society of America*, 99(6):3795–3806, 1996.

[25] S Sekimoto. Effects of formant peak emphasis on vowel intelligibility in frequency compressed. *Ann. Bull. RILP.*

[26] Q. Summerfield, J. Foster, R. Tyler, and P.J. Bailey. Influences of formant bandwidth and auditory frequency selectivity on identification of place of articulation in stop consonants. *Speech Communication*, 4(1-3):213–229, 1985.

[27] G.K. Vallabha and B. Tuller. Systematic errors in the formant analysis of steady-state vowels. *Speech communication*, 38(1-2):141–160, 2002.

[28] T. Wempe and P. Boersma. The interactive design of an f0-related spectral analyser. *Proc. of the 15th Int. Congr. Phonetic Sci. in Barcelona*, 1:343–346, 2003.

Appendix A

Ranking of speakers in terms of their intelligibility on UCL Markham word test

Speaker group	Speaker	Error rank	Triplet error rate (%)	Word error rate (%)	Mean error rate(%)
AF	af-06	1	4.3	2.9	3.6
AF	af-14	2	4.4	5.1	4.8
AF	af-12	5	6.2	5.1	5.6
AF	af-02	6	6.7	5.0	5.9
AF	af-21	7	4.0	8.2	6.1
AF	af-10	9	7.4	5.3	6.4
AF	af-09	11	5.6	7.7	6.7
AF	af-13	14	8.0	6.0	7.0
AF	af-16	16	8.4	5.6	7.0
AF	af-11	17	7.0	7.2	7.1
AF	af-19	19	8.2	6.2	7.2
AF	af-04	20	7.2	7.2	7.2
AF	af-18	24	6.9	9.3	8.1
AF	af-17	25	7.3	9.2	8.3
AF	af-08	35	11.4	9.9	10.6
AF	af-07	37	11.8	10.6	11.2
AF	af-15	41	12.7	17.2	15.0
AF	af-03	43	17.5	16.6	17.1
AM	am-10	3	5.8	3.9	4.8
AM	am-08	4	6.3	4.5	5.4
AM	am-07	8	7.4	5.2	6.3
AM	am-19	13	6.3	7.4	6.8
AM	am-05	21	8.3	6.5	7.4
AM	am-09	23	9.5	6.4	7.9
AM	am-02	26	8.7	8.1	8.4
AM	am-06	27	10.4	7.5	8.9
AM	am-03	29	11.4	8.1	9.7
AM	am-18	30	11.6	7.9	9.8
AM	am-16	33	10.0	10.4	10.2
AM	am-17	34	11.8	9.4	10.6
AM	am-12	40	14.9	15.0	14.9
AM	am-13	44	14.8	19.7	17.3
AM	am-14	45	17.8	19.8	18.8
CF	cf-01	12	6.7	6.7	6.7
CF	cf-04	15	6.5	7.5	7.0
CF	cf-06	28	9.5	9.6	9.5
CF	cf-08	31	9.3	10.3	9.8
CF	cf-03	38	14.4	11.5	12.9
CF	cf-09	39	11.6	17.0	14.3
CM	cm-04	10	6.2	6.7	6.4
CM	cm-05	18	6.9	7.3	7.1
CM	cm-02	22	8.6	6.9	7.7
CM	cm-01	32	11.0	9.3	10.2
CM	cm-03	36	9.1	13.3	11.2
CM	cm-06	42	17.8	16.2	17.0

Figure A.1: Error rates obtained for adult female speakers (AM), adult male speakers (AM), child female (CF) and child male (CM) speakers, aggregated over all listener groups (adults, older children, younger children)

Appendix B

Central frequencies and formants bandwidth of French synthetic vowels

Vowel	Parameter	Formant 1	Formant 2	Formant 3	Formant 4	Formant 5
a	Central frequency	742	1266	2330	3457	4230
	Bandwidth	90	110	170	250	300
e	Central frequency	395	2027	2552	3438	4331
	Bandwidth	90	110	170	250	300
i	Central frequency	252	2202	3242	3937	4419
	Bandwidth	90	110	170	250	300
o	Central frequency	399	829	2143	3445	4191
	Bandwidth	90	110	170	250	300
u	Central frequency	276	733	2171	3506	4064
	Bandwidth	90	110	170	250	300

