# Intelligibility Enhancement of Casual Speech for Reverberant Environments inspired by Clear Speech Properties

*Maria Koutsogiannaki*[1], *Petko N. Petkov*[2], *Yannis Stylianou*[1,2]

[1] Multimedia Informatics Laboratory, CSD, University of Crete, Greece
[2] Cambridge Research Laboratory, Toshiba Research Europe Ltd.

mkoutsog@csd.uoc.gr, petko.petkov@crl.toshiba.co.uk, yannis@csd.uoc.gr

## Abstract

Clear speech has been shown to have an intelligibility advantage over casual speech in noisy and reverberant environments. This work validates spectral and time domain modifications to increase the intelligibility of casual speech in reverberant environments by compensating particular differences between the two speaking styles. To compensate spectral differences, a frequency-domain filtering approach is applied to casual speech. In time domain, two techniques for time-scaling casual speech are explored: (1) uniform time-scaling and (2) pause insertion and phoneme elongation based on loudness and modulation criteria. The effect of the proposed modifications is evaluated through subjective listening tests in two reverberant conditions with reverberation time 0.8s and 2s. The combination of spectral transformation and uniform time-scaling is shown to be the most successful in increasing the intelligibility of casual speech. The evaluation results support the conclusion that modifications inspired by clear speech can be beneficial for the intelligibility enhancement of speech in reverberant environments.

**Index Terms**: Clear Speech, Casual Speech, Intelligibility, Reverberation, Spectral Transformations, Time Modifications, Pause insertion

10.21437/Interspeech.2015-28

## 1. Introduction

Clear speech is the speaking style elicited by speakers when the listener faces a communication barrier with the most common characteristic of slowing down and hyper-articulating. Conversely, casual speech is the type of speech produced when there is no barrier in the communication channel. The intelligibility advantage of clear speech vs. casual speech is proven under noisy and reverberant environments [1] and for various listener populations (hearing-impaired [2, 3, 4], non-native [5], and native listeners in noisy environments [1, 6, 7, 8]). The fact that clear speech can be applicable both in various intelligibility challenging conditions and in quiet motivates the modification of casual speech based on clear speech characteristics with a view to increasing its intelligibility in "noisy" channels and at the same time preserving its quality as these channels vary dynamically in real life.

Spectral transformations based on clear and casual speech differences [9, 10, 11] have been proven advantageous for speech intelligibility [12, 9, 13, 14, 15, 16]. In [16], a simple spectral transformation was proposed that does not require acoustic analysis of the speech to be modified. This method, called mix-filtering, boosts specific spectral regions as appearing on clear speech preserving the overall RMS energy. Previously, the modified casual speech was tested inside Speech Shaped noise (SSN) and was found more intelligible that unmodified speech while preserving its quality outside noise.

The intelligibility benefit of the mix-filtering method for SSN is explored here for reverberant environments. The motivation for proposing this technique in reverberant environments is that the mix-filtered modified speech simulates clear speech in terms of spectral energy distribution, which is resistant to reverberant environments [1]. In such environments, the intelligibility decrease of speech is due to (1) overlap masking effect where the energy of a phoneme is masked by the preceding one [17] (2) self-masking where the information is smeared inside a phoneme possibly as a result of flattened formant transitions [17]. As in clear speech, the mix-filtering approach boosts higher spectral regions, where transient parts are more likely to be found, and "steals" spectral energy from low-frequency energy which usually causes the overlap masking effect on the energy of a preceding phoneme. Other studies that successfully address the problem of intelligibility degradation on reverberant environments use steady-state suppression techniques to reduce steady-state portions of speech like vowel nuclei and to increase transient information [17, 18, 19]. Mix-filtering achieves, with less complexity, a similar acoustic result as steady state suppression and consonant emphasis since it does not require classification of speech portions.

The combination of the mix-filtering spectral technique along with time-scaling is explored, since spectral and time-scaling transformations, either natural (clear speech) or synthetic [20, 21], have been proven advantageous for speech intelligibility. Time-scaling schemes may enhance the intelligibility of unmodified speech through repetition of the information in time, reducing the overlap-masking and self-masking effect. The performance of two time-scaling techniques is evaluated for reverberant environments: 1) Uniform time-scaling 2) Time-scaling based on the Perceptual Quality Measure (PSQ) model. Uniform time-scaling changes the overall duration while respects the "local" speech rhythm. PSQ proposes both an elongation and a pause insertion scheme that could be beneficial inside reverberant environments as the energy of a speech segment falls into pauses and does not mask following segments. Unlike other proposed pause insertion schemes that are used for reverberation [20], this work explores a pause insertion scheme that inserts pauses in acoustically meaningful places.

