INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Modulation Enhancement of Temporal Envelopes for Increasing Speech Intelligibility in Noise

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, speech intelligibility is enhanced by manipulating the modulation spectrum of the signal. First, the signal is decomposed into Amplitude Modulation (AM) and Frequency Modulation (FM) components using a high resolution adaptive quasi-harmonic model of speech. Then, the AM part of midrange frequencies of speech spectrum is modified by applying a transforming function which follows the characteristics of the clear style of speaking. This results in increasing the modulation depth of the temporal envelopes of casual speech as in clear speech. The modified AM components of speech are then combined with the original FM parts to synthesize the final processed signal. Subjective listening tests evaluating the intelligibility of speech in noise showed that the suggested approach increases the intelligibility of speech by 40% on average, while it is comparable with recently suggested state-of-the-art algorithms of intelligibility boosters.

Index terms: Intelligibility, Clear speech, Casual speech, Temporal envelopes, Modulations, Sinusoidal Modeling

1. Introduction

Clear speech has been proven to be more intelligible than casual speech in various adverse listening conditions and for various listening populations (hearing-impaired listeners [1, 2], cochlear implant users [1], non-native listeners [3]). Therefore, enhancing the intelligibility of casual speech based on the acoustic properties of clear speech is quite challenging and promising. Previous acoustic analysis between the two speaking styles revealed lower speaking rate, expanded vowel space, lower spectral tilt and higher modulation depth of the temporal envelopes [4, 5] for clear speech. Among these features, there is a growing evidence that there is a significant contribution from the temporal envelope modulations to the intelligibility advantage of clear speech. This is suggested by neurophysiological and psycho-acoustical studies that link the perception of sounds with modulations [6] and is supported by the study of Drullman et al. [7] which showed an intelligibility degradation of speech after smearing the envelope of the low-frequency modulations (4-16Hz).

Based on the outcome of [7], modulation processing has been introduced to separate speech from noise and therefore, enhance the intelligibility of speech in noisy environments. The denoising algorithms are based on preserving the low-frequency modulations (4-16Hz) of the spectral envelope which are important for intelligibility while discarding other modulations imposed by the masker [8–11].

Rather than denoising the speech signal, other studies focus on re-enforcing the temporal envelope modulations of speech before it is presented in noise, as naturally happens in clear speech. In [12] modulation spectral components between 1-16 Hz are enhanced prior to distortion of speech in reverberant environment. The drawback of the designed modulation filters is that their efficiency depends on the reverberation condition. In [13], the temporal envelope of casual speech is transformed to have higher modulation depth as in clear speech in low modulation frequencies (less than 4Hz) which are considered to be important for phoneme identification. However, this modification technique decreased speech intelligibility due to processing artifacts. Therefore, increasing the modulation depth of the temporal envelopes while enhancing speech intelligibility has not yet been efficiently addressed.

This work explores a novel method for increasing the modulation depth of the temporal envelopes of speech, aiming at enhancing its intelligibility. Specifically, a Sinusoidal model is used to decompose speech into time-varying amplitudes, frequencies and phases. Each time-varying amplitude is considered to be the "temporal envelope" of the signal in the corresponding frequency. Then, a transforming function is applied to the time-varying amplitudes to increase their modulation depth. Finally, the speech signal is synthesized using the modified amplitudes and the initial frequencies and phases.

The proposed modification algorithm has two main advantages over state-of-the-art approaches. First, the analysis and synthesis of the speech signal is performed by the extended adaptive Quasi-Harmonic Model (eaQHM) [14, 15] which can decompose and reconstruct the signal with high accuracy. This alleviates the necessity of performing filterbank analysis and synthesis and temporal envelope estimation using the Hilbert transform. Such techniques reduce the effectiveness of the modulation filters [16] or introduce artifacts to the signal detrimental to its intelligibility. Second, the proposed method does not require the design of modulation filters whose efficiency may depend to the type of noise [12]. Instead, a simple transforming function is used to change the modulation depth of the time-varying amplitudes. The transforming function is designed based on clear speech properties and results in changes to the temporal envelope modulations of speech similar to that of clear speech. Last, compared to other intelligibility boosters, the proposed method emphasizes the harmonic structure of the speech signal not only in perceptually important frequency bands but all over the spectrum, suggesting an intelligibility benefit for various noise maskers.

