# Increasing Speech Intelligibility via Spectral Shaping with Frequency Warping and Dynamic Range Compression plus Transient Enhancement

*Elizabeth Godoy, Yannis Stylianou*

Institute of Computer Science, Foundation of Research and Technology Hellas, Crete, Greece

## Abstract

In order to make speech (natural or synthetic) more intelligible for listeners in real-world noisy environments, various modifications have been proposed that exploit spectral and temporal signal features. Previously, an evaluation campaign involving several approaches illustrated that a Spectral Shaping (SS) and Dynamic Range Compression (DRC) method proved highly successful at increasing speech intelligibility. For the public follow-up campaign (i.e., the Hurricane Challenge), this work introduces additional modifications into SSDRC in an attempt to further enhance intelligibility. First aiming to slow down the articulation rate, the speech is uniformly time stretched to effectively increase signal redundancy. Second, a frequency warping mechanism to expand vowel space is incorporated into the SS. Third, scaling to enhance the transient regions of speech is applied in the time-domain along with DRC. Objective and extensive subjective (i.e., the Hurricane Challenge) evaluations show that the new approach successfully achieves intelligibility gains over natural speech for all of the noise conditions evaluated, though compared to SSDRC, there is less advantage observed at higher SNR.

**Index Terms**: speech intelligibility, spectral shaping, frequency warping, dynamic range compression

## 1. Introduction

With growing numbers of applications (commercial, military, medical, etc) using speech technologies, listeners in real-world scenarios now often hear speech in noisy environments. Consequently, there is great interest in developing intelligibility enhancement methods for devices that use recorded or synthesized speech in order to ultimately increase their effectiveness and relevance. In this vein, a variety of approaches have been proposed that can be generally classified into several groups. First, there are techniques that exploit audio and signal properties, such as the amplitude compression scheme in [1], dynamic range compression in [2] and a method for peak-to-rms reduction in [3]. Second, certain speech intelligibility enhancement methods focus on speech-in-noise and exploit knowledge of the noise masker, such as the optimizations based on a speech intelligibility index in [4] and the glimpse proportion maximization in [5]. Third, in the context of text-to-speech systems, adaptation or synthesis approaches for speech-in-noise have been explored to increase intelligibility, as in [6, 7]. Fourth, certain techniques aim to study and exploit the impact of particular acoustic features of speech with respect to a given speech style. For example, considering Lombard speech, the role of spectral modifications and fundamental frequency was examined in [8] and a modification using a "Lombard" correction filter was examined in [9]. For Clear speech, the relative intelligibility impact of several acoustic features was similarly examined in [10, 11] and an example of time-domain algorithms to slow down speech can be found in [12]. Moreover, several speech intelligibility enhancement approaches combine methods from above, such as the glimpse proportion maximization for Hidden Markov Model (HMM) -based text-to-speech synthesis in [13].

In 2012, an extensive evaluation of the intelligibility impact of a variety of methods was carried out in the campaign described in [14]. Emerging as the most successful modification from this campaign was the combination proposed in [15] of Lombard-like Spectral Shaping (SS) and audio enhancement with Dynamic Range Compression (DRC), with performance being particularly strong in the presence of high noise levels. While originally applied to enhance the intelligibility of natural speech, an additional evaluation campaign for TTS speech intelligibility enhancement has proven that SSDRC is also successful as a post-processing technique for synthetic speech [16]. In this work, SSDRC is used both as a starting-block and comparative standard for the modifications examined.

