SIMPLE AND ARTEFACT-FREE SPECTRAL MODIFICATIONS FOR ENHANCING THE INTELLIGIBILITY OF CASUAL SPEECH

Maria Koutsogiannaki, Yannis Stylianou

Multimedia Informatics Lab, Computer Science Department, University of Crete, Greece
mkoutsog@csd.uoc.gr, yannis@csd.uoc.gr

ABSTRACT

In this paper, the problem of modifying casual speech to reach the intelligibility level of clear speech is addressed. Unlike other studies, in this work modifications on casual speech both consider intelligibility and speech quality. To achieve this, the authors focus on human-like modifications inspired by clear speech. An acoustic analysis performed on clear and casual speech reveals energy differences on specific frequency bands between the two speaking styles. Then, a simple method is used to boost these frequency regions on casual speech. The proposed method, called mix-filtering, uses a multi-band filtering scheme to isolate the information of these frequency bands and then, add this information to the original signal. Our method is compared in terms of intelligibility and quality with unmodified casual speech and with a highly intelligible spectral modification technique, namely the Spectral Shaping and Dynamic Range Compression (SSDRC). Two different objective measures that are highly correlated with subjective intelligibility scores are used for estimating the intelligibility, whereas for evaluating the quality, preference listening tests are performed. Results show that the mix-filtering technique increases the intelligibility of casual speech while maintains its quality. On the other hand, while SSDRC outperforms on intelligibility, it degrades significantly the quality of casual speech.

Index Terms— Clear speech, Casual speech, Intelligibility, Speech quality, Spectral modifications

1. INTRODUCTION

Humans adopt many different speaking styles in order to overcome communication difficulties. Depending on the communication barrier, speakers produce different styles of speech (e.g. Lombard speech, shouted speech). When a speaker is in a non-noisy environment but the listener faces a communication barrier, clear speech strategies are employed. For example, the target listener could be either hearing-impaired or a non-native listener (L2). Despite many differences in speaker strategies, the most common characteristics of clear speech is hyper-articulation, with increased effort on the part of the speaker to enunciate. Conversely, plain or casual speech is the type of speech produced when there is no barrier in the communication channel.

Intelligibility differences between clear and casual speech have been investigated for various listener populations, on a great variety of speech materials and under various conditions (e.g under the presence of noise [1]). Specifically, related studies reported that the average intelligibility of clear speech is higher than that of casual speech across hearing-impaired listeners [2, 3], coclear-implant users[4] and normal-hearing listeners [5, 6, 4, 7]. Even though the degree from the benefit of clear speech varies according to the age of the subjects [3], their linguistic knowledge (e.g native listeners vs. non-native listeners [8]), the noise conditions of the experiment for the normal-hearing population [9], in general clear speech has higher intelligibility than casual speech.

Many studies have been focusing on the intelligibility enhancement of plain speech by seeking and exploiting differences between the two speaking styles [5, 10, 11]. Clear and casual speech, appear to have differences on their spectral and prosodic characteristics. Focusing on the spectral domain, one feature that is possibly associated with the intelligibility of clear speech is an energy increase above 1000Hz [12, 13]. This increase of energy compared to plain speech in similar frequency regions occurs also in other speaking styles, like on Lombard speech (naturally produced speech in noise). It has been shown that performing Lombard-like modifications on plain speech by boosting the frequency region $1-4$kHz while maintaining the overall RMS energy of the signal, has an intelligibility increase [14]. In [10] a similar approach has been used for clear speech, amplifying the energy around F2 and F3 formants on casual speech on voiced segments. Intelligibility tests for normal hearing listeners in noise ($SNR = -1.8dB$) showed that modified speech was more intelligible than unmodified casual speech and less intelligible than clear speech. In addition, other simpler spectral modifications can increase speech intelligibility. Performing high-pass filtering on speech with cut-off frequency $1.5kHz$ increases its intelligibility in noise [15].