2. Methodology

2.1. Decomposition and reconstruction of speech

Generally, the speech signal, containing both voiced and unvoiced segments, can be described as an AM-FM decomposition:

$$x(t) = \sum_{k=-K}^{K} \alpha_k(t) e^{j\Phi_k(t)} \quad (1)$$

http://dx.doi.org/10.21437/Interspeech.2016-500

where $\alpha_k(t)$, $\Phi_k(t)$ are the instantaneous amplitude and the instantaneous phase of the k^{th} component, respectively and K is the number of components that depends on the fundamental frequency and the Nyquist frequency of the speech signal. A means to accurately compute these parameters is the full-band extended adaptive Quasi-Harmonic Model [14, 15] which has been successfully used for analysis and synthesis of speech:

$$\hat{x}(t) = \sum_{k=-K}^{K} (\hat{\alpha}_k + t\hat{b}_k)\hat{A}_k(t)e^{j\hat{\phi}_k(t)} \quad (2)$$

In this model, $\hat{\alpha}_k$ is the complex amplitude and $\hat{b}_k$ is the complex slope of the k^{th} component, and $\hat{A}_k(t), \hat{\phi}_k(t)$ are functions of the instantaneous amplitude and phase of the k^{th} component, respectively [15]. These estimates are iteratively updated via Least Squares until a convergence criterion is met, which is related to the overall Signal-to-Reconstruction-Error Ratio (SRER) [14]. Then, the overall signal is synthesized using Eq.(1) where the estimated phases of $\Phi_k(t)$ are formed by a frequency integration scheme using the estimated phases $\hat{\phi}_k(t)$ [17] and $\alpha_k(t)$ is simply $|\hat{\alpha}_k(t)|$ via linear interpolation.

2.2. Transforming function

The accurate time-varying extraction of the temporal envelopes $\alpha_k(t)$ of clear and casual speech by eaQHM revealed two major findings. First, the very low-energy parts of the temporal envelopes of clear speech are more enhanced compared to that of casual speech. Second, clear speech has increased modulation depth of the temporal envelopes compared to casual speech in midrange frequencies, while in lower frequency regions clear speech has decreased modulation depth compared to casual speech. Fig. 1 and Fig. 2 summarize these two findings. Specifically, Fig. 1 depicts the temporal envelopes of clear and casual speech around 3000 Hz (15^{th} harmonic). As we can see, the low-energy information is more enhanced in clear speech than in casual speech (0-0.8s). Fig. 2 depicts the mean modulation depth, $\overline{D}(t)$, of the temporal envelopes on three acoustic frequency regions for clear and casual speech. The mean modulation depth, $\overline{D}(t)$, is estimated as follows: the time-varying amplitudes that correspond to the three acoustic frequency regions depicted in Fig. 2 are summed to derive the temporal envelope for each frequency region. For example, for the frequency region [200, 600] the first three amplitude components are summed in order to estimate the temporal envelope. Then, the temporal envelope $\hat{p}_i(t)$ can be described by Eq.(1) as an AM-FM signal with modulation amplitudes $d_n(t)$ and modulation phases $\Psi_n(t)$:

$$\hat{p}_i(t) = \sum_{n=-N}^{N} d_n(t)e^{j\Psi_n(t)} \quad (3)$$

EaQHM can estimate the modulation amplitudes $\hat{d}_n$ and phases $\hat{\Psi}_n$ using Eq.(2) where $\hat{x}(t) = \hat{p}_i(t)$ and N is the number of the modulation frequency components. The modulation depth $D(t)$ is the sum of the modulation amplitudes $\hat{d}_n$ in the modulation frequencies 2-8 Hz, with 1Hz resolution, and is also time-varying (Eq.(4)) . The average of $D(t)$ in time, namely $\overline{D}(t)$ is then depicted in Fig. 2.

$$D(t) = \sum_{n=-8, n\neq 0,1}^{8} |\hat{d}_n(t)| \quad (4)$$

Fig. 2 illustrates that clear speech (Clear) has higher mean modulation depth than casual speech (Casual) on midrange frequencies (800-3000Hz, 4^{th}-15^{th} component) while on low midrange frequencies (200-600Hz, 1^{st}-3^{rd} component) clear speech has lower mean modulation depth than casual speech.