Figure B.1: Parameter values for the French synthetic vowels

Appendix C

Perceptual Experiment

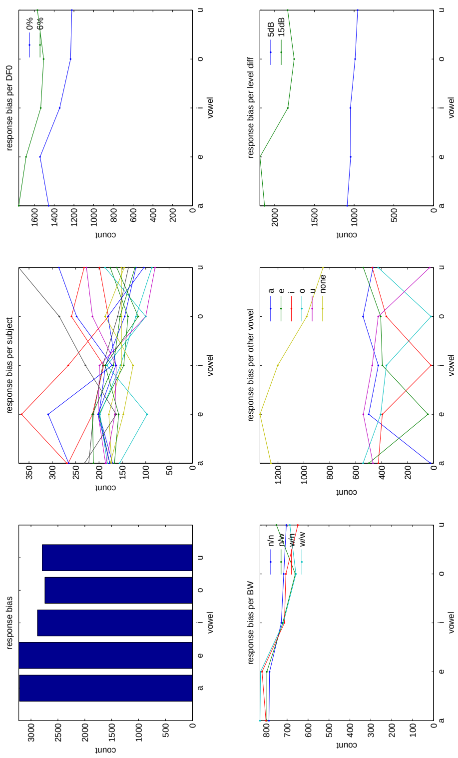


Figure C.1: statistical analysis of the perceptual experiment: part 1

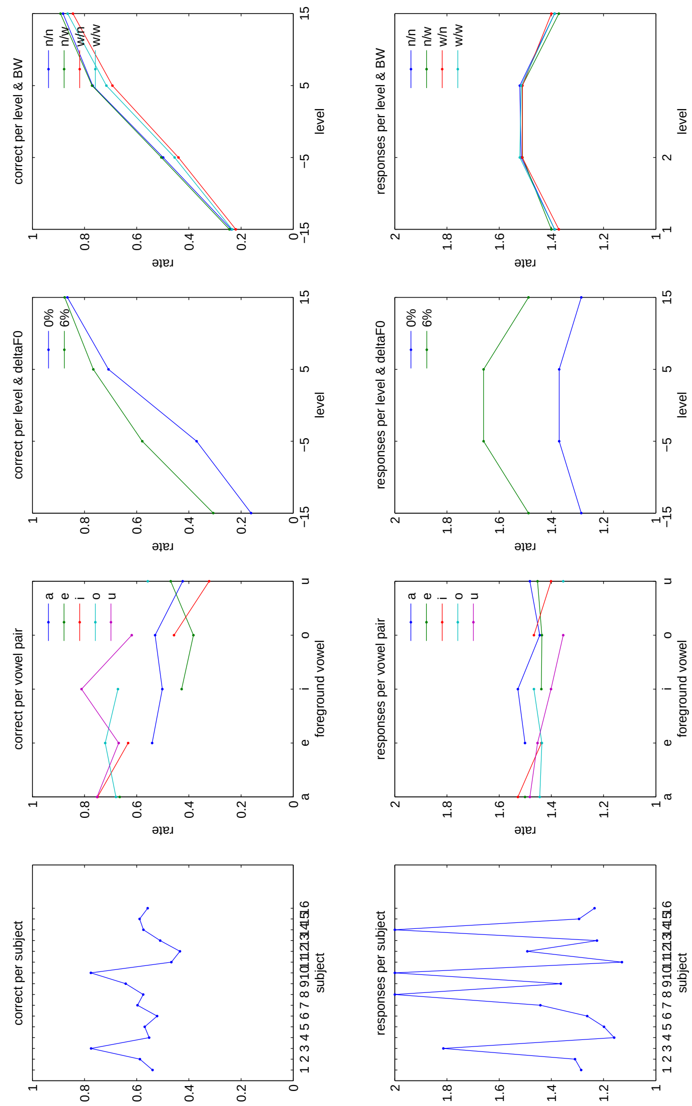


Figure C.2: statistical analysis of the perceptual experiment: part 2

Appendix D

Anova

Source		Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power[a]
bw	Sphericity Assumed	,466	,055	2,595	,223
	Greenhouse-Geisser	,451	,055	2,162	,204
	Huynh-Feldt	,466	,055	2,595	,223
	Lower-bound	,367	,055	,865	,141
F0	Sphericity Assumed	,000	,768	49,643	1,000
	Greenhouse-Geisser	,000	,768	49,643	1,000
	Huynh-Feldt	,000	,768	49,643	1,000
	Lower-bound	,000	,768	49,643	1,000

Figure D.1: ANOVA for relative level of -15dB

Source		Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power[a]
bw	Sphericity Assumed	,000	,328	21,946	,976
	Greenhouse-Geisser	,002	,328	15,976	,930
	Huynh-Feldt	,001	,328	18,804	,957
	Lower-bound	,016	,328	7,315	,715
F0	Sphericity Assumed	,000	,853	87,207	1,000
	Greenhouse-Geisser	,000	,853	87,207	1,000