The aforementioned studies report that spectral modifications of casual speech may be proven beneficial for its intelligibility. However, none of the studies is concerned with the quality degradations imposed to original speech. Previous work that has been done by the authors [16] has shown that spectral shaping and energy reallocation techniques (SSDRC) can increase the intelligibility of casual speech to levels higher than that of clear speech on low SNR. However, the quality of modified speech is quite degraded. The majority of the studies that examine speech intelligibility, test the speech signals in noise, masking all the artefacts that may be introduced on processed speech. For example, the SSDRC modified signals in [16] even though highly intelligible in noise (even higher than clear speech), were quite distorted and this could only be reported if heard outside noise. Even if it is preferable to test the intelligibility of speech in noise, as normal-hearing subjects can be used for evaluations, speech is not always intended in noise. On some applications it is important to preserve the quality of speech (e.g applications for hearing impaired listeners). Therefore, it is under question whether or not spectral modification techniques can increase intelligibility without affecting speech quality.

This work tries to address the problem of increasing the intelligibility of casual speech while maintaining its quality. Motivated by previous studies that achieve to increase intelligibility using spectral modifications, this study also modifies the spectral characteristics of casual signals imposing however, quality restrictions. The pro-

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

posed method, inspired by the properties of clear speech, amplifies the energy of specific frequency bands of original casual speech. The advantage of the method is its simplicity and efficiency. Firstly, unlike other techniques [10, 16] it does not require frame-based analysis and modifications (detection of voiced/unvoiced regions, formant shaping etc). On the contrary, it isolates frequency bands on casual speech by simply performing multi-band filtering and then adds back to the initial signal the filter outputs. Secondly, results show that the proposed modified scheme increases the intelligibility of casual speech while maintains its quality. The evaluation of the modified speech in terms of intelligibility is performed objectively using two different objective measures that predict intelligibility, the Glimpse Proportion (GP) [17] and the Distortion-Weighted Glimpse Proportion (DWGP) [18]. In order to quantify the intelligibility and quality advantage of our proposed method, our modified casual speech is compare not only with unmodified speech but also with SSDRC modified speech, which has similar intelligibility levels with clear speech [16]. Then, the quality of the proposed scheme is evaluated subjectively using a preference test between our modification, unmodified plain speech and SSDRC modified speech.

This paper is organized as follows: Section 2 describes the database of clear and casual speech used for analysis and modifications. Section 3 introduces our proposed method of increasing the intelligibility of plain speech while maintaining its quality. Section 4 presents the evaluations on the intelligibility and quality of modified speech compared to casual speech and SSDRC modified speech. Lastly, Section 5 concludes the paper.

2. SPEECH CORPORA

The corpora used on our analysis is the read clear and read casual speech from the LUCID database [19]. Read casual speech was produced after instructing Southern British English normophonic speakers to read meaningful and simple in syntax sentences "casually as if talking to a friend" whereas for read clear speech the instructions were to speak "clearly as if talking to someone who is hearing impaired" [19, 13]. Read clear speech shows more extreme changes on certain acoustic-phonetic characteristics than spontaneous clear speech [20, 13, 21] and appears to be more intelligible than spontaneous clear speech [22].

From the LUCID corpus 4 Male and 4 Female speakers are selected to form our dataset. The dataset contains 60 sentences per speaker and per speaking style. A preprocessing is performed on the dataset. The preprocessing involves downsampling to $16kHz$ and removal of lowpass noise from breath and lip effects, using a 5-order highpass digital elliptic filter with $80Hz$ cut-off frequency. Then, this dataset is split in two parts. The first part is used as an analysis dataset in order to extract the spectral differences between clear and casual speech (dataset A). It contains 20 sentences per speaker and per speaking style. Then, the second part (dataset B) is used as an evaluation dataset and contains 40 sentences per speaker but only for the casual speaking style. The intersection of the two datasets A and B is null.

