Proceedings eNTERFACE'12

The 8th International Summer Workshop on Multimodal Interfaces

July 2nd - July 27th 2012 ; Supélec, Metz, France

Prof. Olivier Pietquin, Chair

Supélec
Metz Technopôle
2 rue Edouard Belin
57070 Metz, France

Ph.: +33 (0)3 87 76 47 70
Fax : +33 (0)3 87 76 47 00
Olivier.Pietquin@Supelec.fr
http://malis.metz.supelec.fr/~pietquin
http://enterface12.metz.supelec.fr

eNTERFACE 2012 - Project reports

Project	Title	Pages
P1	Speech, gaze and gesturing - multimodal conversational interaction with Nao robot	7-12
P2	Laugh Machine	13-34
P3	Human motion recognition based on videos	35-38
P5	M2M - Socially Aware Many-to-Machine Communication	39-46
P6	Is this guitar talking or what!?	47-56
P7	CITYGATE, The multimodal cooperative intercity Window	57-60
P8	Active Speech Modifications	61-82
P10	ArmBand : Inverse Reinforcement Learning for a BCI driven robotic arm control	83-88.

P8 - Active Speech Modifications

Yannis Stylianou, *Valerie Hazan*, Vincent Aubanel, Elizabeth Godoy, Sonia Granlund, Mark Huckvale, Emma Jokinen, Maria Koutsogiannaki, Pejman Mowlaee, Mauro Nicolao, Tuomo Raitio, Anna Sfakianaki, Yan Tang

1 Introduction

In many intelligibility studies, it was demonstrated that the speaking style referred to as clear speech is significantly more intelligible than conversational (or casual) speech. This intelligibility gain exists for both normal-hearing and hearing-impaired listeners (e.g. elderly persons and linguistically inexperienced listeners like non-native (L2) speakers and children). Also, in a two-way conversation in which one person is affected by an adverse listening condition and one is not (e.g. between one person speaking to another via telephone where the other is in a noisy club, or in a cafeteria, in the street etc.), the person who is not affected still manages to make adaptations (on acoustic-phonetic and linguistic levels) that are quite specifically tailored to counteract the specific communication barrier that the other person is experiencing. These adaptations show that clear speech is not defined in a uniform way, but that there are different styles of clear speech depending on the adverse condition that the speech is heard in. In this context, Active Speech Modifications refer to the speaking-style adaptations or strategies a speaker applies in order to maximize communication effectiveness.

Identification and effective manipulations of the most prominent acoustic-phonetic characteristics of different styles of clear speech can allow for the development of new, signal based, active speech modification algorithms to increase intelligibility. The algorithms can consequently improve speech intelligibility in many situations, such as in the design of hearing aids, telephony, and other speech signal processing technologies and applications (i.e., speech synthesis, recognition, enhancement, etc).

The purpose of this project was to use modern speech analysis and reconstruction algorithms to:

- identify which acoustic-phonetic characteristics are prominent in different styles of clear speech (e.g. babble-countering clear speech, vocoder-countering clear speech, L2-"countering" clear speech) and when they are realized in time.
- model a selection of these aspects so that they can be applied automatically on speech, to enact prosodic changes, changes in amplitude spectrum, modulation frequencies, etc..
- run a series of "proof of concept" perception experiments to see if the "specifically-enhanced" speech is better perceived in the "matched" adverse condition than other types of clear speech (there is evidence that this is the case with the naturally-enhanced speech).

The outcome of the project can be summarized as follows:

- a new speech corpus (P8-Harvard corpus) was linguistically and meta-linguistically annotated and acoustically analyzed with the goal of identifying which acoustic-phonetic characteristics differ between clear and casual speech and also between different styles of clear speech. Moreover, acoustic analyses on specific features were also performed on a different corpus, namely the LUCID database (specifically on read clear and read casual speech signals).
- among the different styles of clear speech, prosodic changes were most apparent. Therefore, signal modification algorithms were developed to mimic human adaptations on prosody in adverse conditions with the aim of increasing intelligibility.
- a user-friendly interface, XPlic8, for a large range of acoustic analyses was developed.
- a set of evaluation experiments was prepared to evaluate the different modifications.

This report is organized as follows. Section 2 describes the P8-Harvard corpus that contains the different speaking styles for analysis. In section 3 the linguistic analysis of the corpus is presented. Section 4 focuses on the analysis of the voice source characteristics between different styles of speech on the P8-Harvard corpus (and on the LUCID corpus to a less extent). In section 5 prosodic differences between the different speaking styles are examined with focus on the number of pauses and the mean word duration. Section 6 introduces two novel time-scaling techniques that try to modify casual speech signals to achieve higher intelligibility scores, mimicking the properties of the elicited clear speech. Section 7 presents a novel tool for a large range of acoustic analyses. Section 8 summarizes the work of this project.

2 P8-Harvard Corpus design and recording

A corpus of materials was recorded and analyzed to provide information about the acoustic phonetic enhancements typically seen in clear speaking styles produced in speech with communicative intent. The aim was to record materials which were controlled and standardized (Harvard sentence lists) but where clear speaking styles were elicited naturally, due to communicative need, rather than via instructions to read materials clearly (LUCID corpus[1]). For that purpose, the first 15 lists of the Harvard sentences (1969) were recorded. These sentences, which are phonetically-balanced and each include 5 keywords, were developed for speech quality evaluations.

2.1 Recording procedure

Two British English speakers, one female and one male were each recorded (as "Speaker A") with a confederate ("Speaker B"). Speaker A had to read a sentence to Speaker B who had to repeat it back to Speaker A. So as to induce Speaker A to make an effort to speak clearly when Speaker B was experiencing a communication barrier, speakers were told that the speaker pair that achieved best "intelligibility scores" would win a prize. Speaker A was told to only say the sentence once even if errors were made by speaker B in repeating it. Two types of communication barrier, following Hazan and Baker (2011), were used in order to elicit clear speaking styles that may differ somewhat in their acoustic-phonetic characteristics. In the "babble" (BAB) condition, Speaker B heard speaker A's voice mixed with 8-speaker babble noise at an approximate level of 0 dB SNR; in the "vocoder" (VOC) condition, Speaker B heard speaker A's voice passed through a three-channel noise-excited vocoder which spectrally degraded the signal. 150 sentences in each of the three conditions ("no barrier" NB, BAB, VOC) were recorded for the two speakers.

Speakers were seated in separate sound-treated rooms. Beyerdynamic DT297PV headsets fitted with a condenser cardioid microphone were used and the speech was recorded on two separate channels at a sampling rate of 44100 Hz (16 bit) using an EMU 0404 USB audio interface and Adobe AUDITION. Only Speaker A's output was analysed here, since speaker A was talking in a non-barrier environment.

3 Linguistic analysis of the P8-Harvard Corpus

For the linguistic analysis of the P8-Harvard corpus, Praat [2] along with several analysis algorithms was used.

3.1 Initial processing

For all sentences, a Praat textgrid was produced with three tiers: tier 1 contains speech (SP) and silent (SILP) regions markers, tier 2 had word aligned markers and tier 3 phoneme-level aligned markers. Sentences (five sentences for Speaker A1 and 12 for Speaker A2) were excluded from the corpus since they contained mispronunciations or hesitations on one or more of the keywords.

3.2 Linguistic Annotation of corpus

The Harvard database [3] is a set of 72 phonetically balanced lists of 10 sentences, each containing 5 keywords. Three lists were recorded for the current project, and in addition to existing keyword coding, the database was enriched with broad/narrow grammatical annotation, lexical frequency and neighborhood density. A summary of the added information to the Harvard database is given in Table 1. Word- and phone-level annotation were semi-automatically carried out and merged with the Harvard

database. The resulting corpus comprises of 2293 manually check words and 6902 segments for the two speakers in the three recording conditions.

Information	Description
word	Orthographic form of the word (punctuation removed)
lemma	Lemma of the word
keyword	Keyword coding of the word (keyword vs. non-keyword)
PoS	Part of speech. Categories are: Adj, Adv, Conj, Det, DetP, Ex, NoC, Num, Prep, Pron, Verb, VMod
freqBNC	BNC[1] frequency of occurrence of the word (inflected form). Occurrence per million in a 100 million spoken and written word corpus
neighPhon	Number of all phonological neighbours that differ from the word by a 1-phoneme substitution, deletion, or addition. Extracted from the de-Cara database[2].
freqCxS	Celex spoken frequency of corresponding lemma. Occurrence per million in a 17.9 million spoken word corpus
freqCxW	Celex written frequency of corresponding lemma. Occurrence per million in a 17.9 million written word corpus

Table 1: Harvard database annotation tagging. 3 lists were annotated for a total of 1066 words.

3.3 Analysis of communication effectiveness

The number of correctly-transmitted keywords was calculated per condition. The percentage of keywords correct in the BAB condition was 88% for A1 and 73 % for A2, while in the VOC condition it was about 40% for both speakers. The VOC condition was therefore harder for both speaker pairs.

Communication breakdowns were defined as sentences in which 3 or more keywords were not correctly repeated by speaker B. Fig. 1 highlights these breakdowns and the sentences immediately following them (see also Section 4.6).