Specifically, in preparation for the Hurricane Challenge (a public follow-up to the evaluation campaign in [14]), a selection of additional time- and spectral- domain modifications was incorporated in SSDRC to create the new "uwSSDRCt." These additional modifications exploit signal processing techniques in an effort to mimic acoustic trends observed in "Clear" speech, which represents a human speaking style that is highly intelligible both in quiet and in noise [17, 18, 19, 20, 21]. First, observing that Clear speech exhibits decreased speaking rate, uniform time stretching is applied, giving the "u" in uwSSDRCt. It should be noted, however, that the acoustic-phonetic traits associated with the decreased speaking rate and greater articulation of human Clear speech are quite complex [12]. Consequently, the uniform time-stretching here takes a simplified approach and functions mainly to slow down the rate at which the listener hears the speech, effectively increasing redundancy. Second, inspired by observations from Clear speech that increased vowel space area reflects greater intelligibility, a frequency warping technique for vowel space expansion is incorporated in SS, giving the "w" in uwSSDRCt. Third, in an attempt to emphasize important acoustic-phonetic cues, as suggested in [22], transient enhancement is applied after DRC, noted by the "t" in uwSSDRCt. Finally, the results from the Hurricane Challenge demonstrate that uwSSDRCt is largely successful at increasing intelligibility, achieving gains over natural speech for all maskers and SNR levels in a manner comparable to SSDRC, though less so at high SNR.

The structure of this article is as follows. Section 2 successively details the new modifications in uwSSDRCt. Section 3 then presents results using an objective extended Speech Intelligibility Index and also the results from the Hurricane Challenge. Finally, Section 4 concludes.

Thanks to EU Future and Emerging Technology (FETOPEN) Project LISTA (The Listening Talker) for support and encouragement.

## 2. Method (uwSSDRCt) Modifications

The following subsections respectively detail the modifications in uwSSDRCt, beginning with the uniform time-stretching and then proceeding to the spectral domain (wSS) and, finally, time-domain (DRCt) modifications. Particular emphasis is given to the new techniques, i.e. frequency warping and transient enhancement, incorporated into SSDRC.

### 2.1. Uniform Time Stretching

As mentioned previously, the goal of the uniform time-stretching applied in uwSSDRCt is mainly to increase the redundancy of the speech that the listener hears in noise. In this way, particularly for a competing speaker masker, the repetition of speech frames could lead the available information into a relative null or region of low noise (e.g., a pause or silence in the competing speech). Consequently, the listener has more of a chance at hearing the speech.

Specifically, each sentence was uniformly time-stretched using WSOLA [23] in order to fill approximately 800ms of available space (note that, in the Hurricane Challenge, 500ms silences were present at the beginning and end of the unmodified speech). Half of the time-stretching interval was filled at the beginning and half at the end of the sentence. In hindsight, it may have been beneficial to fill the ending silence more in order to leave a delay greater than 100ms at the beginning of the sentence, as it has since been suggested to the authors that the human auditory system might not be maximally responsive to speech introduced in noise at this time-delay [24]. In particular, the work in [24] suggests that a delay of 200ms might be more beneficial.

### 2.2. Spectral Shaping with Frequency Warping

The spectral domain modifications in uwSSDRCt begin with the SS proposed in [15], which takes the form of a series of filters applied to the amplitude spectrum of each frame. In particular, the Lombard-inspired fixed filter ($H^r(f)$) in SS significantly increases the audibility of formants by increasing spectral energy significantly in a 1-4kHz frequency band. Additionally, SS also includes adaptive peak sharpening ($H^s(f)$) and pre-emphasis ($H^p(f)$) filters. The details of these respective filters can be found in [15]. In this work, the primary spectral-domain contribution in uwSSDRCt is the incorporation of a frequency warping filter "w" in SS that is designed to expand vowel space.

Specifically, the frequency warping filter $H^W(f)$ is inspired by the observed vowel space expansion in Clear speech that is often largely attributed with the intelligibility gains of the style [25, 26, 27]. The basics of the algorithm are presented here.

First, let the spectral envelope from frame $i$ of unmodified speech be $S_i^X(f)$. For example, the True Envelope is used here with a cepstral order of 48 [28]. The frequency warping filter for frame $i$, $H_i^W(f)$, is then defined as

$$H_i^W(f) = S_i^X(W_i^{-1}(f))/S_i^X(f) \tag{1}$$

where $W_i^{-1}(f)$ is a piecewise linear warping function (the form of which is used for Voice Conversion in [29]), given for $f \in [f_{ni,l}^X, f_{i,l+1}^X]$ by