3. METHODOLOGY

The method proposed for the intelligibility enhancement of casual speech is simple. First, we define which frequency bands are more enhanced naturally on clear speech compared to casual speech. To that purpose, the average smoothed spectral envelopes are estimated for clear and casual speech on the analysis dataset A. The analysis reveals larger differences on two frequency bands between clear and casual. The information corresponding to these frequency bands is isolated and added to the original casual signal with different weighting factors. Then, the modified signal is normalized to have the same energy as the original signal.

3.1. Analysis on clear and casual speech corpora

Clear and casual speech is analyzed in order to reveal possible differences between the two speaking styles. However, unlike other studies [10], analysis is performed on the whole signal and not only on the voiced segments, accounting for the importance of the consonants on speech intelligibility. Every clear sentence and its corresponding casual is analyzed. The analysis is done on dataset A and involves frame-by-frame estimation of the true envelope as proposed by [23, 24] for the voiced segments and spectral envelope estimation directly from the LPC analysis for unvoiced segments. The true envelope estimation is based on cepstral smoothing of the amplitude spectrum. The cepstrum order is set to 10 in order to estimate an overall energy of the frequency bands. For each spectral envelope, the DC component is set to zero. Then, the spectral envelope is normalized by its RMS to eliminate intensity differences between clear and casual speech. The averaged spectral envelopes are computed as the mean of all frames for each speaking style separately. Figure 1 shows the difference of the log average spectral envelopes of clear speech minus casual speech. Positive difference suggests that the energy of clear speech is higher than that of the casual speech. As we can see, clear speech appears to have higher energy in two frequency bands, $B1 = [2000, 4800]$ and $B2 = [5600, 8000]$.

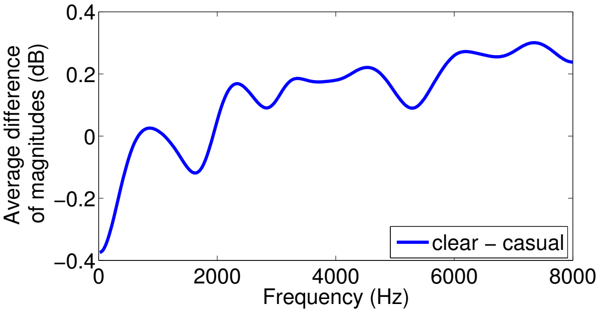


Fig. 1: Difference of the log average spectral envelopes of clear speech minus casual speech: clear speech has higher energy on two frequency bands, $B1 = [2000, 4800]$ and $B2 = [5600, 8000]$.

3.2. Mix-filtering

The above analysis reveals an increased energy on clear speech compared to casual speech on two frequency bands $B_1 = [2000, 4800]$ and $B_2 = [5600, 8000]$. The method proposed on this paper for the enhancement of the intelligibility of casual speech involves the isolation of these important frequency bands B_1 and B_2 and then the addition of their energy to the original signal with different weighting factors for each frequency band. Hopefully, this addition will boost the important frequency regions on casual speech, as it naturally happens in clear speech.

For the isolation of the frequency bands a simple method is used. Casual speech s is filtered with a 5-order bandpass digital elliptic filter with $0.1dB$ of ripple in the passband, and $60dB$ ripple in the stopband and bandpass edge frequencies $[2000, 4800]$. An IIR filter is selected for the frequency band isolation in order to have an abrupt transition from passband to stopband. No phase adjustment is performed. However, distortions due to the non-uniform group-delays are not perceptually noticeable as will be reported later on this paper by the quality evaluations. The output of the filter is signal s_1 which contains information on the B_1 frequency band. Moreover, casual speech s is filtered with a 5-order highpass digital elliptic filter with normalized passband edge frequency $f_c = 5600Hz$. The output of this filter is the signal s_2 which contains information on the frequency band B_2. Then, the original signal s and the filtered signals s_1 and s_2 are combined with different weighting factors to form the modified signal y, which is normalized to have the same RMS energy as original speech:

$$y[i] = w_0 s[i] + w_1 s_1[i] + w_2 s_2[i] \tag{1}$$

$$y_{mixF}[i] = y[i] \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N} s^2[i]}}{\sqrt{\frac{1}{N}\sum_{i=1}^{N} y^2[i]}} \tag{2}$$

where, y_{mixF} is the proposed modified signal, N is the number of samples of the casual signal s and y, and w_0, w_1, w_2 are the weighting factors of the signals s, s_1 and s_2, respectively.