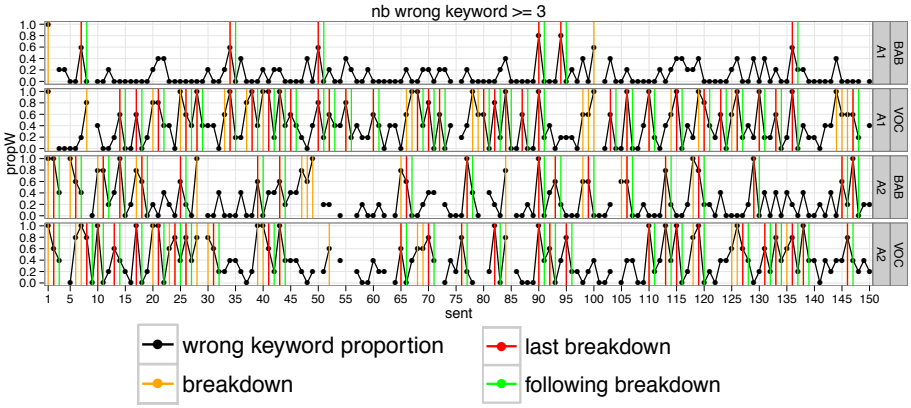

Figure 1: *Identification of communication breakdowns for speaker A1 and A2 in BAB and VOC conditions, defined as sentences in which more than 3 keywords were missed in the interlocutor repetition. A distinction is made between ''breakdowns" (orange lines) and ''last breakdowns" (red lines), the latter depicting breakdowns immediately followed by sentences in which 2 or less keywords were missed (''following breakdowns", green lines).*

[1]British National Corpus. Available online at `http://ucrel.lancs.ac.uk/bncfreq`
[2]de-Cara database. Available online at `http://portail.unice.fr/jahia/page12414.html`

4 Analysis of the voice source and spectral characteristics between different styles of speech

The analysis of voice source characteristics of the P8 corpus included three types of speech: NB, BAB and VOC speech. The idea was, that if there would exist differences between the voice source characteristics of these two voice types, this information could used to convert normal speech into the more intelligible speech in the barrier cases.

The main analysis tools for this task were glottal inverse filtering, pitch detection, glottal closure instant detection, voice source feature extraction and formant detection using Praat. These tools were developed as Matlab scripts for the purpose of the project and details regarding the analysis tools are provided in section 7, since all these analysis algorithms were incorporated in a new proposed analysis tool.

4.1 Glottal flow waveforms and Harmonic analysis

Main findings were that the different voice types did not differ significantly in terms of the use of the voice source. Figure 2 shows the glottal source waveform for the speech signals on the three different conditions, BAB, VOC and NB for the male and female speaker. In Figure 3 the corresponding spectra of the glottal source are depicted. Figure 4 shows a slight decrease of the harmonic-to-noise ratio in the barrier cases for both speakers A1 and A2.

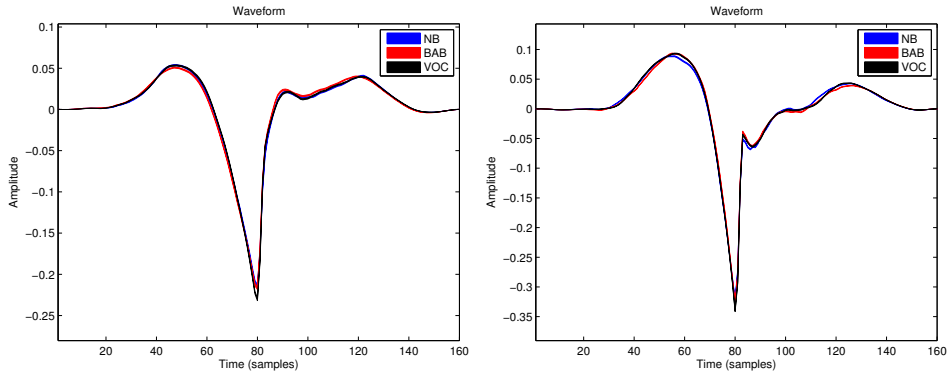

Figure 2: *Glottal source waveform for the speech signals on the three different conditions, BAB, VOC and NB for the male(a) and female speaker(b)*

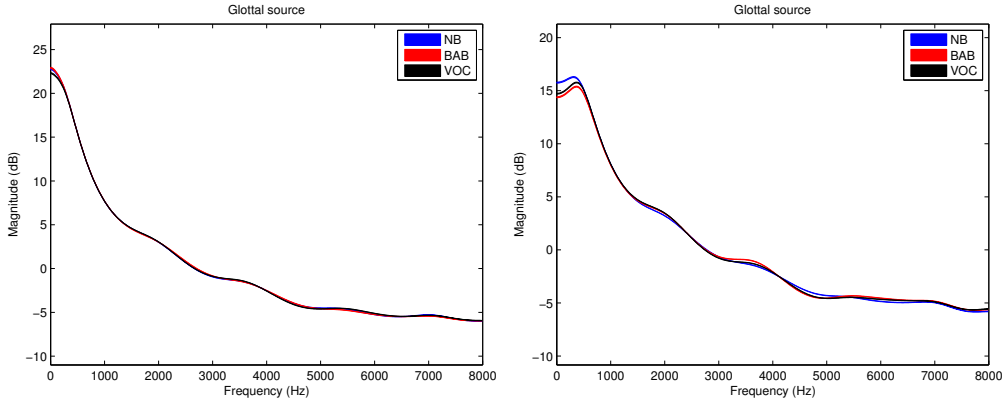

Figure 3: *Glottal source waveform for the speech signals on the three different conditions, BAB, VOC and NB for the male(a) and female speaker(b)*

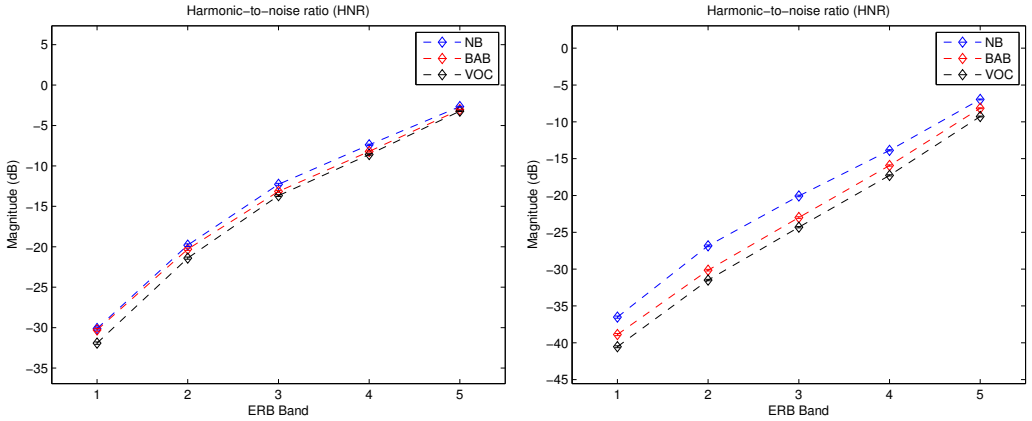

Figure 4: *Harmonic-to-noise ratio for the three different conditions NB, BAB, VOC for the male speaker A2 (left) and the female speaker A1 (right).*

4.2 F0 analysis

The F0 detection is based on glottal inverse filtering and autocorrelation peak detection. The algorithm implemented to extract the F0 and the F0 range from the speech signals is described on section7. These estimated values for the the whole P8-Harvard corpus where statistically analyzed with ANOVA. As expected F0 median was higher for the female speaker $[F(1,137) = 16343.6, p < 0.001]$; it was also higher in the VOC condition than in the NB $[t = 19.3; p < 0.001; df = 132]$ and BAB conditions$[t = -10.7; p < 0.001; df = 132]$, and higher in the BAB than NB conditions $[t = -7.6; p < 0.001; df = 132]$. F0 range also varied across conditions $[F(2,274) = 9.5; p < 0.001]$: it was broader in BAB than in both NB $[t = -2.4; p = 0.018; df = 132]$ and VOC $[t = 4.8; p < 0.001; df = 132]$. However, F0 range did not differ between the NB and VOC conditions $[t = 1.8; p = 0.067; n.s.; df = 132]$. The interaction between speaker and condition was also significant $[F(2,274) = 3.9; p = 0.02]$.

4.3 LTAS

The Long Term Amplitude Spectra (LTAS) were also estimated for the P8-Harvard and the LUCID corpus (the algorithm for the estimation of LTAS is described in Section 7. Previous studies correlate the increase of intelligibility of clear speech with the higher energy in the frequency band 1-3kHz relative to casual speech. Figure 5 depicts the LTAS for speakers A2 (left) and A1 (right) correspondingly for the barrier and no barrier conditions of the P8-Harvard corpus. The male speaker increases his energy above 1000Hz especially for the VOC condition and less on the BAB. For the female speaker there is a slight increase between 2000-4000Hz for the BAB condition and a significant increase above 5000Hz.

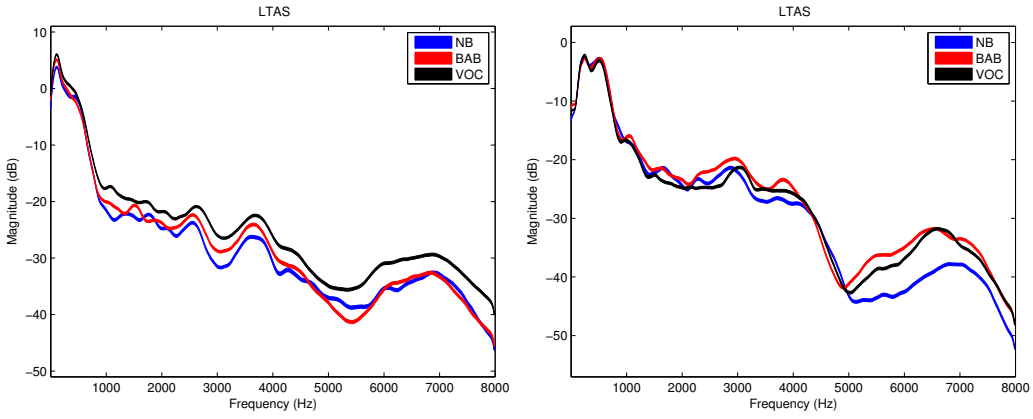

Figure 5: *LTAS of the male (left) and female speaker (right) for three different conditions NB, BAB and VOC*

For the 21 speakers in the LUCID database, averages over all voiced frames of the speaker were computed. The obtained results indicated that for most speakers, the spectral tilt decreases from CV to CL speech. In addition, some energy reallocation to the 1-7 kHz frequency region took place for most speakers. An example of a computed LTAS is shown in Fig. 6 for speaker F38 where the previously mentioned effects can clearly be seen.