Subjective evaluation of modified and unmodified speech is carried out via intelligibility tests on two reverberation times from non-native, native and hearing-impaired listeners. Unlike other studies that use a carrier sentence and non-sense syllables or rhyming words [22, 20, 21] to test word intelligibility, a more realistic scenario is used by testing sentence intelligibility.

Copyright © 2015 ISCA

## 2. Speech corpora

The corpora used for previous and current analysis is the read clear and read casual speech from the LUCID database. Read speech is an exaggerated form of speech relative to spontaneous speech and has higher intelligibility only for the clear style [23, 24]. The speakers participated on the recordings were normophonic Southern British English. The sentences recorded were meaningful and simple in syntax.

## 3. Spectral modifications

The spectral modification scheme explored for enhancing speech intelligibility for reverberant environments is the mix-filtering approach proposed in [16]. In [16], evaluation of the mix-filtered speech using an objective intelligibility metric [25], showed intelligibility enhancement of casual speech in SSN noise without degrading the original signal. Here, the mix-filtering technique is proposed for reverberant environments and is briefly described.

In [16], analysis performed between clear and casual speech revealed spectral differences between the two speaking styles. The analysis involved the extraction of spectral envelopes both for voiced and unvoiced segments for a large number of sentences of 8 different speakers, male and female. Then, averaged spectral envelopes were computed as the mean of all frames for each speaking style separately. Subtracting the average spectral envelopes of clear and casual speech revealed that clear speech has higher energy than casual speech in two frequency bands, $B1 = [2000, 4800]$ and $B2 = [5600, 8000]$ and $B2$ is more enhanced than $B1$. Based on these observations, the following modification scheme is proposed.

### 3.1. Modification algorithm

From the casual speech signal, the information corresponding to the frequency bands $B1$ and $B2$ is isolated and added to the original casual speech signal with different weighting factors. Then, the modified signal is normalized to have the same energy as the original signal. For the isolation of the frequency bands a simple method is used. Original speech $s$ is filtered with a 5-order bandpass digital elliptic filter with 0.1dB of ripple in the passband, and 60dB of attenuation in the stopband and bandpass edge frequencies $[2000, 4800]$. The output of the filter is the signal $s_1$ which contains information on the $B_1$ frequency band. Moreover, original speech $s$ is filtered with a 5-order highpass digital elliptic filter with normalized passband edge frequency $f_c = 5600$Hz. 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:

$$y[i] = w_0 s[i] + w_1 s_1[i] + w_2 s_2[i] \quad (1)$$

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

where, $y_{mixF}$ is the proposed spectrally modified signal, N is the number of samples of the original signal $s$ and $y$, and $w_0 = 0.1, w_1 = 0.4, w_2 = 0.5$ are the weighting factors of the signals $s, s_1$ and $s_2$, respectively. The values of the weighting factors were selected based on the clear and casual speech observations (the spectral energy in $B2$ is greater than in $B1$ and therefore $w_2 > w_1$) and on maximizing an objective intelligibility score, namely the Glimpse Portion, in the presence of low Signal-to-Noise Ratio SSN ($SNR = -10$dB) [26, 27].

The advantage of the method is its simplicity and efficiency and unlike other techniques [9, 13] it does not require frame-based analysis and modifications (detection of voiced/unvoiced regions, formant shaping etc). The mix-filtering technique can be proven advantageous for reverberant environments, as it enhances spectral information based on measured differences between the two speaking styles, clear and casual speech.

## 4. Time-scaling modifications

The performance of two time-scaling modifications are explored in reverberant environments. 1) Uniform time-scaling 2) Time-scaling based on the Perceptual Quality Measure (PSQ).

### 4.1. Uniform time-scaling

Uniform time-scaling is performed by feeding the Waveform Similarity Based Overlap-Add algorithm - WSOLA [28] a constant scale factor, that is the ratio of the casual speech signal duration to that of the clear signal. Then, WSOLA time scales the casual speech signal to match the duration of the clear one.

### 4.2. Perceptual Speech Quality Measure based Time-Scale Modifications

In this work, 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 PSQ measure is based on the basic version of ITU Standard REC-BS.1387-1-2001, a method for objective measurements of perceived speech quality. It estimates features such as loudness and modulations in specific frequency bands, in order to describe the input signal with perceptual attributes. The elongation and pause insertion scheme are described below.