The selection of the proper combination of the weights is important both for intelligibility and quality. In [15] it has been shown that high pass filtering speech above 1.5kHz increases its intelligibility in noise. However, the absence of information on lower frequency bands can degrade the quality of speech. Therefore, this information is contained on the modified speech y_{mixF} by choosing to keep the original speech signal weighted by w_0. Then, the selection of the other two weights is inspired by clear speech properties. Specifically, focusing on the energy differences between clear and casual speech on Figure 1, it can be observed that the energy in B_2 frequency band is greater than that of B_1 on clear speech than on casual speech. Possibly, this happens because the energy of consonants is higher in clear than in casual speech. Therefore, we choose $w_2 > w_1$ to account for the slight higher energy difference of B_2 frequency band compared to B_1 between the two speaking styles.

Summarizing the above, the set of the possible weighting combinations can be described by the following equations:

$$w_0 = 1 - \sum_{i=1}^{2} w_i \tag{3}$$

$$w_2 > w_1 \tag{4}$$

$$w_i \neq 0, i = 0, 1, 2 \tag{5}$$

In order to select one proper weight combination $\{w_0, w_1, w_2\}$ we consider w_0 as a dependent variable. Then, the two variables w_1, w_2 can vary between (0, 1) respecting the restrictions described by equations (3), (4) and (5). As we are interested on enhancing the intelligibility of casual speech, the proper values $\{w_0, w_1, w_2\}$ are those that maximize the intelligibility score of modified speech compared to unmodified speech. To define these values, the casual speech of dataset A is used as a training dataset. Specifically, the casual signals of dataset A are modified using different weight combinations that satisfy the above equations. The intelligibility of the modified sentences using the mix-filtering approach (mixF) and the unmodified casual sentences is evaluated objectively on the presence of low SNR ($SNR = -10dB$) Speech Shaped Noise (SSN). The best combination of weights is the one that maximizes the objective intelligibility difference of the modified speech minus the unmodified speech.

The objective metric used to predict intelligibility is the Glimpse Proportion (GP) [17, 25]. The Glimpse measure comes from the Glimpse model for auditory processing. As an intelligibility predictor, the model is based on the assumption that in a noisy environment humans listen to the glimpses of speech that are less masked. Therefore, the GP measure is the proportion of spectral-temporal regions where speech is more energetic than the noise.

Figure 2 shows for various weight combinations the difference between the intelligibility score of mixF speech minus casual speech given by GP. Note, that w_0 is not present as it is assumed from equation (3) to be the dependent variable. The optimal weight combination that maximizes this difference is {0.1, 0.4, 0.5}. The difference between the average smoothed spectral envelopes of the modified speech mixF that derives from this combination and the casual speech is depicted on Figure 3. The important frequency bands are boosted "stealing" from the lower frequency bands.

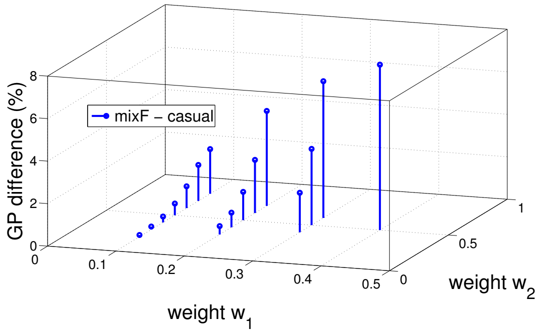


Fig. 2: % difference of GP scores between modified mix-filtered speech (mixF) minus unmodified casual speech. MixF is derived using various weights combinations that verify equations (3), (4), (5). The maximum difference is 7.78% and corresponds to $\{w_0, w_1, w_2\} = \{0.1, 0.4, 0.5\}$.