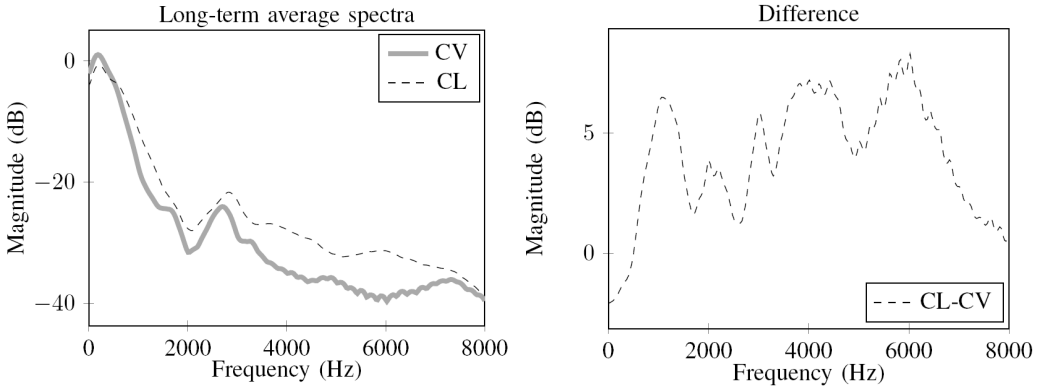

Figure 6: *The long-term average spectra (LTAS) of conversational (CV) and clear (CL) speech and their difference for female speaker F38 in the LUCID database. The LTAS are computed over four sentences for each condition.*

However, the results also varied significantly across speakers. For instance, the energy reallocation patterns were in most cases very different and furthermore, for some speakers the spectral tilt was further increased. This indicates that the speakers used very different strategies to produce clear speech.

A repeated measures ANOVA was done on the measure of intensity (LTAS $1-3kHz$) for the P8-Harvard corpus. LTAS was calculated separately for each sentence. There was a main effect of speaker $[F(1,131) = 22.0; p < 0.001]$, and of condition $[F(2,262) = 547.8; p < 0.001]$; post-hoc paired t-tests show that the BAB condition was greater in intensity ($mean = -3.1$) than the VOC (mean= -3.6) ($t = 4.8; df = 131; p < 0.001$) and NB conditions ($mean = -6.9$) ($t = -26.8; df = 131; p < 0.001$). There was also a significant interaction of speaker and condition $[(F(2,262) = 124.7; p < 0.001]$; post-hoc analyses show that there are significant speaker-specific strategies in terms of intensity ($t = -15.4; df = 131, p < 0.001$): for A1, the BAB condition has a greater intensity than the VOC condition (mean difference between VOC and $BAB = -2.3$), while for A2, the VOC condition has a greater intensity than the BAB condition (mean difference between VOC and $BAB = 1.2$).

4.4 Energy distribution in critical bands

A sinusoidal signal analysis/synthesis mode was used to check the differences between the clear and causal speech on the LUCID corpus. The idea was to investigate the differences between the two speaking styles, clear and casual speech, according to their sinusoidal features (including amplitude and frequency) extracted at the designed critical bands. For this, a pitch-independent sinusoidal model is designed which extracts one sinusoid per critical bands, hence with a fixed dimension equal to the number of critical bands. To design the critical frequency bands we used the 24 center frequencies and bandwidth derived at 16 kHz of sampling frequency. In order to reflect more accurately the subjective loudness of speech signal for the masker noise, the $ITU - R468$ noise weighting filter was taken into consideration. The highest spectral amplitude per frequency band was selected to avoid sidelobe peak problem. This modified the center frequency and bandwidth of some of the critical bands. The sinusoidal model designed as such showed a hardly distinguishable difference between the re-synthesized and the original signal.

Experiments were conducted on voiced frames of length 16 ms with a frame shift of 4 ms for two speakers, M8 (male) and F22 (female), of the LUCID database. Figure 7 shows the histogram of the amplitude (top) and frequency (bottom) of clear (blue) and casual speech (red) at a specific critical sub-band characterized with its center frequency and bandwidth for the male speaker M8. Figure 8 shows the histogram of the the amplitude (top) and frequency (bottom) of clear (blue) and casual speech (red) at a specific critical sub-band characterized with its center frequency and bandwidth for the female speaker F22. The center frequency and bandwidth for each critical band is shown at top of each subplot.

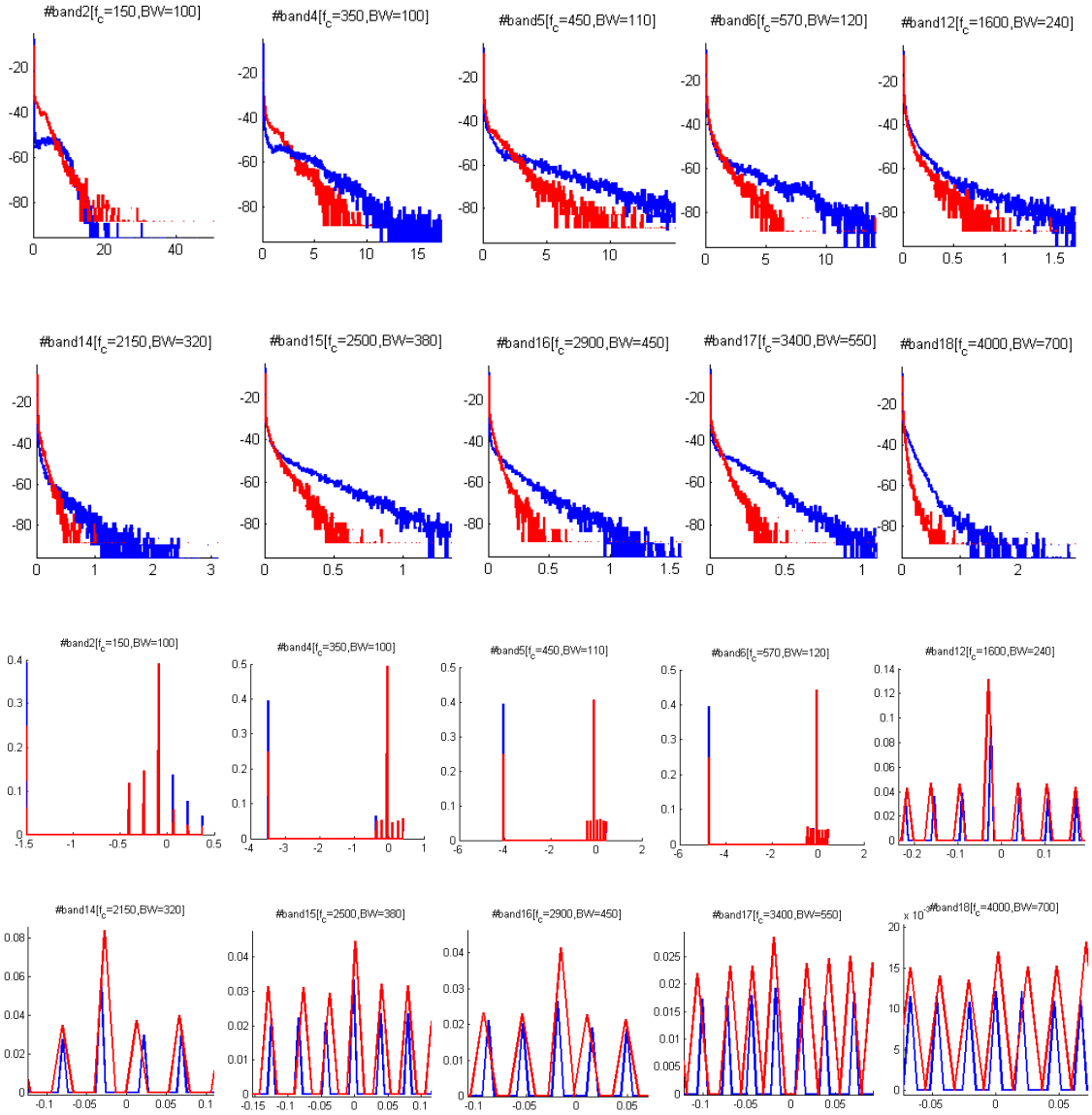

Figure 7: *Histograms of the amplitude (top) and frequency (bottom) of clear (blue) and casual speech (red) at a specific critical sub-band characterized with its center frequency and bandwidth for the male speaker M8*