#### *4.2.1. Elongation of voiced parts of speech*

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 frequency bands and the loudness modulations in high frequency 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, this metric is not sufficient for distinguishing stationary from non-stationary parts of speech, as some voiced stop consonants have high energy in low frequency bands. Time-scaling voiced stop consonants can cause distortion, probably not noticeable in reverberation but it may also reduce the phoneme's intelligibility. Therefore, the loudness is not the appropriate metric to decide which parts of speech should be elongated.

The combination of the loudness with another metric, namely the loudness modulations of high frequency bands (around 4000Hz) is proposed as a more efficient technique to detect stationary. 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. Subtracting the modulation values (near zero for vowels, high for unvoiced and voiced consonants) from the loudness values (high for vowels, near zero for unvoiced consonants and high for voiced consonants) the stationary parts are detected more efficiently than based on a purely loudness metric (after subtraction values are high for vowels, negative for unvoiced and near zero

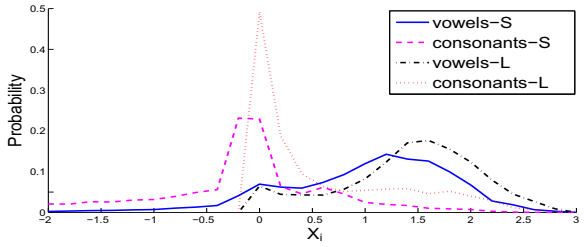



Figure 1: *Normalized histograms of vowels and voiced consonants {b,g,d,l} based on the values of the two metrics, the loudness metric L and the proposed metric S. The proposed metric makes a smaller classification error for the voiced consonants*

for voiced consonants). The efficiency of the metric to classify voiced consonants as non-stationary is estimated. Specifically, 100 sentences are annotated for a speaker in our database to distinguish consonant-frames from vowel-frames. Then, the average perceived loudness in low frequency bands $L$ and loudness modulations in high frequency bands $M$ are calculated per frame. The difference $S$ between these values for each frame is calculated and the normalized histograms of $S$ for vowel and the voiced consonant frames {b, g, w, l} and the corresponding normalized histograms of L are depicted on Figure 1. Selecting a decision threshold $T > 0.5$ for consonant and vowel classification, the misclassification error of the proposed metric S for the consonants (the area below the consonant curve on the interval [0.5, 3]) is lower than that of the L metric. Therefore, the proposed scheme decides to elongate a speech frame if the S value is above $T = 1$. Each frame allowed to be elongated by the S metric is time-scaled by 20% of its original duration. The time-scaling is performed by WSOLA.

#### *4.2.2. Pause Insertion*

Pause insertion is also implemented using the PSQ model. The proposed pause insertion scheme is purely unsupervised and takes into consideration the acoustic properties of casual speech. Specifically, the perceived loudness of the speech signal in the whole frequency band is estimated (in dB SPL). Then, loudness is normalized by the maximum loudness of the signal and all valleys are detected on the normalized loudness curve. PSQ adds pauses on valleys with less than 20% of the normalized loudness. The valleys are usually in the middle of word boundaries and are appropriate for inserting pauses without distorting the signal. A pre-processing of the signal before and after the location of the valley is performed; the signal is time-scaled around the location where the pause will be inserted and a hamming window is applied on the center of the valley, so that the transition from speech to silence will be more smooth. Inserted pauses have a fixed length of $90ms$ based on average pause duration on clear speech.

## 5. Evaluations

In this section the proposed modifications are evaluated in reverberant conditions. Reverberation is simulated using a room impulse response (RIR) model obtained with the source-image method [29]. The hall dimensions are fixed to $20\ \mathrm{m} \times 30\ \mathrm{m} \times 8\ \mathrm{m}$. The speaker and listener locations used for RIR generation are $\{10\ \mathrm{m},\ 5\ \mathrm{m},\ 3\ \mathrm{m}\}$ and $\{10\ \mathrm{m},\ 25\ \mathrm{m},\ 1.8\ \mathrm{m}\}$ respectively. The propagation delay and attenuation are normalized to the direct sound. Effectively, the direct sound is equivalent to the sound output from the speaker. Convolving the modified speech signals with RIR produces the signals for evaluation.