Therefore, we propose a transforming function that modifies casual speech towards these two clear speech characteristics; the enhanced modulation depth in midrange frequencies and the increased energy in the low-energy parts of the temporal envelopes. The transforming function is a compression function applied in each component besides the first three components as suggested by our analysis (Fig. 2):

$$m_k(t) = \hat{\alpha}_k(t)^{\gamma}, \;\; |k| = 4, .., K \quad (5)$$

where $\frac{1}{2} \leq \gamma < 1$. After modifying the time-varying amplitudes of casual speech using Eq.(5), the signal is synthesized using Eq.(1) where $\alpha_k(t) = m_k(t)$. Then, the synthesized signal is normalized to have the same Root Mean Square energy (RMS) as the original unmodified signal. The transforming function with $\gamma = 0.5$ (DMod) follows the desired clear speech characteristics; it effectively enhances the low-energy parts of speech (Fig. 1) and increases the modulation depth of casual speech significantly above the midrange frequencies while decreases its modulation depth on low midrange frequencies (Fig. 2). Trials and errors along with informal listening tests on values of γ has shown that the intelligibility of casual speech significantly increases around the area of $\gamma = 0.5$. For higher values of γ the modulation depth of the modified time-varying amplitudes is less than that for lower values and the modified signal is acoustically closer to the original signal. For higher SNR levels a higher value of 0.5 for γ can be selected. Therefore, the parameter γ can be proportional to the noise level. However, very low values of γ should be avoided since they create signal distortions that may affect speech intelligibility. Fig. 3 presents the spectrogram of the original signal and the modified signal using the transforming function with $\gamma = 0.5$. It is worth noticing how the harmonic structure is emphasized.

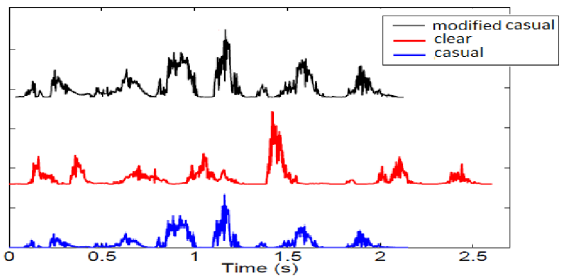

Figure 1: *Time-varying amplitude of 15 quasi-harmonic (around 3000Hz) estimated by eaQHM for the same sentence uttered in clear and casual style and for the modified casual signal by DMod with* $\gamma = 0.5$

3. Subjective Evaluation

For comparison purposes, our proposed algorithm is compared with two other intelligibility enhancing techniques, the Spectral Shaping and Dynamic Range Compression (SSDRC) [18] and the Mix-filtering [19]. SSDRC has been proven to be the most successful modification from a challenge task, containing extensive evaluation of various intelligibility enhancement techniques [20]. SSDRC performs energy reallocation in spectral and time domain to increase the intelligibility of speech in

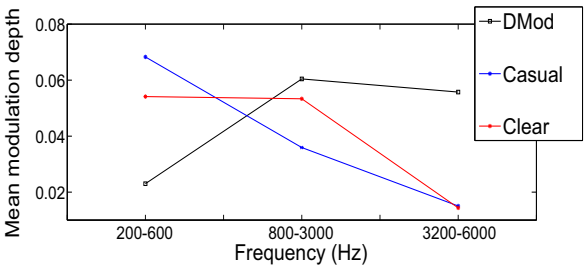

Figure 2: *Mean modulation depth of low frequency modulations (2-8Hz) estimated by eaQHM on the temporal envelopes of {Clear, Casual, DMod ($\gamma = 0.5$)} on three acoustic frequency regions for the same sentence.*

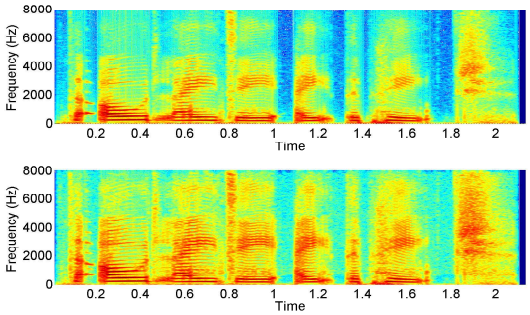

Figure 3: *Spectrogram of the casual (upper panel) and the modified casual signal (lower panel) using the transforming function with $\gamma = 0.5$.*

noise. The SSDRC modified casual speech has been shown to reach the intelligibility levels of clear speech for 0dB and 5dB SNR Speech Shaped noise (SSN), and exceed them for -3dB [21], thus SSDRC serves as an upper bound for intelligibility. The Mix-filtering proposed in [19] implements spectral energy reallocation based on clear and casual speech differences. It also increases the intelligibility of the original speech while preserves its quality. Four sets of signals are evaluated: (1) original speech (OR), (2) the proposed modified speech using modulation-depth enhancement (DMod) (3) the SSDRC modified speech (SSDRC) (4) the mix-filtering modified speech (MixF). The term Categories will be used to refer to the 4 sets of signals, {OR, DMod, SSDRC, MixF}.

