On spectral and time domain energy reallocation for speech-in-noise intelligibility enhancement

Tudor-Cătălin Zorilă[1], *Yannis Stylianou*[1,2]

[1]Computer Science Department, University of Crete, Heraklion, Crete, Greece
[2]Toshiba Cambridge Research Lab, UK

ztudorc@gmail.com, yannis.stylianou@crl.toshiba.co.uk

10.21437/Interspeech.2014-466

Abstract

This paper addresses the problem of increasing speech-in-noise intelligibility under the constraint of energy preservation. Two recently proposed algorithms which have been shown to be very successful in this problem according to two large formal listening tests are reviewed and a hybrid system which combines the properties of the two methods is suggested. The first technique, which is a frequency domain approach, is re-implemented providing clarifications on its energy reallocation strategy. Based on objective measures well correlated with human perception, we show that our implementation performs similarly to the original approach. Moreover, this is combined with a dynamic range compression algorithm from the second method to allow reallocation of energy over time as well. Experiments with speech shaped noise (SSN) and competing speaker (CS) noise maskers at various SNRs indicate that the hybrid system outperforms the individual algorithms in terms of intelligibility scores.

Index Terms: speech-in-noise intelligibility enhancement, spectral contrast enhancement, dynamic range compression

1. Introduction

Speech communication has as goal to convey information. However, in most of the real communicative situations speech signals get corrupted by the acoustical background noise (ABN) found where the speaker and/or listener is located. Above certain levels of noise maskers, the intelligibility of speech received by listeners in such conditions may drop substantially, thus making the information transfer to become unreliable. Everyday voice-based services such as telephonic and news broadcasting (at train stations, airports etc.) systems are frequently operating in strong ABN environments while they are delivering speech messages from ordinary conversations to last-minute changes in flight schedules at airports. Moreover, there are cases where high intelligibility of speech in noise is a critical goal to achieve, i.e., the radio communication between a plane pilot and the air traffic controller. Although natural speech redundancy may partially compensate for some intelligibility losses, that is not always possible or acceptable. Thus, it is most useful to find signal processing techniques aimed at artificially boosting speech intelligibility. These also facilitate the advances of other speech-oriented technologies, such as speech synthesis, speech recognition or speech coding.

This paper addresses the problem of artificially increasing the intelligibility of clean speech received by listeners located in noisy environments (near-end listening enhancement [1]), under the constraint of equal signal's power before and after enhancement. In this context, studies of natural speech uttered by speakers in the presence of ABN have revealed a series of guidelines to follow when designing such algorithms. It is a known fact that human talkers adapt their speaking style in acoustically challenging environments [2]. Changes in speaking style may be as simple as repeating the message (often with a different prosody, i.e. slower speaking rate, more frequent pauses, emphasizing parts of speech etc.) or may involve substantial modifications of normal speech production (e.g., spectral tilt flattening, increased F0, hyper-articulation etc.) caused by an increased vocal effort as response to acoustical perturbations (the so-called Lombard effect [3]). Consequently, many algorithms are designed to mimic similar changes in speech production of talkers exposed to strong acoustical perturbations [4] and/or to take advantage of the means listeners perceive the mix of speech and noise in the same circumstances [5]. Furthermore, the studies concerning hearing-impaired listeners have contributed greatly to the understanding of human speech intelligibility internals [6, 7, 8].

Recently, two research challenges were organized asking participants to propose new methods to improve speech intelligibility in noise by using signal processing [4, 9]. They were provided with a full set of unmodified (plain) natural speech sentences and two noise maskers (SSN and CS) at various SNRs. The goal was to modify the plain speech such that its intelligibility over specified noise conditions improves, but keeping the same energy of the signal before and after modification. Large scale listening tests were organized to evaluate the intelligibility gains of each algorithm. Among the performers of these challenges, two algorithms achieved very good results in terms of intelligibility improvements over a wide range of noise conditions [4, 9]. The first method (denoted SSDRC) readjusts signal's energy both in frequency and time domains following the observations in clear and Lombard speech, but also in hearing-aid studies [10]. For this purpose, it uses two subsystems connected in a cascade form, one working in frequency domain (spectral shaping, denoted SS) and the other working in time domain (dynamic range compression, denoted DRC). The second algorithm uses exclusively spectral energy reallocation (SER) to improve the intelligibility of speech in noise, by combining spectral tilt flattening, spectral contrast enhancement (SCE) and low-frequency spectrum preservation [11]. This last technique is interchangeably referred to as (bs)SER (baseline SER) or SER.