Seven sets of signals are evaluated: (1) the clear speech (CL), (2) the casual speech (CV) (3) the mix-filtering spectrally modified casual speech signal (M) (4) the uniformly time-scaled casual speech signal (U) (5) the PSQ-based time-scaled casual speech signal (P) and the combinations of the above modifications (6) uniform time-scaling and mix-filtering of casual speech (UM) (7) PSQ-based time-scaling and mix-filtering of casual speech (PM). The term Categories will be used to refer to the seven sets of signals. 56 randomly selected distinct sentences from the LUCID corpus are presented to the listeners, uttered from 2 Male and 2 Female speakers (14 sentences per speaker, 8 sentences per set of signals, 4 sentences per Category per reverberant condition). The reverberation times are $RT_1 = 0.8s$ and $RT_2 = 2s$ to simulate low and high reverberant environments, respectively. A "header" of 4 sentences is added to the listening test to serve as a preparation set for the listeners to the reverberant environment (these sentences are not evaluated). The listener hears each sentence once and is instructed to write down whatever he/she perceives to have heard.

As sentence difficulty may affect the intelligibility scores (especially for the non-native population), 7 different listening scenarios have been created to ensure that each sentence will be presented in a {CL, CV, M, U, P, UM, PM} manner to different listeners (as each listener cannot hear the same sentence twice). For example, if a specific sentence is presented to the listener in CL manner on $RT_1$ condition on the listening Scenario 1, then the same sentence will be presented to another listener in CV manner on the same reverberant condition on listening Scenario 2 etc. This allows us to "denoise" the performance evaluation from the sentence dependency.

### 5.1. Evaluation part

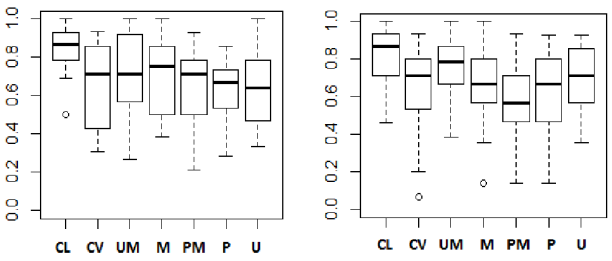
32 listeners participated in the intelligibility test, 7 native speakers, 4 hearing-impaired listeners, and 21 non-native speakers with good perception of English (this was also verified in the listening test with 5 difficult sentences presented without reverberation conditions). As the majority of the listeners are non-native, explicit statistic analysis is presented for this population.

#### *5.1.1. Non-native speakers*

Performance evaluation for the non-native speakers contains three parts of analysis. The first part presents the intelligibility scores of each Category across listeners, in order to reveal possible intelligibility benefits of the proposed modifications for the non-native population. The second part of analysis computes the intelligibility scores of each Category across sentences, to parcel out the possible variability due to sentence difficulty and reveal the Category main effect. Lastly, the third part of analysis presents the intelligibility scores of each Category across the two different reverberant conditions.

For each reverberant condition, the ratio of the correct keywords to the number of total keywords per sentence is estimated per listener and per Category. Then, the mean of the ratios for all sentences is estimated per listener and per Category. Figure 2 shows the {min, $1^{st}$ quartile, median, $3^{rd}$ quartile, max} of intelligibility scores per Category across all listeners. CL appear to have a higher intelligibility advantage over all Categories for both reverberant conditions while the UM seems to have a benefit over CV on $RT_2$ (Figure 2).

In order to evaluate the statistical significance of these re-




(a) (b)

Figure 2: *Intelligibility scores per Category across listeners on (a)* $RT_1 = 0.8s$ *(b)* $RT_2 = 2s$

sults, a repeated-measures ANOVA is performed on intelligibility with Category nested within each listener. Results reveal significant intelligibility differences among Categories, for both reverberant conditions $RT_1$ ($F(6, 20) = 5.601, p < 0.001$) and $RT_2$ ($F(6, 20) = 7.167, p < 0.001$). Post-hoc comparisons using pairwise paired t-tests reveal that the mean intelligibility score of CL ($M = 0.86, SD = 0.13$, M stands for mean and SD for standard deviation) is significantly different ($p < 0.001$) from CV ($M = 0.67, SD = 0.22$) in $RT_1$ while in $RT_2$ both CL ($M = 0.83, SD = 0.16$) and UM ($M = 0.77, SD = 0.17$) have significantly different means ($p < 0.01$) from CV ($M = 0.64, SD = 0.23$). No significant difference between means of CL and UM are reported ($p = 0.07$).