$$W_i(f) = A_{i,l}f + B_{i,l} \tag{2}$$

where $f_{i,0}^X = 150Hz, f_{i,M_i+1}^X = 3500Hz$, and

$$A_{i,l} = \Delta(f_{i,l+1}^X) - \Delta(f_{i,l}^X) \tag{3}$$

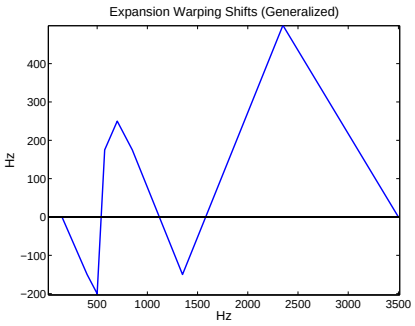


Figure 1: $\Delta(f)$ - *Generalized curve of warping shifts used in the frequency warping for vowel space expansion.*

$$B_{i,l} = f_{i,l}^X(1 - A_{i,l}) + \Delta(f_{i,l}^X) \tag{4}$$

and $f_{i,l}^X$ indicates the frequency of the $l^{th}$ spectral envelope peak detected in frame $i$, $l = 1, ...M_i$. Note that the peaks $f_i^X$ are detected as maxima (following a minima lower by at least 10%) of the tilt-normalized (the first two cepstral coefficients are removed) spectral envelope. Also, the curve given in Fig. 1 of exaggerated warping shifts $\Delta(f)$ is drawn from the first and second formant shifts observed in going from conversational to Clear speech. Specifically, the curve $\Delta(f)$ exaggerates formant shifts in order to overcome the harmonic structure in the amplitude spectrum. Additionally, the slope of $\Delta(f)$ were adjusted to ensure that formant overlaps were avoided (i.e. all warped frequency axes have non-negative slope). As can be seen from Fig. 1, this curve essentially shifts low/high formants (F1 and F2) down/up in order to ultimately expand the vowel space.

Now, considering the full combination of the SS and frequency warping filters, the wSS modification for frame $i$, with amplitude spectrum $E_i(f)$, is given by

$$\hat{E}_i(f) = E_i(f)H_i^p(f)H_i^s(f)H^r(f)H_i^W(f) \tag{5}$$

where the first three filters are from the SS given in [15] and the frequency warping filter $H_i^W(f)$ is described above. As described in [15], the signal resynthesis is achieved via overlap-add using the modified amplitude spectrum $\hat{E}_i(f)$ and the original phase spectrum.

### 2.3. DRC plus Transient Enhancement

The DRC described in [15] is an audio enhancement that effectively reduces the peak-to-rms ratio across the sentence, augmenting loudness. This DRC can be described as a scaling, here noted by $g_{DRC}(n)$, multiplying the speech signal $s(n)$. In addition to this scaling, uwSSDRCt incorporates a transient enhancement noted by $g_t(n)$. This transient enhancement is motivated by the cue enhancement work in [22], with the principle idea being to increase the energy around speech transients, as they hold important acoustic-phonetic information. In this work, the scaling $g_t(n)$ is determined based on a non-stationarity metric described in [30] and used as a classification for speech stationarity in [9]. Specifically, the process for calculating $g_t(n)$ is as follows.

To begin with, in order to estimate the signal non-stationarities, the transition rate is calculated using both signal energy in time and spectral information, as described in [30]. This transition rate is then spline-interpolated between frames so that it has a value at each sample. Let $T(n)$ represent this resulting stationarity metric (normalized so that the maximum

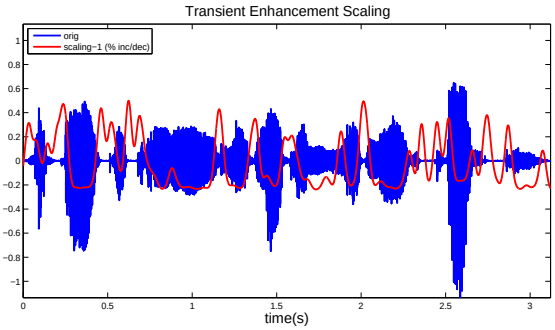


Figure 2: *Transient Enhancement Scaling, shown as a percent increase or decrease in energy at each sample (specifically* $g_t - 1$ *in red).*

value in the sentence is 1). Since the goal of the transient enhancement is to emphasize transitions in the signal, or regions around which there are non-stationarities, the curve $T(n)$ is broadened by convolution with a 80ms Gaussian window $G(n)$