Although both SSDRC and SER algorithms have reported very good results at respectively 2012 and 2013 Hurricane Challenges [4, 9], so far they have not been directly compared in terms of their intelligibility benefits. Furthermore, the original description of (bs)SER in [11] was found to contain several obscure algorithmic details (see Section 2) which makes a chal-

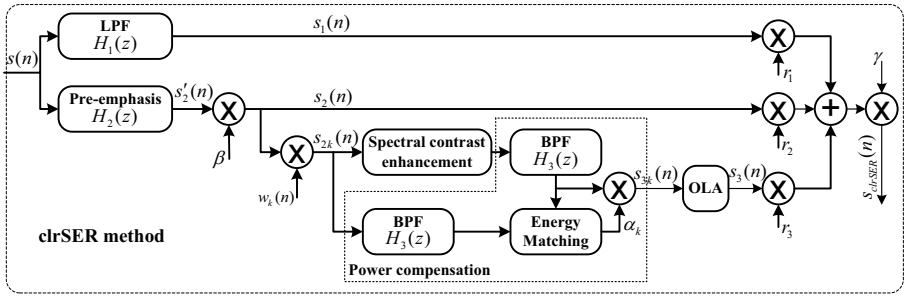

Figure 1: Block diagram of the (clr)SER method.

lenging task for someone to replicate the system and may cause significant perceptual quality degradation depending on how they are implemented. Thus, in this paper first we try to clarify these issues by providing a straightforward re-implementation of the baseline SER algorithm, re-implementation henceforth denoted as (clr)SER (clarified SER). Then, we suggest a further step in boosting speech intelligibility by extending SER with a secondary energy redistribution, not in frequency but in time domain. Hence, inspired by the work of Baer et al. [6] and by our own studies [10], we consider a hybrid SER followed by DRC system (SER+DRC). Finally, for the first time we present the results of direct comparison between intelligibility of SSDRC, SER, (clr)SER and SER+DRC modified speech in terms of objective scores well correlated with human perception. These experiments validate our implementation of SER algorithm and indicate that a further temporal energy re-adjustment may additionally boost speech intelligibility, findings which are confirmed by a series of informal listening tests.

The rest of the paper is structured as follows. Section 2 first describes (clr)SER, our re-implementation of (bs)SER method which gives additional clarifications of several important algorithmic details missing from the original description, then it briefly reviews the reference SSDRC algorithm and, finally, it introduces the SER+DRC system. Section 3 describes the evaluation procedure and presents the results, while the paper is concluded in Section 4.

2. Methods

2.1. Clarified Spectral Energy Reallocation method

In the context of 2013 Hurricane Challenge, Takou et al. have proposed an effective noise independent solution to increase speech in noise intelligibility [9, 11]. This method applies different signal transformations over time domain to redistribute spectral energy of speech sounds, and consists of three stages, i.e., low-frequency preservation, spectral tilt flattening and spectral contrast enhancement.

However, the original description of (bs)SER in [11] was found to be deficient of several algorithmic details of significant importance to a straightforward re-implementation. Among these obscure points could be mentioned: a) missing details regarding the discrete implementation of the pre- and post-filters of SCE; b) whether band-pass filtering is applied at the end of SCE and which are the specifications of that filter; c) whether power compensation is applied as an integrated stage of SCE on a frame-by-frame basis or is applied only once on the reconstructed overlap-added speech signal; d) missing indications if and when zero-phase filters were used for different stages of (bs)SER. Thus, in what follows we detail (clr)SER, a more clarified re-implementation of the system introduced by Takou et al., which was found to produce similar signal modifications with those of (bs)SER. The clarified SER method has the block diagram shown in Fig. 1. In this paper we assume sampled signals, with n being the discrete time index.