A repeated-measures ANOVA on intelligibility with Category nested within each sentence is performed to remove possible dependencies of the intelligibility scores on sentence difficulty. ANOVA null hypothesis of equal means of the intelligibility scores for every Category, is rejected using the F-test for $RT_1$ ($F(6, 27) = 6.634, p < 0.001$) and $RT_2$ ($F(6, 27) = 7.268, p < 0.001$). Post-hoc comparisons using pairwise paired t-tests reveal that the mean intelligibility score of CL ($M = 0.87, SD = 0.15$) is significantly different ($p < 0.01$) from CV ($M = 0.66, SD = 0.20$) in $RT_1$ while in $RT_2$ both CL ($M = 0.83, SD = 0.18$) and UM ($M = 0.75, SD = 0.21$) have means different from CV ($M = 0.63, SD = 0.26$) and this result is statistical significant ($p < 0.001$ for CL, $p < 0.01$ for UM). No significant differences are reported between the means of CL and UM ($p = 0.07$). The mean of UM is significantly different from the means of all other modifications ($p < 0.01$). Last, pairwise paired t-tests showed no significant difference between means per Category in $RT_1$ with their corresponding in $RT_2$.

#### *5.1.2. Native listeners and Hearing-Impaired*

Subjective evaluations are also performed by 7 native listeners. As the sentences were meaningful, the content helped the native listeners to understand both CL and CV speech almost 100%. One listener appeared to have intelligibility score below 70% for both speaking styles. That listener benefits from all modification techniques in $RT_2$, and in $RT_1$ from all modifications except uniform-time scaling. Repeated measures ANOVA showed no statistical significant differences between Categories both for $RT_1$ ($F(6, 6) = 1.544, p = 0.192$) and $RT_2$ ($F(6, 6) = 1.781, p = 0.131$).

Subjective evaluations were also performed by 4 non-native hearing impaired listeners. CL speech was more intelligible than CV speech in $RT_1$ ($M_{CL} = 0.87, SD_{CL} = 0.16$, $M_{CV} = 0.60, SD_{CV} = 0, 22$) and $RT_2$ ($M_{CL} = 0.80, SD_{CL} = 0.23$, $M_{CV} = 0.57, SD_{CV} = 0.17$). In $RT_2$ condition, modification schemes failed to increase the intelligibility of casual speech. However, for $RT_1$, all listeners showed an intelligibility increase of modified casual speech with the mix-filtering modification ($M = 0.86, SD = 0.26$). Repeated measures ANOVA showed no statistical significant differences between Categories for $RT_1$ ($F(6, 3) = 1.754, p = 0.166$) and $RT_2$ ($F(6, 3) = 3.228, p = 0.0248$).

### 5.2. Discussion

Subjective evaluations presented in this experiment confirm that clear speech is more intelligible than casual speech in reverberant conditions for the non-native listeners. Indeed, CL outperforms CV by 19% in 0.8s and 2s reverberant time. Non-native listeners also report that the combination of uniform time scaling and mix-filtering technique is advantageous for $RT_2$ since the intelligibility benefit is 13%, 6% lower from the upper bound (CL). However, in less reverberation time, the benefit of this modification drops. This inefficiency is possibly due to the selection of the uniform-time scaling factor. Figure 2(a) shows that the mix-filtering technique has a slight advantage over casual speech. Then, when uniform-time scaling is combined with the spectral boosting, the median intelligibility score drops and the variance increases. Therefore, this result indicates that the time-scaling factor is important for reverberant environments and its selection should be proportional to the reverberation time. Also, the PSQ-based modification fails to increase intelligibility of casual speech. One possible reason for this is the change of rhythm between speech segments and the extreme elongation in some cases. A more conservative time-scaling factor could be proven more advantageous for the time-scaling techniques and is to be explored in the future.

The hearing-impaired population is rather small to draw any concrete conclusions. However, the clear speech intelligibility advantage is 23% and 27% higher than that of casual for $RT_1$ and $RT_2$, respectively and the mix-filtering in $RT_1$ increases the intelligibility of casual speech by 26%.

Finally, native listeners do not benefit from the transformations as the intelligibility of CV is as high as that of CL, highlighting the importance of the semantic content and/or the amount of reverberation, above which their perception is degraded (possibly on higher reverberation times).