	Huynh-Feldt	,000	,853	87,207	1,000
	Lower-bound	,000	,853	87,207	1,000

Figure D.2: ANOVA for relative level of -5dB

Source		Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power[a]
bw	Sphericity Assumed	,000	,474	40,482	1,000
	Greenhouse-Geisser	,000	,474	29,469	,997
	Huynh-Feldt	,000	,474	34,684	,999
	Lower-bound	,002	,474	13,494	,929
F0	Sphericity Assumed	,000	,764	48,511	1,000
	Greenhouse-Geisser	,000	,764	48,511	1,000
	Huynh-Feldt	,000	,764	48,511	1,000
	Lower-bound	,000	,764	48,511	1,000

Figure D.3: ANOVA for relative level of 5dB

Source		Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power[a]
bw	Sphericity Assumed	,001	,299	19,201	,955
	Greenhouse-Geisser	,006	,299	12,242	,858
	Huynh-Feldt	,004	,299	13,988	,893
	Lower-bound	,023	,299	6,400	,658
F0	Sphericity Assumed	,315	,067	1,082	,164
	Greenhouse-Geisser	,315	,067	1,082	,164
	Huynh-Feldt	,315	,067	1,082	,164
	Lower-bound	,315	,067	1,082	,164

Figure D.4: ANOVA for relative level of 15dB

Appendix E

Lombard effect analysis

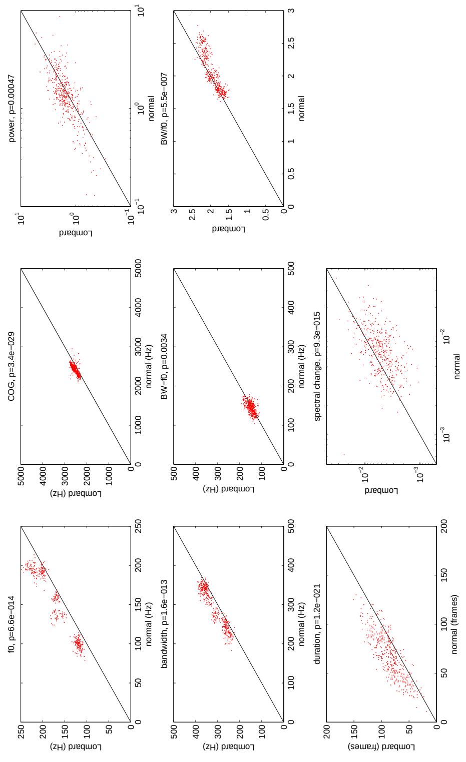


Figure E.1: Statistical analysis of the Lombard database