2.1.1. Low-frequency preservation

At this stage the low-frequency content below 400 Hz of the entire speech signal $s(n)$ is selected by a low-pass filter (LPF), denoted $H_1(z)$ in Fig. 1. $H_1(z)$ has the cutoff frequency of 400 Hz and it is designed as a 30-th order FIR filter using a Blackman window. In our implementation this filter is applied twice for the entire signal in a zero-phase configuration (once in the forward and once in the reverse time directions), resulting $s_1(n)$. Takou et al. have specified a similar FIR-based LPF, but of 60-th order and with no indication about zero-phasing. However, we consider the zero-phase filtering approach to be important at preventing audible artifacts caused by potential phase mismatchings from signal modifications in three different parallel stages whose outputs are summed.

2.1.2. Spectral tilt flattening

Inspired by the Lombard speech studies, the second stage in Fig. 1 is used to reduce signal's spectral tilt. Following the indications in [11], this is done using a pre-emphasis filter $H_2(z)$,

$$H_2(z) = 1 - \alpha z^{-1} \tag{1}$$

where $\alpha = 0.97$ at a sampling rate of 16 kHz. The pre-emphasis filter is applied over the entire signal, followed by a global energy adjustment to keep the same power of speech, before and after pre-emphasis. If $s_2'(n)$ is the pre-emphasized speech in Fig. 1, then the energy-adjusted signal $s_2(n)$ is computed as

$$s_2(n) = \beta s_2'(n), \;\; \beta = \sqrt{\frac{\sum_n s^2(n)}{\sum_n s_2'^2(n)}} \tag{2}$$

2.1.3. Spectral contrast enhancement

This is the last stage of SER and it is mostly implemented following the guidelines from the work of Turicchia and Sarpeshkar [12], as it is mentioned in [11]. Spectral contrast is defined as the difference in dB between adjacent peaks and valleys in the spectrum [12]. Thus, increasing spectral contrast also sharpens formants, which is known from studies with clear speech to improve intelligibility in noise. Moreover, the solution proposed by Turicchia and Sarpeshkar is known to keep a stable relation between the levels of the formants, thus providing high quality perceived speech.

SCE is done on a frame-by-frame basis and it has the architecture presented in Fig. 2. Pre-emphasized speech signal $s_2(n)$ is divided in 75% overlapping frames of length 32 ms using Hann windows, $w_k(n)$. Then, each frame $s_{2k}(n)$ is passed

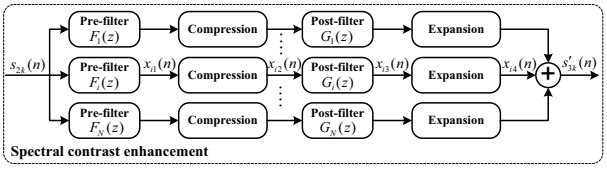

Figure 2: Spectral contrast enhancement architecture.

through a filter bank with $N = 50$ channels which applies SCE. Each channel 'i' contains a pre-filter $F_i(z)$, a compression block, a post-filter $G_i(z)$ and an expansion block.

The pre- and post-filters have the same resonant frequencies, logarithmically-spaced between 250 Hz and 4000 Hz, while their Laplace transfer functions are defined as

$$F_i(s) = F_i^{'4}(s) = \left(\frac{2\left(\frac{\tau_i}{q_1}\right) s}{\tau_i^2 s^2 + 2\left(\frac{\tau_i}{q_1}\right) s + 1} \right)^4 \quad (3)$$

$$G_i(s) = G_i^{'4}(s) = \left(\frac{2\left(\frac{\tau_i}{q_2}\right) s}{\tau_i^2 s^2 + 2\left(\frac{\tau_i}{q_2}\right) s + 1} \right)^4 \quad (4)$$

and they are implemented in two stages, as follows. First, the digital versions of $F_i'(s)$ and $G_i'(s)$ are computed by employing bilinear transformations with resonant frequency pre-warping, and then they are applied four times in a zero-phase configuration. These information are not specified in the work of Takou et al. [11] and they answer point (a) from the list with obscure points mentioned earlier.

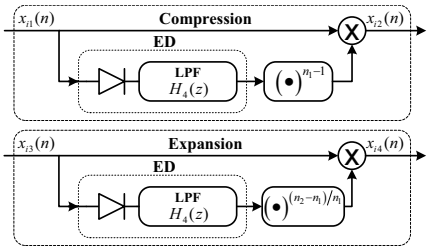

Figure 3: Details of companding strategy used for spectral contrast enhancement [12].