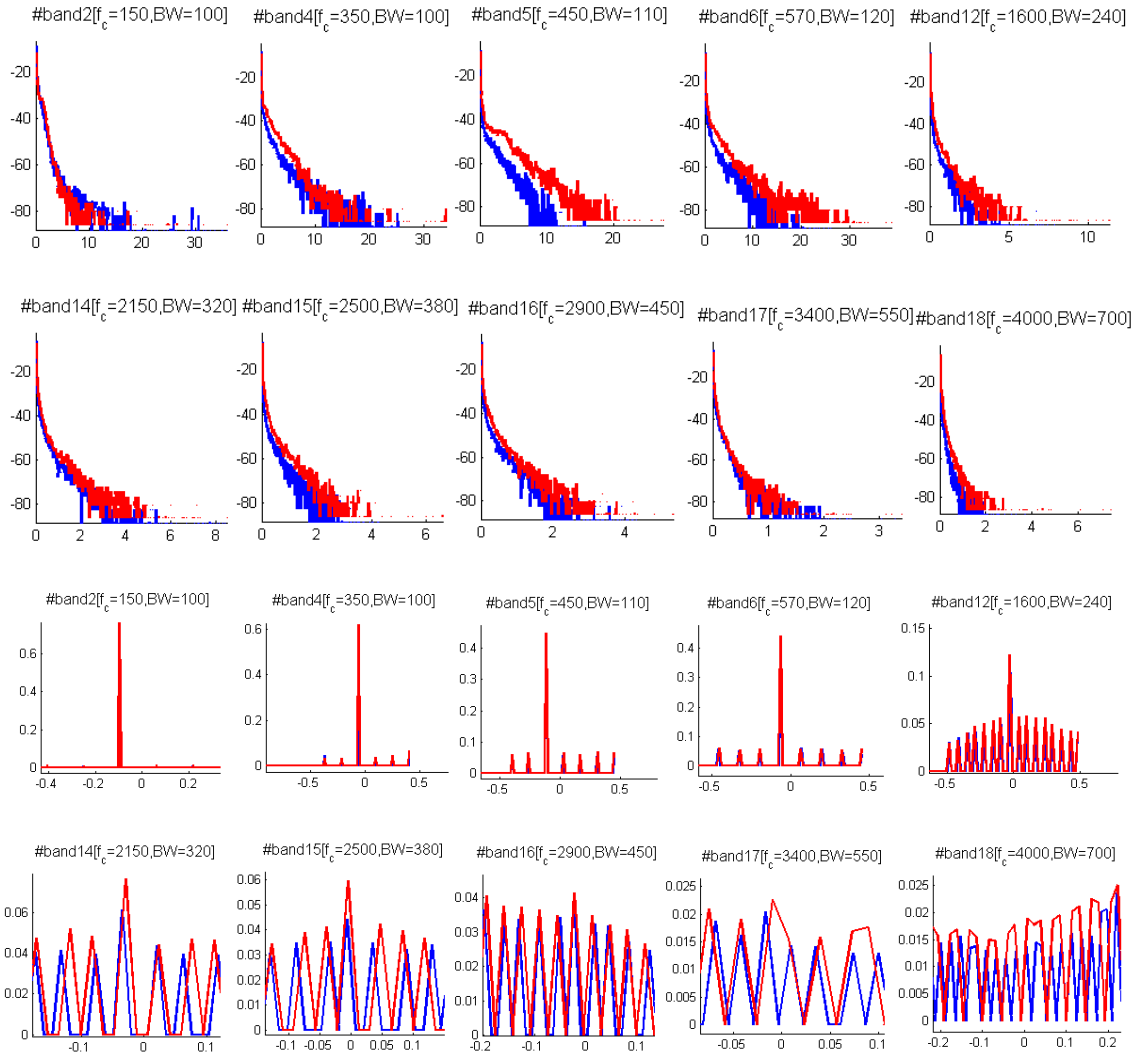

Figure 8: *Histograms of the amplitude (top) and frequency (bottom) of clear (blue) and casual speech (red) at a specific critical sub-band characterized with its center frequency and bandwidth for the female speaker F22*

The x-axis of each subplot is the range for amplitude or frequency. For frequency case, the x-axis is the frequency deviation from the center frequency (f_c) normalized by the bandwidth (BW) to make it a standard random variable called ($f_{standard}$):

$$f_{standard} = (f - f_c)/BW \tag{1}$$

The amplitude histogram figures, indicate the amount of energy difference per critical band between clear speech and casual speech. It is observed that for clear speech we have significantly more energy contribution than that for casual speech. This is well pronounced for frequency bands lying higher than 450 Hz. Looking at the changes in the frequency of clear and casual speech at critical bands, it is observed that the two speech styles have differences at frequency bands between 2000 and 4000 Hz (critical bands 13 to 18).

Future analysis can be performed in this domain. The idea is to find a way to model these differences between the barrier and no-barrier speech. Using the learned statistics, the final goal is to modify the barrier speech (causal speech), in terms of its sinusoidal parameters at critical bands, in order to improve the speech intelligibility. One possible idea is to increase the energy distribution of the causal speech at certain critical bands.

4.5 Vowel space

In order to visualize and quantize the vowel pronunciation of different speakers and styles, vowel spaces are useful. The vowel space is a plot of the mean of vowel instances in a 2D plane defined by the first and second formant frequencies. The area that the observed vowels span in this space then reflects the discriminability of the vowels. Previous studies report the expansion of vowel space in the case of clear elicited speech versus casual speech. The vowel spaces have been generated as follows. First, in order to isolate the vowel instances in the corpora, all of the speech was segmented using an HTK-based audio-to-text aligner. No manual corrections were performed. For each vowel instance, formant analysis is performed using the Praat algorithm [2]. The representative pair of F1 and F2 values for each vowel instance is then taken as the values at the center of the speech segment. For each vowel, the mean over all of the vowel instances is trimmed, with 95% of the data kept, in order to limit the influence of potential outliers. Then, the convex hull (i.e., a polygon fit that encompasses all of the data points) is calculated in order to represent the vowel space area. The convex hull is selected to represent the vowel space area in this work because it effectively captures the maximum area that the points in the vowel space span. Figure 9 depicts the vowel space in the three conditions for speakers A1 and A2 as defined by the largest-area polygon fit (convex hull) for the 4 tense and 6 lax vowels (95% trimmed means).

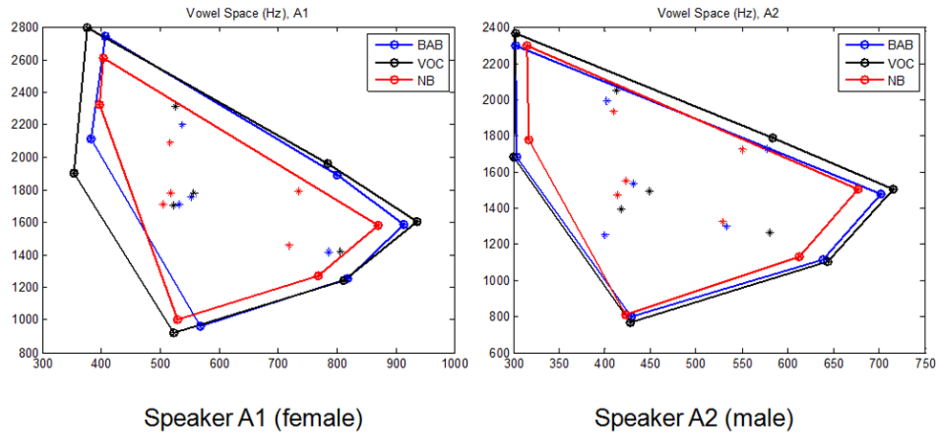

Figure 9: *Vowel space in the three conditions for speakers A1 and A2 as defined by the largest-area polygon fit (convex hull) for the 4 tense and 6 lax vowels (95% trimmed means).*

A per-vowel analysis was run on the measures using a mixed-model ANOVA, with vowel as a between-subjects factor, and condition (NB, BAB, VOC) as a within-subjects factor. The analysis showed a significant condition effect on all three vowels /i/, /ɒ/ and /ɔ/ for Speaker A1 ($p = 0.0398$) but no effect for Speaker B. So, for Speaker A, vowel space expands as follows: $NB < BAB < VOC$. Additionally, no significant interaction was found between vowel type and condition for either speaker.

4.6 Analysis of speech produced post communication breakdown

Sentences with more than two keywords incorrectly perceived were classified as having caused "communication breakdown". To find out whether there are significant differences between the pre- and post-breakdown sentences in terms of acoustic characteristics, for Speaker A1, acoustic analyses (i.e., sentence duration, LTAS at 1-3 kHz and 5-8 kHz, F0 median and range) were compared for breakdown sentences and post-breakdown sentences where all keywords were correct. In the BAB condition, there is a trend for longer sentence duration ($p = 0.163$) and significantly higher LTAS (5-8 kHz) post breakdown ($p < 0.05$). In VOC condition, effects were not significant but a trend for lower F0 median and higher F0 range post breakdown can be discerned. Although no correlation between sentence duration and communication effectiveness was found, a gradual increase in sentence duration was observed as time progressed in BAB condition for Speaker A1.

5 Examining prosodic differences between speech styles

The P8-Harvard corpus was also analyzed on time-domain, focusing on the number of pauses, mean word duration and the "rhythm" between the different speech styles.

5.1 Number and duration of pauses

In order to detect the number of pauses in the sentences of the whole P8-Harvard corpus, an algorithm was implemented to detect parts of speech signal with no proper speech content(NS, Not-Speech) such as pause between words, or even closure within stop consonants, etc. The silence detector relies on a *low-loudness detection* function based on the Perceptual Speech Quality measure. First the total loudness of the speech signal is computed by PSQ (ITU Standard REC-BS.1387-1-2001) and then the normalized loudness is computed dividing by the maximum loudness of the signal. A frame of the signal is considered NS if its normalized loudness is less than 15%. After cross-validation using a subset of files with manually-detected pauses (50 files from the P8-Harvard corpus) and it was found consistent. According to the linguistic context where the low-loudness part was located, the function could address the following type of Not-Speech:

- S: part of signal with loudness above threshold
- NS: generic low-loudness part of signal
- NS_{sil}: low-loudness part of signal at the beginning-end of the sentence
- NS_{sc}: low-loudness part of signal, which is part of a stop consonant inside a word
- NS_{iw}: low-loudness part of signal between two separate words (Inter-Word pause)
- $NS_{iwsc} : NS_{iw}$ in which the second word begins with stop consonant and therefore it is not possible to say if it is a pause or the closure of the consonant.

Applying the automatic detector to the P8-Harvard database, it was possible to compare the number of Not-Speech in different conditions. Table 2 contains the number of Not-Speech for each category, speaker and condition and Figure 10 has the average number of the total number of inter-word pauses per each utterance.

	A1			A2		
Type of NS	NB	BAB	VOC	NB	BAB	VOC
sc	437	492	*529*	433	470	*520*
iw	32	65	*155*	17	24	*75*
iwsc	176	209	*220*	130	162	*182*
sil	295	293	*298*	*277*	276	276
nc	1161	1386	*1615*	947	1059	*1247*

Table 2: *Number of instances of different type of NS*

The first results showed an increasing number of NS parts in the speech along with the difficulties in the communication. $\#NS_{VOC} > \#NS_{BAB} > \#NS_{NB}$ for both speakers, even though the male speaker tends to compensate less to the adverse conditions, as confirmed by other analysis.

A significant increase is worth to mention in the number of intra-words NS (NS_{iw}) between the VOC barrier and the other two conditions in both speakers as Figure 10 explicitly shows. This confirms that when the communication channel is really destructive and there is no direct feedback of it, the speaker focus the main part of his/her effort to greatly decrease the speaking rate.