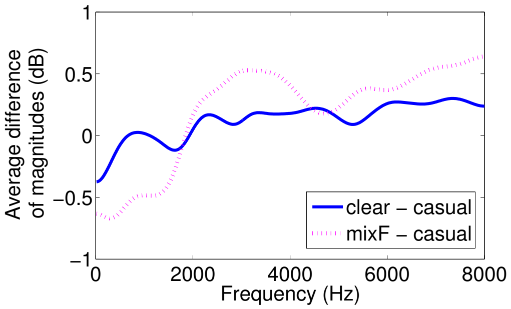


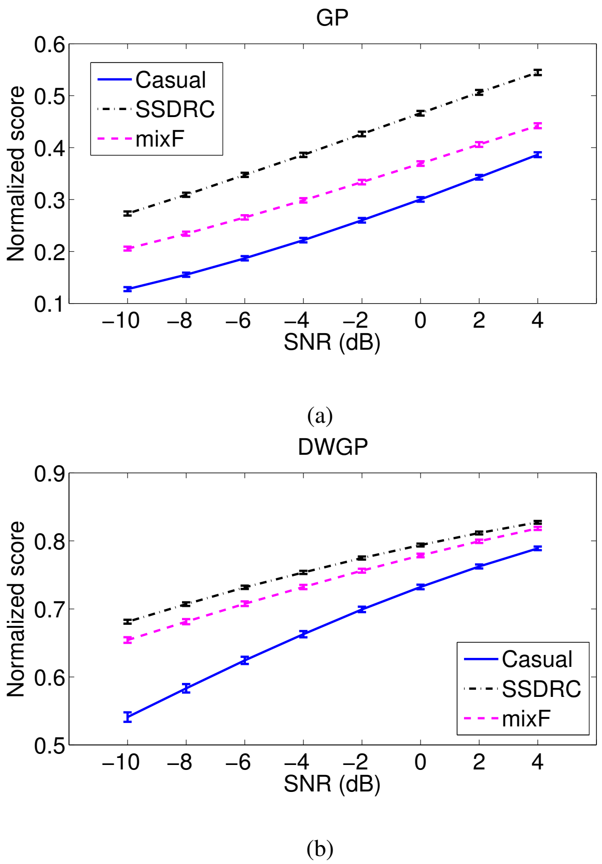
Fig. 3: Difference of the log average spectral envelopes.

4. EVALUATIONS

The modified casual speech derived from the mix-filtering approach is compared in terms of intelligibility and quality with unmodified casual speech and with SSDRC modified casual speech. The evaluation of mixF in terms of intelligibility is done using two different objective measures, the GP measure described above [17, 25] and Distortion-Weighted Glimpse Proportion (DWGP) [18]. DWGP has been shown to have a better correlation with subjective intelligibility evaluations than GP [18]. The DWGP measure computes the correlation between frequency bands of clean speech and speech in noise, weighting these correlations according to the importance of each frequency band. The prediction of intelligibility is estimated by the correlation which gives a measure of how much noise affects the signal. Then, for the evaluation of the quality of speech a preference test is made between three different speech signals.

4.1. Objective evaluations on intelligibility

On the testing dataset B, GP and DWGP scores are extracted for the three categories of speech, casual speech, mixF and SSDRC modified speech. Speech Shaped Noise (SSN) of various SNR levels is used for evaluating objectively the intelligibility of each category in noise. Figures 4a and 4b depict the objective scores predicted by GP and DWGP respectively for SNR levels varying from -10 to 4 dB. GP reports that the SSDRC outperforms in terms of intelligibility while our proposed scheme increases the intelligibility of casual speech by 8% on low SNR. On the other hand, DWGP predicts that the intelligibility advantage of our proposed method is more than 10% on casual speech on low SNR $(-10dB)$, approaching the intelligibility scores of SSDRC. Overall both objective scores predict an intelligibility increase of our proposed scheme for every SNR, with DWGP reporting intelligibility levels of mixF close to those of SSDRC.