The compression and expansion algorithms of each channel are those suggested in [12], and they are depicted in Fig. 3. They both require an envelope detector (ED) which is implemented by using a full-wave rectifier followed by a first-order low-pass filter, referred to as $H_4(z)$ in Fig. 3. The poles of $H_4(z)$ are chosen to scale with the resonant frequency of each channel $f_i = 1/(2\pi\tau_i)$, such that

$$\tau_i^{ED} = w\tau_i \quad (5)$$

while this filter is applied twice, in a zero-phase configuration. For the rest of the parameters in equations (3)-(5) and Fig. 3, (clr)SER uses the same values of those mentioned in [11], respectively: $q_1 = 2$, $q_2 = 12$, $w = 40$, $n_1 = 0.3$ and $n_2 = 1$.

The spectral contrast enhanced speech frame $s'_{3k}(n)$ is obtained by summing the contributions of each channel in Fig. 2. Power compensation is then used to adjust the energy of current band-limited and contrast enhanced speech frame to that before enhancement. This last step is not part of the SCE method presented in [12], but it is an extension suggested by Takou et al. [11]. However, its description is incomplete and may lead to different interpretations (see points (b)-(d) discussed above). Thus, our experiments have shown that the power compensation approach depicted in Fig. 1 produces similar results to (bs)SER. First, $s'_{3k}(n)$ is band-pass filtered (BPF) by a 30-th order FIR filter whose cutoff frequencies are 250 and 4500 Hz, and which is designed using the Blackman window. This filter (denoted $H_3(z)$ in Fig. 1) is applied twice, in a zero-phase configuration. Next, the response of $H_3(z)$ to the enhanced speech frame $s'_{3k}(n)$ is scaled to match the energy of unmodified band-limited $s_{2k}(n)$. Thus, similarly to equation (2), the energy matching block computes a correction factor α_k per each frame, which is used to generate $s_{3k}(n)$. Finally, the energy compensated frames $s_{3k}(n)$ are overlap-added (OLA), resulting $s_3(n)$.

The clr(SER) method is concluded by weighting and adding the signals from the previous three stages, followed by a last scaling γ to keep the same average power of modified and original speech signals. The weights are those suggested in [11], respectively $r_1 = r_2 = r_3 = 1$.

2.2. Spectral Shaping and Dynamic Range Compression

In our previous work we have shown that combining frequency and time domain energy redistributions (by means of spectral shaping and dynamic range compression, respectively) is an effective solution to boost speech in noise intelligibility under the constraint of equal signal power before and after enhancement [10, 4]. SSDRC has two sub-systems. The following subsections briefly present them, while the details can be found in [10].

2.2.1. Spectral Shaping

Spectral shaping is the first sub-system of SSDRC and it is designed to rearrange signal's spectral energy following the observations in clear and Lombard speech studies, where formant sharpening and spectral tilt flattening were found to boost intelligibility in noise. To alleviate the artifacts caused by the processing of unvoiced speech segments, the former modifications are adapted to the probability of voicing on a frame-by-frame basis. Speech analysis and synthesis are done by means of DFT, inverse DFT and overlap-add [10].

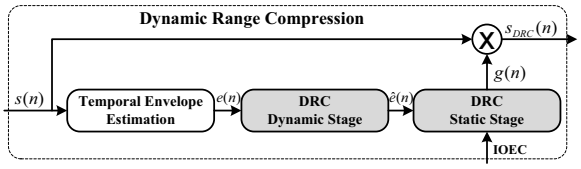

Figure 4: Block diagram of DRC algorithm.

2.2.2. Dynamic Range Compression

The second sub-system of SSDRC was inspired by the compression techniques in audio broadcasting and hearing-aid amplification [13]. It augments intelligibility in noise by redistributing energy over time domain such that signal's temporal envelope variations are reduced. Thus, low-energy segments of speech (e.g., nasals, onsets and offsets) are amplified, while more energetic voiced sounds are attenuated. This operation promotes intelligibility because low-energy segments of speech are more susceptible to noise masking [14]. A simplified DRC block diagram is shown in Fig. 4. Speech samples are re-scaled by means of time-varying gains $g(n)$ which are computed from