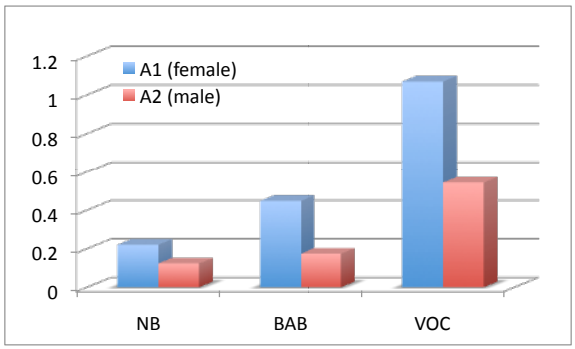

Figure 10: *Average number of inter-word pauses (NS_{iw}) for each utterance in different conditions.*

Further insight can be gained by looking at the durations of the different silence categories: Fig. 11 shows that, apart from leading/trailing silences (NS_{sil}), all types of silences undergo durational increase from NB to BAB to VOC, particularly the interword pauses (NS_{iw}). In contrast, speech part durations remain stable, highlighting a possible speaker strategy of reducing speech rate by detaching the words.

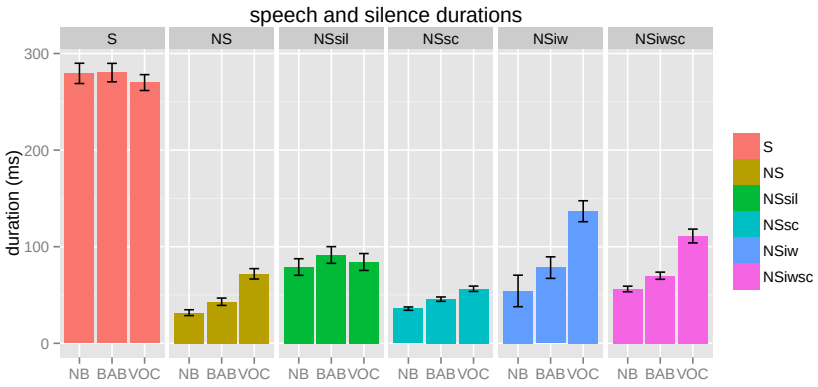

Figure 11: *Mean speech and silence durations for speakers A1 and A2 across NB, BAB and VOC. Errorbars are 95% confidence interval.*

5.2 Mean Word Duration analysis

The Mean Word Duration (MWD) for each type of condition was measured accurately using the silent detector, since the inter-word durations within utterances could be identified and subtracted to the word durations.

The results plotted in Figure 12 and Figure 13 display the change of duration in the VOC and BAB condition with respect to the No-Barrier condition during the experiments sessions. First observation is that all speakers elongate their speech production, especially in the worst condition (VOC barrier). This evolves along the sessions. However, this is not consistent between the two speakers for the BAB condition. Speaker A2 maintains mean word duration and mean content word duration stable. Speaker A2 was found to be less effective in the compensation, he slightly elongated the speech ($\sim 20\%$), only in the VOC barrier case but he didn't adjusted his speech any further. This lack of efficiency was confirmed by the amount of the errors the listener made which were much more compared to the errors he made during the session of speaker A1.

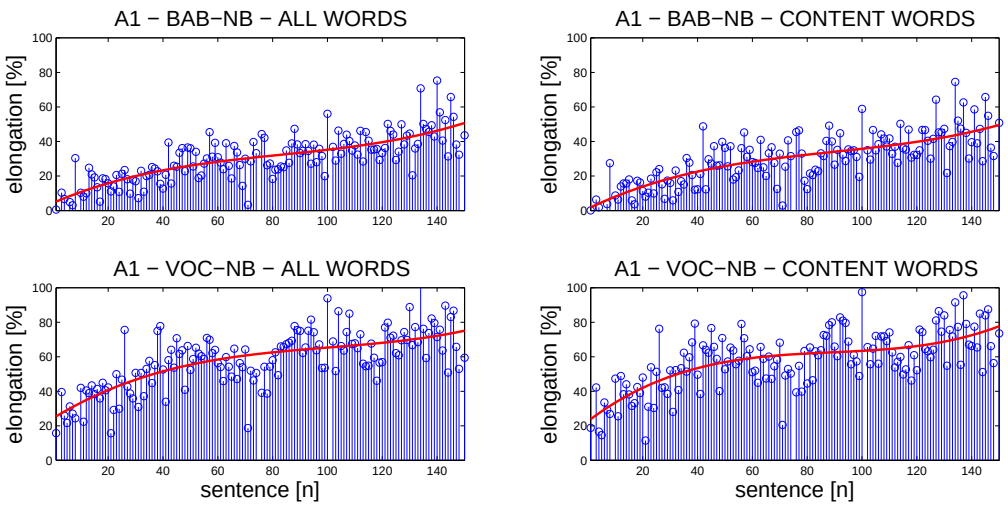

Figure 12: *Elongation strategy of female speaker A1 in the experiment sessions. On the left-hand side the mean word duration related to all words is shown, whereas on the right-hand side there is the mean content-word duration only.*

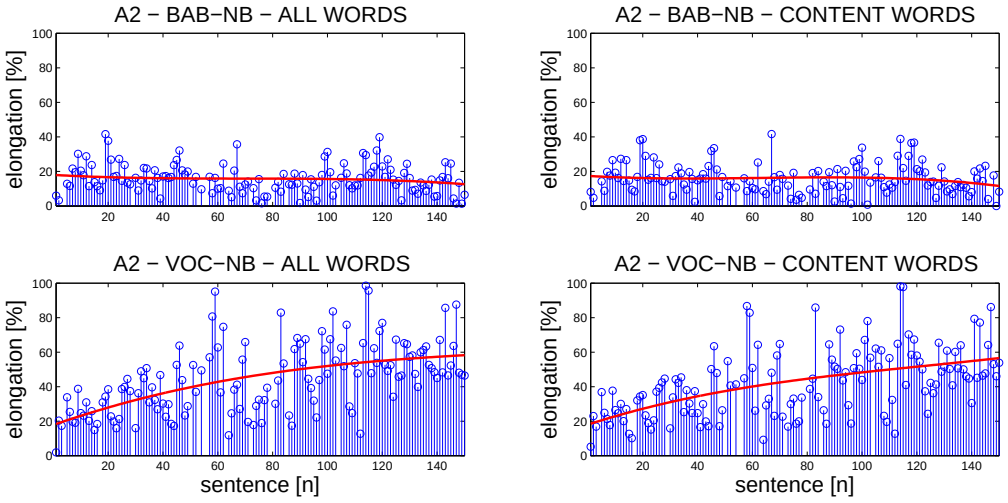

Figure 13: *Elongation strategy of male speaker A2 in the experiment sessions. On the left-hand side the mean word duration related to all words is shown, whereas on the right-hand side there is the mean content-word duration only.*

In Figure 12 and Figure 13 the red line is a 3rd-order polynomial curve that fits the data. Based on the shape of the line, three different stages emerge in all sessions, particularly for speaker A1 and the most stressful condition, i.e. the VOC barrier. At the beginning, the speaker starts with almost the same mean word duration as the NB condition, but as soon as he/she received intelligibility feedbacks from the listener, he/she increased the effort (i.e word duration) and hence a steep slope is seen at the beginning of the session. In the central part, the curve is flatter and the hypothesis is made that the current elongation is effective for the condition and the listener and no further adaptation is needed. In the final part, speaker A1 increased mean word duration further and it is hypothesised that she was trying to compensate the listener's tiredness, whereas, in the same conditions, speaker A2 seemed to cease making the effort to elongate, maybe due to a lack of motivation towards the end of the session.

Some correlations were investigated between the increase/decrease of mean word duration in a utterance and the number of listener's errors, but no clear relationship was found yet due to difficulties in comparing the two completely different data domains.

5.2.1 Rhythmogram analysis

The rhythmic patterns which differentiate barrier and no barrier speech was also investigated. For this task the rhythmogram [4] was employed. The Rhythmogram is a hierarchical representation of speech rhythm, from which one can extract the locations of relative prominences in the speech signal. This is achieved in a first step by computing auditory-based energy envelope with different time windows, and, by linking the peaks at different scales in a subsequent stage, enabling the identification of global (e.g., sentence-level) prominences (see Fig. 14). The detected prominences might undergo different

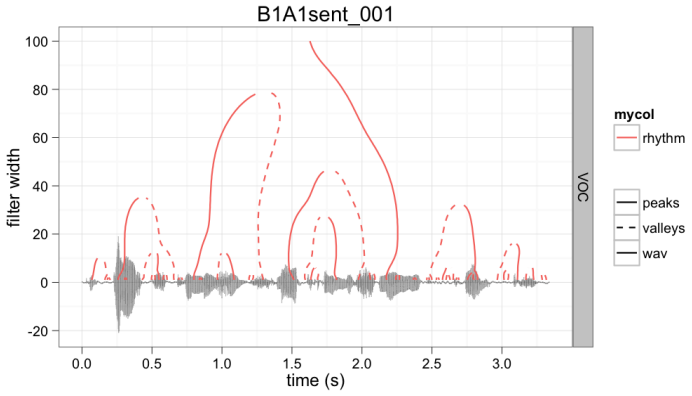

Figure 14: *Rhythmogram analysis for the sentence* "The birch canoe slid on the smooth planks". *Plain line stems identify relative prominences, dashed stems relative silences. Prominences and silences strength is determined by their highest value on the y-axis, and their location in the speech signal by their minimum value (smallest filter width).*

modifications by talkers in the reduction processes from clear to casual, and suggested a comparative analysis between clear and casual speech.

Using the manually annotated temporal mapping between clear and casual speech on a different Database (LUCID database), we assessed whether prominences and silences were treated differentially by talkers. Results on 69 pairs of matched casual/clear speech sentences showed that speech segments containing silences were significantly more compressed than prominences. ($p<.001$), Fig. 15. This shows that the nonlinearities observed in the temporal reduction from clear to casual can be explained by the rhythmic properties of speech: whereas silences appear to be suppressed from clear to casual, prominences tend to be preserved.