$$T_w(n) = T(n) * G(n) \tag{6}$$

The scaling factor for the transient enhancement is then given as

$$g_t(n) = 0.75(T_w(n) + 1) \tag{7}$$

so that the region of maximal non-stationarity in the signal is enhanced by 50% and no part of the signal is reduced by more than 25%. That said, the transient-enhanced signal is ultimately coupled with the DRC gain and the result is rms-normalized to match the energy of the unmodified signal. Consequently, the specific decreases and increases in signal energy depend on the DRC gain as well as the energy distribution of the signal over time. An example of an original sentence (normalized by its maximal absolute value) and the calculated scaling $g_t(n) - 1$ is given in Fig. 2. The gain is shown minus 1 in order to indicate the relative percent increase/decrease dictated by the transient enhancement scaling. At the same time, visually, the scaling now lies on-top of the signal (in a range $[-.25, .5]$ correspondingly to a 25% decrease and 50% increase, respectively) and can more readily highlight detected non-stationarities.

Now, given the DRC $g_{DRC}$ and transient enhancement $g_t$ scaling, the DRCt modification of $s(n)$ is given by

$$s_{DRCt}(n) = g_{DRC}(n)g_t(n)s(n) \tag{8}$$

where $s_{DRCt}(n)$ is the modified signal.

### 2.4. Full uwSSDRct Modification

Combining all of the above, the full modification with uwSSDRCt thus begins with uniform time scaling using WSOLA, followed by frame-by-frame spectral modification given by Eq. 5. As described in subsection 2.2, the signal is then resynthesized using the modified amplitude spectrum with the original phase spectrum and overlap-add. Next, the time-domain gain filters are applied, as in Eq. 8. Finally, the final signal is rms-normalized to match that of the unmodified signal. Fig. 3 shows an example of the proposed speech modification, with the first sentence in the Harvard corpus from the speaker used in the Hurricane Challenge. The sampling frequency is 16kHz. Below the original sentence is the speech modified by uwSSDRCt. The most apparent modifications that can be seen in the view shown

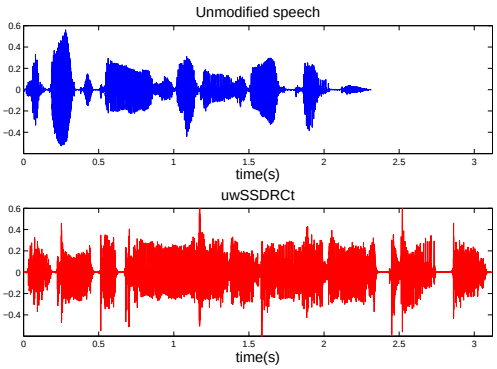


Figure 3: *An example of applying the suggested algorithm on a speech signal: original (above) and modified with uwSSDRCt (below).*

in Fig. 3 relate to the time-domain. Specifically, the uniform time stretching is indicated by the longer signal, the DRC is apparent in the signal "flattening" in time, and the slight transient enhancement yields more apparent peaks in energy around transitions.

## 3. Objective and Subjective Evaluations

The original evaluation campaign (via listening tests) described in [14] assessed the intelligibility impact of various speech modifications. The subsequent Hurricane Challenge follows the same criteria and considers the same sentences, noise maskers and Signal-to-Noise Ratio (SNR) conditions as its predecessor. Specifically, two types of noise maskers are examined: competing speaker (CS) and speech shaped noise (SSN). The SNRs were then determined in order to represent [Lo,Mid,Hi] levels of noise, with [-1,-4,-7] dB for the CS and [-7,-14,-21] dB for SSN.