## 6. Conclusions

Different time and spectral techniques for increasing the intelligibility of casual speech for reverberant environments are explored, inspired by clear speech which is proven to be robust in reverberant conditions. The proposed modification uses a combination of spectral boosting and uniform time-scaling. Our spectral transformation applies a multi-band filtering on casual speech, boosting information from important frequency bands indicated by clear speech and it has low computational complexity as it does not require detection of steady-state portions. The mix-filtering and uniform time-scaling combination increases the intelligibility of casual speech in high reverberant environments ($RT = 2s$) for the non-native population. Results indicate that modifications based on clear speech properties can be beneficial for the intelligibility enhancement of casual speech in reverberant environments. A more refined selection of the uniform-time scaling factor according to the degree of reverberation is to be explored in the near future.

## 7. References


[1] K. Payton, R. 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.*, vol. 95(3), pp. 1581–92, 1994.

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

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

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

[5] A. Bradlow and T. Bent, "The clear speech effect for non-native listeners," *J. Acoust. Soc. Amer.*, vol. 112, no. 1, pp. 272–284, 2002.

[6] S. Ferguson, "Talker differences in clear and conversatoinal speech: Vowel intelligibility for normal-hearing listeners," *J. Acoust. Soc. Amer.*, vol. 116, no. 4, pp. 2365–2373, 2004.

[7] 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.*, vol. 112, pp. 2165–72, 2002.

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

[9] J. Krause and L. Braida, "Evaluating the role of spectral and envelope characteristics in the intelligibility advantage of clear speech," *J. Acoust. Soc. Amer.*, vol. 125, no. 5, pp. 3346–3357, 2009.

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

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

[12] R. 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*, vol. 24, no. 4, pp. 277–282, 1976.

[13] 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.

[14] T. Zorila, V. Kandia, and Y. Stylianou, "Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression," *Interspeech 2012, Portland Oregon, USA*, pp. 635–638, September 2012.

[15] E. Godoy, M. Koutsogiannaki, and Y. Stylianou, "Approaching speech intelligibility enhancement with inspiration from lombard and clear speaking styles," *Comput. Speech Lang.*, vol. 28, no. 2, pp. 629–647, 2014.

[16] M. Koutsogiannaki and Y. Stylianou, "Simple and artefact-free spectral modifications for enhancing the intelligibility of casual speech," in *ICASSP*, May 2014, pp. 4648–4652.

[17] A. K. Nabelek, T. R. Letowski, and F. M. Tucker, "Reverberant overlap- and self-masking in consonant identification," *J. Acoust. Soc. Amer.*, vol. 86, no. 4, pp. 1259–1265, 1989.

[18] T. Arai, K. Kinoshita, N. Hodoshima, A. Kusumoto, and T. Kitamura, "Effects of suppressing steady-state portions of speech on intelligibility in reverberant environments," *Acoust. Sci. Tech.*, vol. 23, no. 4, pp. 229–232, 2002.

[19] N. Hodoshima, D. Behne, and T. Arai, "Steady-state suppression in reverberation: a comparison of native and non-native speech perception," *Interspeech 2006*, pp. 873–876, 2006.

[20] T. Arai, "Padding zero into steady-state portions of speech as a preprocess from improving intelligibility in reverberant environments," *Acoust. Sci. Tech.*, vol. 25, no. 5, pp. 459–461, 2005.

[21] T. Arai, Y. Nakata, N. Hodoshima, and K. Kurisu, "Decreasing speaking rate with steady-state suppression to improve speech intelligibility in reverberant environments," *Acoust. Sci. Tech.*, vol. 28, no. 4, pp. 282–285, 2007.

[22] Y. Nakata, Y. Murakami, N. Hodoshima, N. Hayashi, Y. Miyauchi, T. Arai, and K. Kurisu, "The effects of speech-rate slowing for improving speech intelligibility in reverberant environments," *International Workshop on Frontiers in Speech and Hearing Research, Technical Report of IEICE Japan*, 2006.

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

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

[25] Y. Tang, M. Cooke, and C. Valentini-Botinhao, "P16: A distortion-weighted glimpse-based intelligibility metric for modified and synthetic speech," *Speech in Noise Workshop*, 2013.

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

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

[28] M. Demol, K. Struyve, W. Verhelst, H. Paulussen, P. Desmet, and P. V. Author, "Efficient non-uniform time–scaling of speech with wsola for call applications," *Proc. of InSTIL/ICALL2004–NLP and Speech Technologies in Advanced Language Learning Systems, Venice*, 2004.

[29] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.