Given this result, prominence and silence locations were further characterized in terms of what sound class they fell in the P8-Harvard database comprising the NB and the two clear speech eliciting communicative barrier conditions BAB and VOC. The results presented in Fig. 16 show that most of the detected prominences fell into sonorant sound classes, i.e., vowels, nasals, and to a lesser extent, liquids. On the other hand, silences were found in stops, fricatives and annotated silences (occurring mainly in VOC condition, cf. Section 5.1). It should be noted here that the silences detected in the stop segments sound classes were mainly "low-level" silences in the rhythmogram hierarchy, and are not captured by

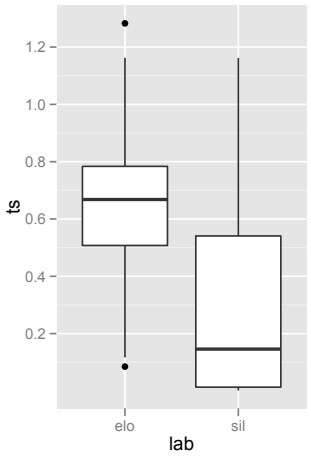

Figure 15: *Time scale factor of speech parts containing silences and prominences in clear speech.*

the global/salient silences detection. This analysis shows which segment would mainly benefit from an intrinsic time-scale modification based on the rhythmogram (cf. Section 6.2).

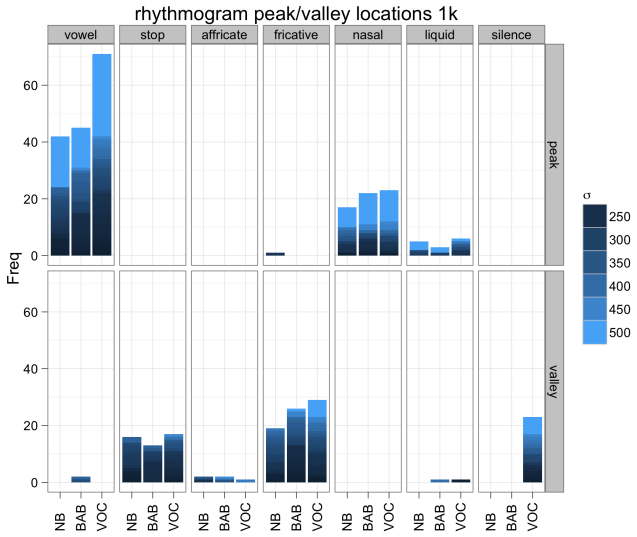

Figure 16: *Prominences and silences locations in sound classes in one non-barrier condition (NB) and two clear speech eliciting conditions (BAB and VOC). The filter width is controlled by σ: the higher the number, the more global the prominence / silence.*

These analyses provided a first pass characterization of the rhythmic properties of clear speech over casual or less clear speech styles. Future directions for assessing the specific places which differ between the speech styles could include acoustic analyses in the vicinity of detected prominences/silences, and global rhythmogram pattern analyses.

6 Proposed Time-scale modifications

Analysis of the P8-Harvard corpus showed a consistency of the speakers to elongate content words and add more pauses in the barrier cases. These adaptations result to a lower speaking rate of the speech signal. Therefore, two time-scaling modification techniques were developed in order to mimic the adaptations that speakers A1 and A2 make when they elicit clear speech in the barrier conditions. These time-scaling

techniques elongate the non-stationary parts of speech in order not to introduce artifacts to the speech signal and insert pauses to the signal. The first time-scaling technique is based on the Rhythmogram of speech and the second on the Perceptual Speech Quality Measure (ITU Standard REC-BS.1387-1-2001).

6.1 Perceptual Speech Quality Measure based Time-Scale Modifications

The Perceptual-Speech-Quality measure (PSQ) is used to elongate the stationary parts of casual speech and to define where to insert pauses to the signal. The Perceptual Speech Quality measure is based on the basic version of $ITUStandardREC - BS.1387 - 1 - 2001$, a method for objective measurements of perceived speech quality. It estimates features such as loudness and modulations in specific bands, in order to describe the input signal with perceptual attributes.

Two metrics of the PSQ model are used to detect the stationary parts of speech, where time-scaling can be applied: the perceived loudness of the signal in low bands and the loudness modulations in high bands. Analytically, PSQ estimates the perceived loudness on the low frequency bands (0-300Hz) of the signal, where unvoiced speech is less likely to be present. However, some voiced stop consonants, e.g. /d/, have high energy in low frequency bands. Time-scaling voiced stop consonants would cause distortion. Therefore, the loudness metric is not sufficient to decide which parts should be elongated. Then, another metric is introduced, namely the loudness modulations of high frequency bands (around 4000Hz). The loudness modulations in high frequency bands are strongly correlated with the non-stationarity of the signal and are able to detect voiced stop consonants. Therefore, the combination of the two metrics is proposed, called the Elongation Index (EI), defined as:

$$EI = \begin{cases} L-M, & L-M < threshold \\ -1, & L-M > threshold \end{cases} \quad (2)$$

where L is the average perceived loudness in low bands and M the loudness modulations in high frequency bands. EI is calculated for each frame of the signal. If EI exceeds a threshold then the frame is allowed to be elongated. The lower the threshold, the more likely is to capture non-stationary parts. EI does not depend on the energy of the signal and its threshold is defined between [1.3 - 1.4].

An example of how EI is calculated for a speech signal is shown on Figure 17. The speech signal depicted on Figure 17 corresponds to the phrase "made a sign." The loudness in low frequency bands, as calculated by PSQ, is depicted with a green curve whereas the modulations in high frequency bands are in red. Voiced phonemes like /ey/ and /e/ have high loudness on low bands and low modulations on high bands. In these cases, EI is above the threshold and these phonemes are allowed to be elongated (Figure 17b). Phoneme /d/, as a voiced consonant, has high loudness in low bands as well as high modulations in high bands, so it is not allowed to be elongated. For consonants like /s/ the loudness metric is lower than the modulation metric. Therefore, they will not be elongated either. Notice that the value of EI in Figure 17b is not important, rather, the sign of EI is taken into account.

Each frame that can be elongated is now the center of a window with duration 20msec. Then, for this frame, a time-scale factor of 120% is created. The time-scale factor for the total sentence duration consists of the time-scale factors only for the frames that will be elongated. The time scale factors for these voiced frames are given as input to WSOLA [5], which then time-scales the signal.

6.1.1 Pause Insertion

Pause insertion is also implemented using the PSQ model. The perceived loudness of the speech signal in the whole band is estimated. Then, loudness is normalized by the maximum loudness of the signal and on this loudness curve, the valleys that are 30% lower than the maximum loudness of the signal are detected. The valleys with very low values, less than 10% of the normalized loudness of the signal, can be considered silences. On the other hand, it is observed that the valleys that fall within the loudness interval (10%, 20%] usually are in the middle of word boundaries and are appropriate for inserting pauses without distorting the signal. The pauses that result from these valleys are called aggressive pauses to distinguish them from the pauses derived from the valleys with very low values of loudness (non-aggressive). The PSQ algorithm adds both non-aggressive and aggressive pauses to the signal. The reason for the distinction between aggressive and non-aggressive pauses is that the algorithm uses different techniques to do the insertion. First, the non-aggressive pauses are inserted on the signal by adding a constant pause of 90 ms duration. Then, in order to insert the non-aggressive pauses on the location where the signal has higher loudness, a pre-processing of the signal before and after the location of the valley must be made. The pre-processing involves a time-scaling of the signal around the location where the gap will be inserted, if this is allowed by the stationarity restriction. Then, after scaling, a

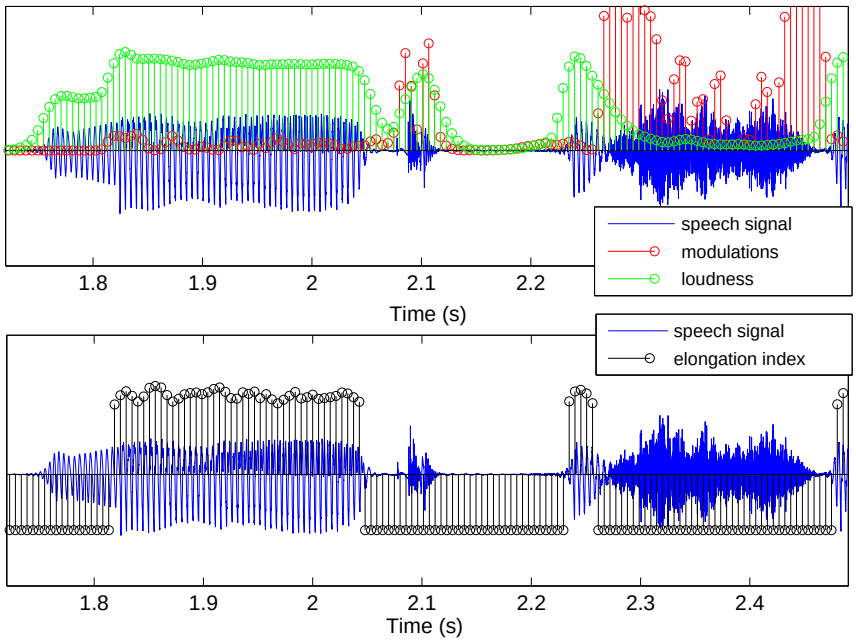

Figure 17: *Detection of non-stationary parts using PSQ model on the sentence "made a s(ign)" a) Loudness in low frequency bands and modulations in high frequency bands (top) b) Elongation index (bottom)*

hamming window is applied on the center of the valley so that the transition from speech to silence will be more smooth.

6.2 Rhythmogram-Inspired Time-Scaling and Pause Insertion

The speech rhythmogram has been proposed by Todd and Brown [6, 7] in order to model prosody perception. In order to generate the rhythmogram, the speech signal is first pre-processed to simulate the processing of the auditory periphery. In particular, the speech signal is first rectified and then raised to the one-third power. This processing approximates the loudness of the speech. Then, for the rhythmogram generation, multi-scale filtering is carried out by convolving the pre-processed speech with Gaussian windows of varying length in time. The peaks or prominences of the levels (corresponding to different Gaussian window lengths) are then linked in order to capture and visualize the overall rhythm of the speech. The following describes how this rhytmogram analysis of speech is used to inspire a time-scaling and pause insertion algorithm for speech.