### 3.1. Objective Extended Speech Intelligibility Index

Before proceeding to the Hurricane Challenge Results, the following offers an initial objective evaluation of uwSSDRCt using the extended Speech Intelligibility Index (extSII) described in [31] and used for objective evaluations in [15]. Since the number of Hurricane Challenge entries was limited and only uwSSDRCt in its entirety was evaluated, the objective evaluations here serve to highlight components within uwSSDRCt. Fig. 4 shows the objectively intelligibility via extSII for the different noise conditions and levels evaluated in the Hurricane Challenge. In Fig. 4, the extSII for the plain (unmodified) speech is given as well as that for the proposed uwSSDRCt. Then, in order to indicate the relative objective gains of the time and spectral domain modifications, wSS and DRCt are examined, with both being applied on the uniformly time-stretched speech "u" in order to control for this initial modification. As can be seen in Fig. 4, the extSII indicates a significant gain of uwSSDRCt over plain speech, specifically compounding the relative gains of the time (DRCt) and spectral (wSS) domain modifications, with even more gain indicated for the SSN. It should also be noted that the objective difference between uwSSDRt and the original SSDRC was not significant, so these methods are compared based on their relative subjective scores in the next subsection.

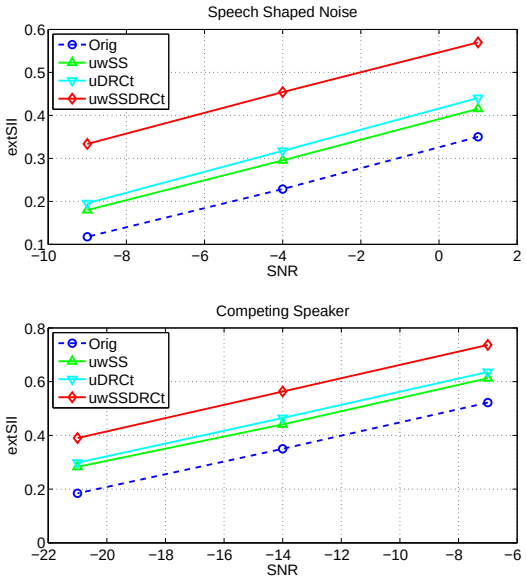


Figure 4: *extSII Results for the noise conditions evaluated in the Hurricane Challenge. The objective results for plain speech are given along with those for uwSSDRCt and splitting the temporal and spectral modifications respectively (after time-stretching) in uwSS and uDRCt.*

### 3.2. Performance in the Hurricane Challenge

In the Hurricane Challenge, 219 listeners participated before being screened for auditory deficiencies and non-nativeness. Of these participants, the results of 175 listeners were used for the final evaluations. During the test (cf [14]), listeners heard speech samples (plain, TTS and a variety of modified) in the specified noise conditions and then transcribed what they think they heard. The results for the uwSSDRCt modification (applied on plain speech) are given in Table 1 in mean percentage points, representing the percent of correct keyword identification, plus or minus the standard error. The results for the intelligibility of plain speech are also provided in Table 1 as a reference. The relative gains of uwSSDRCt over plain speech in percentage points is also provided, calculated based on fits to psychometric functions (i.e. intelligibility versus SNR) for each of the two masker types [14]. Additionally, the corresponding gains of the original SSDRC evaluated in [14] are given for comparison.

As can be discerned from Table 1, uwSSDRCt yields intelligibility gains at all SNRs for both of the noise maskers. In particular, the advantage of uwSSDRCt is more significant at low SNR (i.e., high noise levels). This advantage is no doubt due to the increased audibility and loudness of the modified speech, as in SSDRC. That is, these methods are most effective at enhancing intelligibility in the presence of significant noise levels that mask a wider range of cues in the speech signal. On the other hand, at high SNR (in the presence of little noise), listeners can pick up on more (secondary) cues that help to determine the intelligibility. In these cases, the proposed uwSSDRCt is less effective, as is the original SSDRC. Comparing the gains of uwSSDRCt and SSDRC, they are comparable overall. However, it is evident that the additional modifications examined in this work are not positively influencing intelligibility at higher SNRs, particularly for the SSN masker. Finally, in comparing the objective extSII scores with those of the listening tests, the greatest discrepancy is in distinguishing between the different SNR levels. That is, the extSII exhibits relatively constant gains across SNRs while the listener responses vary significantly based on the noise levels.