(a)

(b)

Fig. 4: Objective scores for predicting intelligibility of each speech category in speech-shaped noise: mean values and 95% confidence intervals.

4.2. Subjective evaluations on quality

For evaluating the quality of our method, mixF, casual and SSDRC are compared in quality using preference listening tests that have been conducted without the presence of noise. 10 random distinct sentences from dataset B were presented to 18 listeners. Each sentence was modified by SSDRC and mixF and was heard 6 times, two times for each pair {casual-mixF, mixF-SSDRC, SSDRC-casual}. Listeners had to select from -3 to 3 the degree of preference between those pairs in terms of quality with 0 corresponding to the same quality and 3 (-3) to the much better (worse) quality of the one signal compared to the other. Despite the fact that the energy of the signal was the same for the three categories, the loudnesses was higher for SSDRC and mixF. Therefore, all signals were normalized in loudness using ACTIVLEV (ITU-T P.56).

Figure 5 summarizes the scores of preference of each category against the two others. Confidence intervals are also provided. As we can see, casual and mixF appear to have similar scores of preference whereas SSDRC gives negative quality scores against the other two categories, casual and mixF. The proposed mix-filtering approach preserves the quality of casual speech.

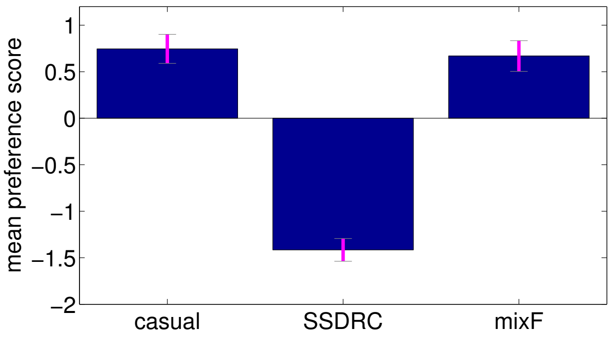


Fig. 5: Subjective quality evaluation: mean values and 95% confidence intervals of the preference scores of each category against the two others.

5. CONCLUSIONS AND FUTURE WORK

We propose a simple method for increasing the intelligibility of casual speech while preserving its quality. Our method, called mix-filtering, applies a multi-band filtering on casual speech, isolating information from important frequency bands indicated by clear speech. Then, the filtered signal is added to the original signal boosting the energy of these frequency bands. Therefore, it does not require frame-by-frame modifications and does not introduce processing artefacts. Objective evaluations that predict the intelligibility of speech in noise show an intelligibility increase compared with unmodified casual speech. As the proposed method is less intrusive, the intelligibility benefit is less compared to SSDRC. However, unlike SSDRC, the mix-filtering approach does not degrade the speech quality, as reported by subjective quality tests. A possible intelligibility increase from the combination of our proposed method with other spectral and/or temporal modifications is to be explored in the near future.

Acknowledgments

The authors would like to thank Yan Tang and Martin Cooke for providing the objective intelligibility score DWGP, Gilles Degottex for fruitful discussions and George Kafentzis for his support on the subjective evaluations.

6. REFERENCES

[1] K.L. Payton, R.M. Uchanski, and L. D. Braida. Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing. *J. Acoust. Soc. Amer.*, 95(3):1581–92, 1994.

[2] M.A. Picheny, N.I. Durlach, and L.D. Braida. Speaking clearly for the hard of hearing II: acoustic characteristics of clear and conversational speech. *J. of Speech and Hearing Research*, 29:434–446, 1986.

[3] S.H. Ferguson and D. Kewley-Port. Vowel intelligibility in clear and conversational speech for normal-hearing and hearing-impaired listeners. *J. Acoust. Soc. Amer.*, 112:259–271, 2002.