Given the previously described observations on the differences between clear and casual speech, a PSQ-based algorithm for time-scaling and pause insertion was proposed. The rhythmogram provides a simple way to approximate the PSQ-based algorithm, in that it also elongates louder parts of speech while largely avoiding non-stationarities. Moreover, valleys in the rhythmogram level curves are used to detect where to insert pauses. Simplifying the processing by removing the need for calculation of the PSQ measure then frees up the rhythmogram-based approach (in terms of complexity) to provide additional pause enhancement using a WSOLA-based interpolation scheme. Explicitly, the rhthymogram-based time-scaling and pause insertion can be broken down and described in the following steps.

6.2.1 Pause Detection and Insertion

First, in order to approximate loudness, step 1 is to rectify the speech signal and raise it to the one-third power, mimicking processing in the auditory peripher. A "gross" Gaussian window (50msec length) is then convolved with the processed signal. The valleys in the resulting envelope then represent the longest pauses, or silences in the signal. This envelope is normalized so that its maximum value is one. The

location of the deepest valleys, defined as those more than 40% lower than the envelope maximum, are then used to indicate where zeros are inserted in the signal (see Figure 18). The length of the insertions are inversely proportional to the envelope valley depth, with the lowest valley being elongated the most (80msec).

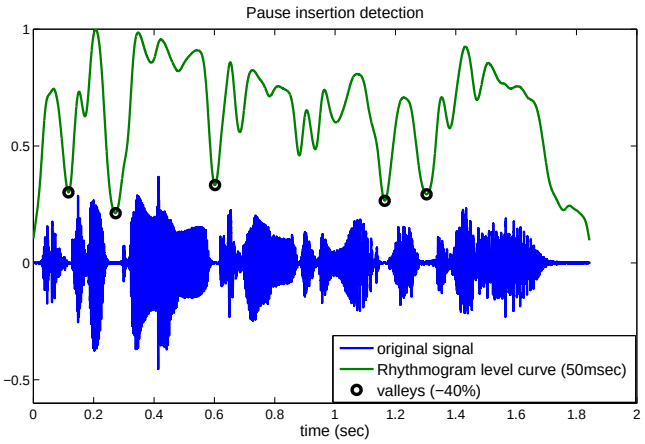

Figure 18: *Rhythmogram-based pause detection.*

6.2.2 Time-Scaling

A similar process to that used for the pause detection and insertion can also be used for time-scaling. In particular, the speech signal (with inserted pauses) is processed and the envelope (rhythmogram level curve) extracted in the same way. However, in this case, the time-scaling seeks to elongate prominences (peaks) and also silences (valleys) in the envelope. Like for the PSQ-based algorithm and uniform scaling described in this work, WSOLA is used for the time-scaling. Consequently, the normalized envelope from the rhytmogram level will determine the time-scaling factors that are input to WSOLA. In particular, the mean of the normalized envelope is first removed. The result is then rectified, so that the valleys become peaks. With this rectification, the parts of the envelope corresponding to transitions in speech (e.g., non-stationarities) lie near zero, as they have energy lower than the loudest parts of speech, yet higher than silences. The rectified envelope is then scaled by a maximum scaling factor, so that the time-scaling will not involve a factor larger than this amount. The scaling factors input to WSOLA are then one plus the scaled, rectified envelope. So, the non-stationary parts of speech will have a scaling factor near one and the rest of the speech will have a scaling factor above one, to elongate the signal, but below the specified maximum scaling factor (the maximum scaling currently limits the time-scaling factor to 2). Figure 19 shows an example of the input to WSOLA based on the rhythmogram-inspired time-scaling.

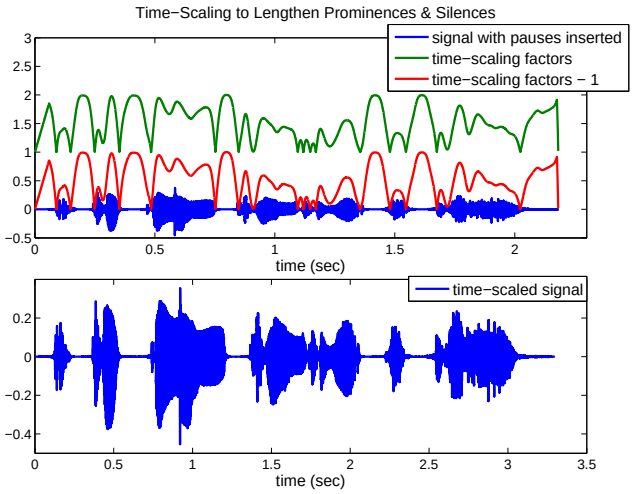

Figure 19: *Rhythmogram-based time-scaling.*

7 The GUI: XPlic8

As one of outputs of P8, XPlic8 is a MATLAB-based graphic tool for carrying out a series of analyses on single or batch of signals. It comprises of a set of functions for acoustic-phonetic measurement of speech, as typically used in speech science and phonetics research.

XPlic8 is able to perform seven acoustic-phonetic analyses and two visualization methods on sentence, word and phoneme levels. These are listed below:

- Analysis
  - Duration (s)
  - F0 median (Hz)
  - F0 range (Hz)
  - LTAS energy between a specified frequency range db SIL
  - Spectral tilt (dB/oct)
  - Vowel space (F1 (Hz), F2 (Hz))
  - Centre of gravity (Hz)
- Visualization
  - Source features analysis [8], [9]
    - LPC Spectrum
    - Harmonic-to-noise (HNR) ratio plot
    - Average glottal flow waveform
  - Vowel space plots
    - Plot of F1/F2 of tense vs. lax vowels
    - Plot of mean F1/F2 for all vowels
    - Plot of centre of gravity for /i/-/ɒ/-/ɔ/

Note that some analyses can only be performed on certain levels and also rely on the existence of corresponding annotation files for the signals. The detailed results from the analyses can be exported in plain text format that can be used as direct input for statistical applications such as SPSS or R for further analysis.

7.1 Analysis algorithms

The analysis algorithms developed during this project and incorporated in the GUI XPlic8 are described below.

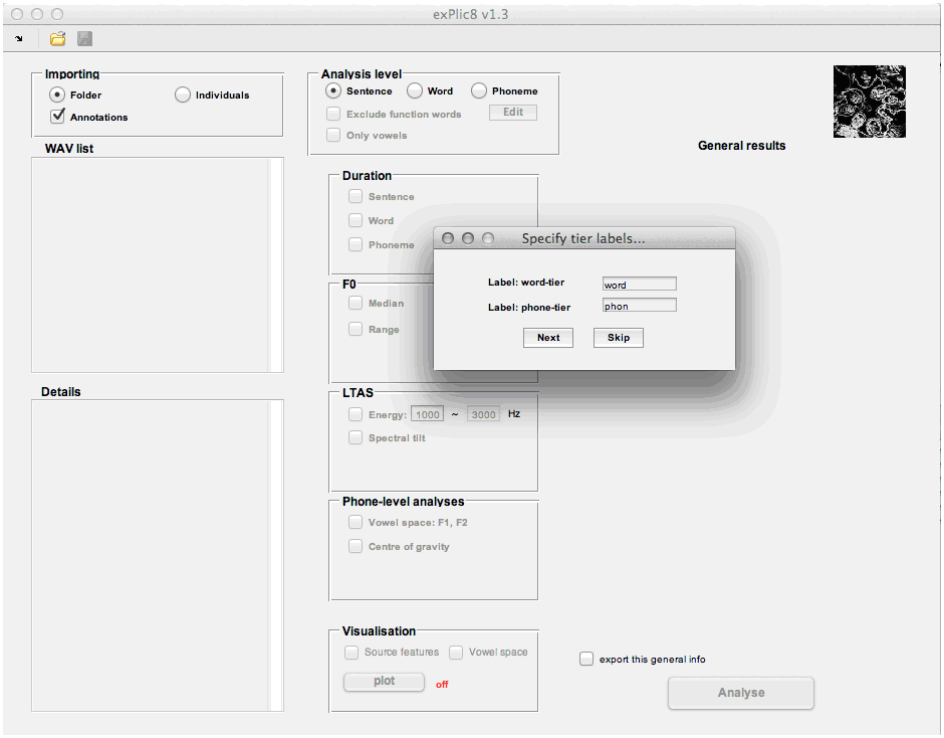

Figure 20: *The GUI XPlic8*

7.1.1 F0 estimation

A rough F0 trajectory prediction is performed prior to actual pitch detection. This is done in two stages: The first stage is to high-pass filter the speech signal in order to remove possible low frequency noise, followed by defining the rough F0 range. This is performed by using simple inverse filtering of the speech signal in order to remove most of the formants and then integrating the signal in order to get a signal close to glottal flow. This is done frame-wise with a 40-ms window. The rough fundamental period is estimated by evaluating the autocorrelation sequence of the signal and then finding the maximum peak that corresponds F0 between 50 and 500 Hz. Those frames with low energy or high zero-crossing rate (ZCR) are classified as unvoiced. F0 range is defined as:

$$F0_{min} = median(f_0)^{\frac{1.2}{5}} \tag{3}$$

$$F0_{max} = 2.2median(f_0) \tag{4}$$

The actual pitch detection takes place after the initial estimation of the F0 range. The analysis window size is adjusted to the estimated F0 range so that it is twice the lowest fundamental period ($2/F0_{min}$). The glottal inverse filtering method used in F0 estimation is iterative adaptive inverse filtering (IAIF) which estimates the glottal flow signal of the frame using linear prediction such that the fundamental period from the vibratory glottal flow waveform can be estimated. The fundamental period is estimated again finding the maximum peak of the autocorrelation sequence.