Table 1: *Hurricane Challenge Results for uwSSDRCt. Values are given in percentage points (pp) +/- standard error. The gain of uwSSDRCt over plain (natural) speech, calculated from fits to psychometric functions, is indicated in bold. The gains of SSDRC, similarly calculated from the original campaign in [14], are given in italics.*

| mask,SNR | plain | uwSSDRCt | gain | *gain(I)* |
|---|---|---|---|---|
| CS, Hi | 85.1 +/- 1.5 | 87.5 +/- 1.2 | **2.3** | *5.5* |
| CS, Mid | 57.0 +/- 2.4 | 71.2 +/- 1.9 | **14.2** | *14.0* |
| CS, Lo | 24.8 +/- 1.9 | 40.4 +/- 2.1 | **15.5** | *15.4* |
| SSN, Hi | 88.31 +/- 1.3 | 88.8 +/- 1.2 | **0.5** | *7.6* |
| SSN, Mid | 63.0 +/- 2.2 | 83.1 +/- 1.5 | **20.1** | *29.3* |
| SSN, Lo | 17.3 +/- 1.8 | 53.8 +/- 2.2 | **36.6** | *36.5* |

### 3.3. Discussion

One way to interpret the motivations underlying the new modifications introduced in uwSSDRCt is the attempt to mimic acoustic trends observed in human Clear speech in order to enhance intelligibility, particularly at high SNR (where SSDRC is less performant). However, the acoustic-phonetic cues that speakers produce and that listeners exploit in their perception of speech functions on many levels, for example, from the segmental acoustic characteristics to particular Vowel-Consonant (CV or VC) pairings, prosody and stress, lexical structure and word confusability, to name a few considerations. Unfortunately, the purely-acoustic modifications adopted in this work to enhance SSDRC were apparently unsuccessful in furthering intelligibility gains. Consequently, the present results would suggest that alternative or additional acoustic-phonetic levels should be examined in an effort to achieve gains at high SNR, similar to those observed for Clear speech.

## 4. Conclusions

This work detailed temporal and spectral modifications that were incorporated into SSDRC in an attempt to further enhance speech intelligibility for the Hurricane Challenge. The proposed uwSSDRCt included acoustic modifications based generally on observations from Clear speech, specifically uniform time-stretching, frequency warping for vowel space expansion and transient enhancement. While uwSSDRCt ultimately achieved intelligibility gains at all SNR levels for both noise maskers, the relative gain compared to SSDRC (as evaluated in the original evaluation campaign that inspired the public Hurricane Challenge) was less at higher SNR. Rather than focus purely on acoustic features, future work will seek to incorporate more phonetic and linguistic information in speech analyses and corresponding modifications, for example, considering keyword vowel-formant transitions in isolated CV-VC contexts.

# 5. References


[1] R. Niederjohn and J. Grotelueschen, "The enhancement of speech intelligibility in high noise levels by high-pass filtering followed by rapid amplitude compression," *IEEE Trans Audio, Speech, Lang Processing*, vol. 24, no. 4, pp. 277–282, 1976.

[2] B. Blesser, "Audio dynamic range compression for minimum perceived distortion," *IEEE Trans. Audio Acoust.,*, vol. 17, no. 1, 1969.

[3] T. Quatieri and R. McAulay, "Peak-to-rms reduction of speech based on a sinusoidal model," *IEEE Trans. on Signal Processing*, vol. 39, no. 2, pp. 273–288, 1991.

[4] B. Sauert and P. Vary, "Near end listening enhancement: speech intelligibility improvement in noisy environments," in *ICASSP*, 2006, pp. 493–496.

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

[6] B. Langner and A. Black, "Improving the understandability of speech synthesis by modeling speech in noise," in *ICASSP*, vol. I, 2005, pp. 265–268.

[7] T. Raitio, A. Suni, M. Vainio, and P. Alku, "Analysis of HMM-based lombard speech synthesis," in *Interspeech*, 2011, pp. 2781–2784.

[8] Y. Lu and M. Cooke, "The contribution of changes in f0 and spectral tilt to increased intelligibility of speech produced in noise," *SpeechComm*, no. 51, pp. 1253–1262, 2009.

[9] E. Godoy and Y. Stylianou, "Unsupervised acoustic analyses of normal and lombard speech, with spectral envelope transformation to improve intelligibility," *Accepted Interspeech 2012, Portland Oregon, USA*, 2012.