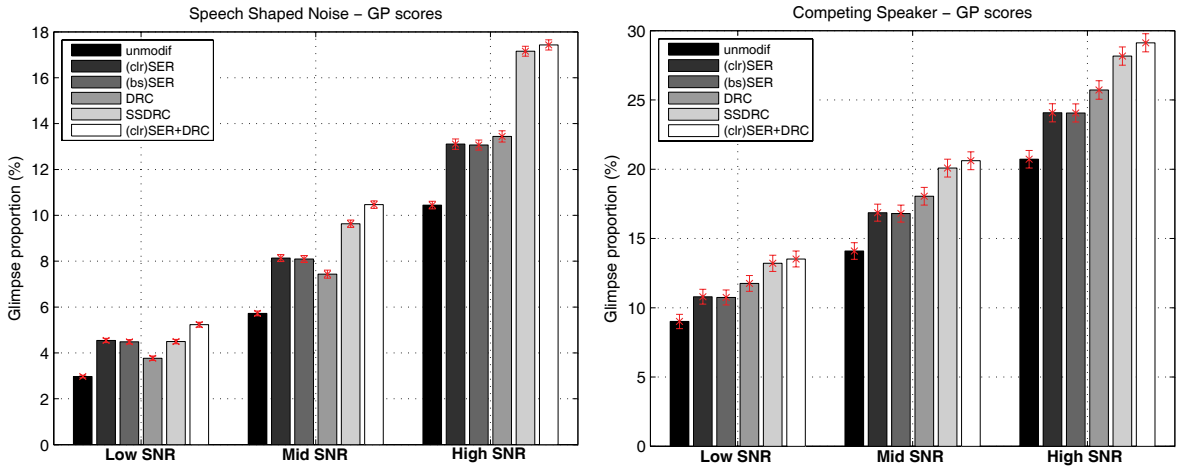

Figure 5: Results in terms of GP scores for all considered methods and noise conditions. 95% confidence intervals are also provided.

the dynamically and statically compressed temporal envelope. First, the temporal envelope is estimated by taking the magnitude of analytical signal corresponding to speech waveform. Fast envelope fluctuations are reduced by dividing the previous estimated signal into non-overlapping segments of 2.5 times speaker's mean pith period, then 95% of the maximum value in each frame is saved to generate $e(n)$. An average 150 Hz pitch frequency was used in our experiments. Next, $e(n)$ is dynamically compressed with 2 ms release and an instantaneous attack time constants [10]. During the static stage, 30% of the maximum value of dynamically compressed envelope $\hat{e}(n)$ is used as reference level to convert this signal in dB, then a predefined input/output envelope characteristic (IOEC) is applied [10]. The gains resulted from the previous operations are used to re-scale the speech samples. At the end, a global power correction is applied to keep the same energy of the signal before and after DRC.

2.3. SER followed by DRC

Inspired by the work of Baer et al. [6] regarding SCE and time domain compression for listeners with hearing impairments, but also by our own experiments with SSDRC [10], in this paper we suggest merging SER with the DRC stage from SSDRC (described earlier). Hence, the hybrid SER followed by DRC algorithm reallocates signal's energy both over frequency and time domains, so a higher intelligibility for the enhanced speech in noise is expected, against individual SER and DRC techniques.

3. Evaluation & Results

The speech and noise data employed in our experiments were provided by the organizers of Hurricane Challenges and they are the same as those used to report results in [4, 9]. In our experiments, speech data consisted of the first 100 Harvard sentences, while SSN and CS noise maskers were applied at low, mid and high SNRs, respectively as follows: -9dB, -4dB and 1dB for SSN, and -21dB, -14dB and -7dB for CS. The first 100 baseline SER modified Harvard sentences were used as reference. All signals have 16 kHz sampling frequency and 16 bits quantization.

Speech in noise intelligibility was objectively predicted using a metric well correlated with human perception, namely the glimpse proportion (GP) [15]. The GP score gives a percentage of spectro-temporal regions of excitation pattern in which the local SNR exceeds a given threshold (set here to 3 dB). Bigger GP scores indicate more intelligible speech signals. Thus, the following speech styles were evaluated: plain (unmodified), (clr)SER, (bs)SER, DRC, SSDRC and SER+DRC modified speech.