For post-processing, two highest peaks are saved: First, the post-processing involves forming a continuous trajectory from the two trajectories. This is based on the relative jump of the trajectories compared to a local F0 median. Second, 5-point median filtering is applied to smooth out outliers. Third, the unvoiced parts are set to zero based on the energy, ZCR, autocorrelation peak value, and gradient index. Fourth, the F0 trajectory is filtered with a 3-point medial filter. Finally, the median F0 is defined as the median of the non-zero values of the trajectory. The $F0_{min}$ and $F0_{max}$ are defined as the minimum and maximum non-zero F0 values of the trajectory.

7.1.2 LTAS energy in specified frequency ranges

The energy is computed as the intensity in SIL (sound intensity level) dB on the specified frequency range. The input sample is windowed with a 5-ms rectangular window without overlap and a 1024-length Fourier transform (using the FFT function) is computed for each frame. To obtain the normalized intensity for each frame, the energy in the specified frequency range is normalized by the length of the FFT, the length of the window (in samples) and the sampling frequency. Finally, the normalized intensities of all the frames are summed and the corresponding decibel value is computed by using the reference value $I_0 = 10e^{12}$.

7.1.3 Spectral tilt

The average spectral tilt is computed by fitting a regression line to 1/3-octave band energies of the LTAS (long-term average spectrum) in logarithmic scale. The LTAS is computed in 5-ms frames without overlap. For each frame, a 2048-length Fourier transform (with the fft function) is computed and the LTAS is obtained as the mean of the absolute values of the Fourier transforms over all frames. The average energy in the LTAS for each third-octave band is computed and normalized with the width of the band. These values are then transformed to logarithmic scale and a first-degree polynomial fit is estimated (using function polyfit). The average spectral tilt (in dB/octave) is three times the value of the first coefficient of the polynomial.

7.1.4 Vowel space (F1, F2)

The formant extraction tool returns the formant values in the middle point of the selected segment. It uses Praat [2] to extract the formant values for each consecutive frame in the selected speech segment and the cheapest paths through those values. Then, the values related to the centre of the time interval are chosen. This function returns formant info for every selected phone and this data is also used to plot the vowel space. Most of the analysis options are already optimised and cannot be changed: Time step = 0.01 s, Maximum formant number = 7, Number of paths to tracks = 5, Formant search range ceiling = 6500 Hz, Pre-emphasis filter lower limit = 50 Hz, Duration of the analysis window (0.025 s). For a detailed description of these parameters, please refer to the online Praat manual (Sound to Formant (Burg) and Formant Track)

7.1.4.1 Formant extraction The sound is re-sampled (Sound: Resample) to a frequency of twice the value of maximum formant and a pre-emphasis filter is also applied (Sound: Pre-emphasize (in-line)). For each analysis window, a Gaussian-like window is applied and the LPC coefficients are as per the algorithm by Burg, as (Childers, D.G., 1978) and (Press, W.H. et al., 1992). The number of "poles" in this algorithm is set as twice the maximum number of formants. The algorithm finds the best peaks in the selected range of frequency (between 0 Hz and the maximum formant value). Then, all formants below 50 Hz and above the ceiling minus 50 Hz are removed because very low frequency (near 0 Hz) and very high frequency (near the maximum) peaks cannot usually be associated with the vocal tract resonances and they are likely to be artifacts of the LPC algorithm.

7.1.4.2 Formant tracking After the formant candidate extraction, a tracking on these values is performed in order to rearrange the peaks to obtain the best formant tracks. This command uses a Viterbi algorithm with multiple planes and chooses the cheapest path through all the previously selected peaks (Formant Track). The cost function for one track (e.g. 2) with proposed values $F_{2,i}$ ($i = 1...N$, where N is the number of frames) is:

$$\begin{aligned} CostFunction = & \sum_{i=1}^{N} frequencyCost \frac{|F_{2,i} - referenceF_2|}{1000} \\ & + \sum_{i=1}^{N} bandWidthCost \frac{B_{2,i}}{F_{2,i}} + \\ & + \sum_{i=1}^{N-1} transitionCost |log_2 \frac{F_{2,i}}{F_{2,i+1}}| \end{aligned} \tag{5}$$

where frequencyCost, bandWidthCost, transitionCost, and referenceF2 values are fixed and all set to 1. Analogous formulas compute the cost of other tracks. The procedure will assign those candidates that minimize the sum of all-track costs.

7.1.5 Centre of gravity (CoG)

The Centre of Gravity is a measure of the spectrum energy distribution. The average spectrum on the speech segment is computed. It uses the Praat software [2]. Given the complex spectrum, S(f), f is the frequency, the CoG is computed by

$$\int_0^\infty f|S(f)|^p df \tag{6}$$

divided by the"energy"

$$\int_0^\infty |S(f)|^p df \tag{7}$$

The value of p is chosen to be 2. For further details please refer to the online Praat manual (Spectrum: Get the centre of gravity).

7.1.6 Source features

For details of F0 prediction refer to F0 estimation. The polarity is estimated by comparing the positive and negative energy of the glottal flow derivative signal. If the negative energy is greater, the speech signal most likely has positive polarity (and vice versa). After F0 and polarity detection, a suitable window size is selected for estimating the parameters ($3/F0_min$). Iterative adaptive inverse filtering (IAIF) is applied to the speech signal to separate the vocal tract transfer function and the voice source signal. Then, various parameters are extracted, such as:

- F0 and voiced/unvoiced decision [3]
- LPC and FFT spectra of voiced speech
- LPC and FFT spectra of unvoiced speech
- LPC and FFT spectra of vocal tract
- LPC and FFT spectral of voice source
- Speech energy
- Harmonic-to-noise ratio (HNR)
- H1-H2 value of the glottal flow signal
- Normalized amplitude quotient (NAQ)
- Individual glottal flow pulses and their average

The harmonic-to-noise ratio is evaluated by peak picking of the harmonics and then comparing the magnitude difference between the harmonics and the inter-harmonic valleys. These values are averaged to five equivalent rectangular bandwidth (ERB) bands. Normalized amplitude quotient is evaluated for each glottal flow pulse and thus averaged to one value for each frame. Finally, all the estimated unique glottal flow pulses are interpolated to constant length and averaged to estimate the average glottal flow waveform. Parameters are post-processed with median filtering. Statistics of the parameters are evaluated with 95% confidence intervals.

[3]Only available when single WAV file is selected and the analyses are performed on sentence level.

8 Summary and Conclusions

A new speech corpus, the P8-Harvard corpus, was linguistically and meta-linguistically annotated and acoustically analyzed. The aim was to identify which acoustic-phonetic characteristics differ between clear and casual speech then to modify casual speech to sound as intelligible as clear speech.

The P8-Harvard corpus contains, for each of two speakers, 150 sentences produced in a casual and two clear speaking styles. It is provided with word- and phoneme-level annotation, as well as pause annotations. Communication was harder in the communication barrier conditions, as shown by a decrease in keywords correctly transmitted, with the VOC condition being harder for both speakers. Acoustic-phonetic analysis revealed that sentence duration increased significantly in the barrier conditions, but that this was mainly due to an increase in pause duration, with a greater number of inter-word pauses seen in the more difficult (VOC) condition. Speakers also altered their F0 in the barrier conditions (higher F0 median in both conditions, broader F0 range in BAB condition only), and increased speech intensity (mid-frequency region), especially in the BAB condition. Speaker A1 hyperarticulated her vowels in the barrier conditions but no significant vowel space expansion was seen in male speaker A2. Evidence of communication-barrier specific strategies was seen. There was also evidence of differences in enhancement strategies across the two speakers for most dimensions.

Analysis of the P8-Harvard corpus showed a consistency of the speakers to elongate content words and add more pauses in the barrier cases. These adaptations result to a lower speaking rate of the speech signal. Therefore, two time-scaling modification techniques were developed in order to mimic the adaptations that speakers A1 and A2 make when they elicit clear speech in the barrier conditions. These time-scaling techniques elongate the non-stationary parts of speech in order not to introduce artifacts to the speech signal and insert pauses to the signal. The first time-scaling technique is based on the Rhythmogram of speech and the second on the Perceptual Speech Quality Measure (ITU Standard REC-BS.1387-1-2001). a set of evaluation experiments was prepared to evaluate the different modifications. The evaluation must be done by native listeners therefore no listening tests were conducted during the Enterface2012.

Finally, a significant outcome of P8 is XPlic8, a MATLAB-based graphic tool for carrying out a series of analyses on speech databases. It comprises of a set of functions for acoustic-phonetic measurement of speech, as typically used in speech science and phonetics research.

References

[1] V. Hazan and R. Baker. Does reading clearly produce the same acoustic-phonetic modifications as spontaneous speech in a clear speaking style? *DiSS-LPSS*, pages 7–10, 2010.

[2] Boersma P. and Weenink D. Praat: doing phonetics by computer [computer program]. 2012.

[3] IEEE. IEEE recommended practice for speech quality measurements. Technical Report No. 297, 1969.

[4] Neil P McAngus Todd and Guy J Brown. Visualization of Rhythm, Time and Metre. *Artificial Intelligence Review*, 10:253–273, 1996.

[5] M. Demol, K. Struyve, W. Verhelst, H. Paulussen, P. Desmet, and P. Verhoeve Author. Efficient non-uniform time-scaling of speech with WSOLA for call applications. *Proceedings of InSTIL ICALL2004 - NLP and Speech Technologies in Advanced Language Learning Systems*, 2004.

[6] N.P. Todd and G. Brown. A computational model of prosody perception. *ICLSP*, 10:127–130, 1994.

[7] N.P. Todd and G. Brown. Visualization of rhythm time and meter. *Artificial Intelligence Review*, 10:91–113, 1996.

[8] T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, M. Vainio, and P. Alku. HMM-based speech synthesis utilizing glottal inverse filtering. In *IEEE Transactions on Audio, Speech, and Language Processing*, volume 19, pages 153–165, 2011.

[9] T. Raitio, A. Suni, H. Pulakka, M. Vainio, and P. Alku. Utilizing glottal source pulse library for generating improved excitation signal for HMM - based speech synthesis. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4564–4567, 2011.