[10] J. Krause and L. Braida, "Investigating alternative forms of clear speech: The effects of speaking rate and speaking mode on intelligibility." *JASA*, vol. 112(5), pp. 2165–2172, 2004.

[11] ——, "Acoustic properties of naturally produced clear speech at normal speaking rates," *JASA*, vol. 115, no. 362-378, 2004.

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

[13] C. Valentini-Botinhao, J. Yamagishi, and S. King, "Can objective measures predict the intelligibility of modified hmm-based synthetic speech in noise?" *Interspeech 2011*, Florence, Italy, 2011.

[14] M. Cooke, C. Mayo, C. Valentini-Botinhao, Y. Stylianou, B. Sauert, and Y. Tang, "Evaluating the intelligibility benefit of speech modifications in known noise conditions," *Speech Communication*, 2013, http://dx.doi.org/10.1016/j.specom.2013.01.001.

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

[16] C. Valentini-Botinhao, E. Godoy, Y. Stylianou, B. Sauert, J. Yamagishi, and S. King, "Improving intelligibility in noise of hmm-generated speech via noise-dependent and -independent methods," *ICASSP 2011*, Vancouver, Canada, 2013.

[17] A. R. Bradlow, N. Kraus, and E. Hayes., "Speaking clearly for learning-impaired children: sentence perception in noise." *Journal of Speech, Language, and Hearing Research*, vol. 46, pp. 80–97, 2003.

[18] 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." *Journal of Speech and Hearing Research*, vol. 29, pp. 434–446, 1986.

[19] ——, "Speaking clearly for the hard of hearing iii: an attempt to determine the contribution of speaking rate to differences in intelligibility between clear and conversational speech." *Journal of Speech and Hearing Research*, vol. 32, pp. 600–603, 1989.

[20] V. Hazan and R. Baker, "Acoustic-phonetic characteristics of speech produced with communicative intent to counter adverse listening conditions," *Journal of the Acoustical Society of America*, vol. 130, no. 4, pp. 2139–2152, 2011.

[21] C. Mayo, V. Aubanel, and M. Cooke, "Effect of prosodic changes on speech intelligibility," in *Proc. Interspeech*, Portland, OR, USA, 2012.

[22] V. Hazan and A. Simpson, "Cue-enhancement strategies for natural vcv and sentence materials presented in noise," *Speech, Hearing and Language*, vol. 9, pp. 43–55, 1996.

[23] 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," *Proceedings of InSTIL/ICALL2004 NLP and Speech Technologies in Advanced Language Learning Systems*, Venice 17-19 June, 2004.

[24] M. Cooke, M. L. G. Lecumberri, O. Scharenborg, and W. A. van Dommelen, "Language-independent processing in speech perception: Identification of english intervocalic consonants by speakers of eight european languages," *Speech Comm.*, vol. 52, pp. 954–967, 2010.

[25] S. H. Ferguson and D. Kewley-Port., "Vowel intelligibility in clear and conversational speech for normal-hearing and hearing-impaired listeners." *Journal of the Acoustical Society of America*, vol. 112, pp. 259–271, 2002.

[26] 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.

[27] C. Davis and J. Kim, "Is speech produced in noise more distinct and/or consistent?" in *Speech Science and Technology*, 2012, pp. 46–49.

[28] A. Roebel and X. Rodet, "Efficient spectral envelope estimation and its application to pitch shifting and envelope preservation," in *Digital Audio Effects (DAFx)*, 2005, pp. 30–35.

[29] E. Godoy, O. Rosec, and T. Chonavel, "Voice conversion using dynamic frequency warping with amplitude scaling, for parallel or nonparallel corpora," *IEEE Trans Audio, Speech, Lang Processing*, vol. 20, no. 4, pp. 1313–1323, 2012.

[30] D. Kapilow, Y. Stylianou, and J. Schroeter, "Detection of non-stationarity in speech signals and its application to time-scaling," in *Eurospeech*, 1999, pp. 2307–2310.

[31] K. S. Rhebergen, N. J. Versfeld, and W. A. Dreschler, "Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise." *J. Acous. Soc. Am.*, vol. 120, no. 6, pp. 3988–3997, 2006.