The results in terms of GP scores are shown in Fig. 5. They indicate that SER+DRC outperforms in terms of GP scores all other methods considered in the present experiment and over all noise conditions. In the context of speech in noise intelligibility enhancement, these findings strengthen the importance of additional energy redistribution over time domain by means of DRC. Furthermore, the same results in Fig. 5 suggest that SSDRC-modified speech is considerable more intelligible than SER-modified speech in 5 out of 6 noise conditions, while (clr)SER and (bs)SER perfom similarly in terms of GP points. The later assertion was confirmed by informal listening tests and parallel visual inspection of short-time spectra of (clr)SER and (bs)SER modified signals. The informal listening tests in a noise-free scenario indicated that (clr)SER and (bs)SER enhanced speech sounds are perceptually very challenging to separate. In support of these claims, several speech samples are provided at the web-page mentioned in [16].

4. Conclusions

In this paper we have suggested a hybrid algorithm to improve speech in noise intelligibility by combining the properties of two recently proposed methods from the literature. This approach was found to outperform the individual methods in terms of objective scores designed to predict speech in noise intelligibility. In the same context, this paper presents the results of the first objective comparison between the two reference algorithms. Furthermore, we have re-implemented the spectral energy reallocation method providing helpful clarifications on some of its signal processing. As future work, we plan to organize a formal listening test to confirm the results presented in this paper. The findings will be reported at the conference.

5. Acknowledgments

Authors would like to thank the organizers of Hurricane Challenge for providing the data for evaluation.

6. References

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

[2] V. Hazan and R. Baker, "Acoustic-phonetic characteristics of speech produced with communicative intent to counter adverse listening conditions," *J. Acoust. Soc. Am.*, vol. 130, pp. 2139–2152, 2011.

[3] J.-C. Junqua, "The Lombard reflex and its role on human listeners," *J. Acoust. Soc. Am.*, vol. 93, pp. 510–524, 1993.

[4] 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*, no. 55, pp. 572–585, 2013.

[5] S. Ferguson and D. Kewley-Port, "Vowel intelligibility in clear and conversational speech for normal-hearing and hearing-impaired listeners," *J. Acoust. Soc. Am.*, vol. 112, no. 1, pp. 259–271, 2002.

[6] T. Baer, B. Moore, and S. Gatehouse, "Spectral contrast enhancement of speech in noise for listeners with sensorineural hearing impairment: effects on intelligibility, quality, and response times," *J. Rehab. Res. Dev.*, vol. 30, no. 1, pp. 49–72, 1993.

[7] T. Ching, H. Dillon, and D. Byrne, "Speech recognition of hearing-impaired listeners: Predictions from audibility and the limited role of high-frequency amplification," *J. Acoust. Soc. Am.*, vol. 103, no. 2, pp. 1128–1140, 1998.

[8] A. Bhattacharya, A. Vandali, and F.-G. Zeng, "Combined spectral and temporal enhancement to improve cochlear-implant speech perception," *J. Acoust. Soc. Am.*, vol. 130, no. 5, pp. 2951–2960, 2011.

[9] M. Cooke, C. Mayo, and C. Valentini-Botinhao, "Intelligibility-enhancing speech modifications: the Hurricane Challenge," in *Proc. Interspeech*, 2013, pp. 3552–3556.

[10] T. Zorilă, V. Kandia, and Y. Stylianou, "Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression," in *Proc. Interspeech*, 2012, pp. 635–638.

[11] R. Takou, N. Seiyama, and A. Imai, "Improvement of speech intelligibility by reallocation of spectral energy," in *Proc. Interspeech*, 2013, pp. 3605–3607.

[12] L. Turicchia and R. Sarpeshkar, "A bio-inspired companding strategy for spectral enhancement," *IEEE Trans. Speech and Audio Process.*, vol. 13, no. 2, pp. 243–253, 2005.

[13] B. Blesser, "Audio dynamic range compression for minimum perceived distortion," *IEEE Trans. on Audio and Electroacoustics*, vol. 17, no. 1, pp. 22–32, 1969.

[14] J. Allen, *Articulation and intelligibility*. Morgan & Claypool Publishers, 2005.

[15] M. Cooke, "A glimpsing model of speech perception in noise," *J. Acoust. Soc. Am.*, vol. 119, no. 3, pp. 1562–1573, 2006.

[16] http://sites.google.com/site/tczorila/interspeech2014.