[4] S. Liu, E.D. Rio, A.R. Bradlow, and F.G. Zeng. Clear speech perception in acoustic and electric hearing. *J. Acoust. Soc. Amer.*, 116(4):2374–2383, 2004.

[5] R.M. Uchanski, S.S. Choi, L.D. Braida, C.M. Reed, and N.I. Durlach. Speaking clearly for the hard of hearing IV: further studies of the role of speaking rate. *J. of Speech and Hearing*, 39:494–509, 1996.

[6] J. Krause and L. Braida. Investigating alternative forms of clear speech: The effects of speaking rate and speaking mode on intelligibility. *J. Acoust. Soc. Amer.*, 112:2165–72, 2002.

[7] S.H. Ferguson. Talker differences in clear and conversatoinal speech: Vowel intelligibility for normal-hearing listeners. *J. Acoust. Soc. Amer.*, 116(4):2365–2373, 2004.

[8] A.R. Bradlow and T. Bent. The clear speech effect for non-native listeners. *J. Acoust. Soc. Amer.*, 112(1):272–284, 2002.

[9] R. Smiljanic and A. Bradlow. Speaking and hearing clearly: Talker and listener factors in speaking style changes. *Language and Linguistic Compass*, 3(1):236–264, 2009.

[10] J.C. Krause and L.D. Braida. Evaluating the role of spectral and envelope characteristics in the intelligibility advantage of clear speech. *J. Acoust. Soc. Amer.*, 125(5):3346–3357, 2009.

[11] S.H. Mohammadi, A. Kain, and J.Santen. Making conversational vowels more clear. *Interspeech, Portland*, 2013.

[12] J. Krause and L. Braida. Acoustic properties of naturally produced clear speech at normal speaking rates. *J. Acoust. Soc. Amer.*, 115:362–378, 2004.

[13] 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.

[14] E. Godoy and Y. Stylianou. Unsupervised acoustic analyses of normal and lombard speech, with spectral envelope transformation to improve intelligibility. *Interspeech*, 2012.

[15] R.J. Niederjohn and J.H.Grotelueschen. The enhancement of speech intelligibility in high noise levels by high-pass filltering followed by rapid amplitude compression. *IEEE Trans. Acoust. Speech Signal Process*, 24(4):277–282, 1976.

[16] M. Koutsogiannaki, M. Pettinato, C. Mayo, V.Kandia, and Y. Stylianou. Can modified casual speech reach the intelligibility of clear speech? *Interspeech, Portland Oregon, USA*, 2012.

[17] M.Cooke. A glimpsing model of speech perception in noise. *J. Acoust. Soc. Amer.*, 119:1562–1573, 2006.

[18] Y. Tang and M. Cooke. Personal communication.

[19] R. Baker and V. Hazan. Lucid: a corpus of spontaneous and read clear speech in british english. *DiSS-LPSS*, pages 3–6, Tokio, 2010.

[20] G. Laan. The contribution of intonation, segmental durations, and spectral features to the perception of a spontaneous and a read speaking style. *Speech Comm.*, 22(1):43–965, 1997.

[21] V. Hazan and R. Baker. Acoustic - phonetic characteristics of speech produced with communicative intent to counter adverse listening conditions. *J. Acoust. Soc. Amer.*, 130(4):2139–52, 2011.

[22] C. Smith. Differences between read and spontaneous speech of deaf children. *J. Acoust. Soc. Amer.*, 72(4):1304–06, 1982.

[23] S. Imai. Cepstral analysis synthesis on the mel frequency scale. *ICASSP*, 8:93–96, 1983.

[24] A. Roebel, F. Villavicencio, and X. Rodext. On cepstral and all-pole based spectral envelope modeling with unknown model order. *Pattern Recognition Letters*, 28(11), 2007.

[25] Y. Tang and M. Cooke. Subjective and objective evaluation of speech intelligibility enhancement under constant energy and duration constraints. *Florence, Italy*, pages 345–348, 2011.