The database for evaluating the proposed modification contains sentences with one keyword inside the carrier sentence uttered in Greek "Lége [léksi klidí] padú" ("Say [keyword] everywhere") [22–25]. The keyword is a CVCV word. This database is used to test word intelligibility by native (Greek) listeners. The sentences are presented in Speech Shaped Noise (SSN) of low ($\text{SNR}_1 = -8$dB) and mid ($\text{SNR}_2 = -2$dB) levels of SNR. The listeners are asked to write down the keyword that they hear. Each keyword is presented only once to the listeners. 16 distinct sentences are presented to the listeners for evaluation uttered by a female and a male speaker (8 sentences per speaker (8x2 = 16 sentences), 4 sentences per speaker per noise level (4x2x2 = 16 sentences), 2 sentences per Category per noise level (2x4x2 = 16 sentences)). All sentences are normalized to have the same RMS energy and then noise is added to the sentences. First, the low SNR sentences are presented to the listeners and next the sentences on higher SNR. 4 sentences are presented as a header to the listeners to adjust their hearing to the noise level (20 sentences in total). The "header sentences" are not evaluated. The scoring system is based on previous research on English intelligibility tests [26–28], supported also by researchers for Greek language [29]. Each word is considered incorrect even if there is a mismatch in one phoneme e.g "fíki" instead of "thíki" ("seaweed" instead of "case"). However, incorrect person of verb and number of noun is considered half-correct e.g. "dóra" instead of "dóro" ("gifts" instead of "gift").

As word difficulty may affect the intelligibility scores of the algorithms, 4 different listening scenarios have been created to ensure that all algorithms will be evaluated on the same words. For example, if a specific word is presented to the listener in OR manner in SNR_1 on the listening Scenario 1, then the same word will be presented to another listener in SSDRC manner in the same SNR condition on listening Scenario 2 etc. This allows us to "denoise" the performance evaluation of our algorithms from the word dependency. For each listener, a listening scenario is randomly selected when he/she starts the test.

60 listeners (15 listeners per scenario), all native Greek speakers, participated in the intelligibility test. Performance evaluation contains two 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 native population. The second part of analysis computes the intelligibility scores of each Category across sentences, to parcel out the possible variability due to word difficulty.

For each SSN condition, the score of all correct and half correct keywords to the score of the total keywords is estimated per listener and per Category. Fig. 4 shows the {min, 1st quartile, median, 3rd quartile, max} of intelligibility scores per Category across all listeners. As it is expected, there is a high variability across listeners due to different word difficulty and different perception of sounds in noise by listeners [30]. Mean values are also depicted (rhombus symbol). SSDRC appears to have a higher intelligibility advantage over all Categories for both SSN conditions. Our proposed modification DMod has higher intelligibility score, that is 36% and 48% (mean values) vs unmodified speech (0% and 18%) for SSN_1 and SSN_2 respectively, approaching the intelligibility benefit of SSDRC (44% and 62% respectively). MixF achieves lower intelligibility scores than DMod (25% for SSN_1 and 35% for SSN_2).

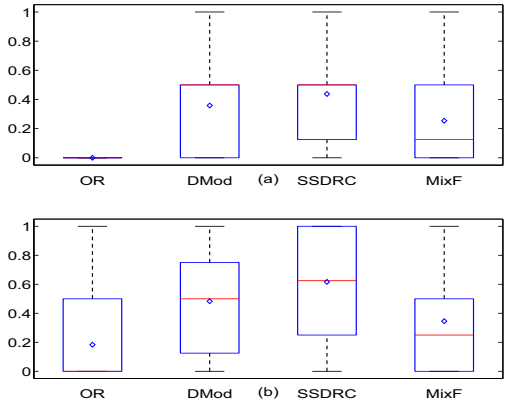

Figure 4: *Intelligibility score across listeners per Category (a) SNR=-8dB (b) SNR=-2dB*

In order to evaluate the statistical significance of these results, a repeated-measures ANOVA is performed on intelligibility with Category nested within each listener. Results re-

veal significant intelligibility differences among Categories, for both SSN conditions SSN_1 ($F(3;59) = 35.91$; $p < 10^{-15}$) and SSN_2 ($F(3;59) = 15.46$; $p < 10^{-8}$). Post-hoc comparisons using pairwise paired t-tests with Holm adjustment reveal that the mean intelligibility scores of DMod ($M = 0.36$ (mean); $SD = 0.26$ (standard deviation)), SSDRC ($M = 0.44$; $SD = 0.32$) and MixF ($M = 0.25$; $SD = 0.28$) are significantly different ($p_{DMod} < 10^{-14}$, $p_{SSDRC} < 10^{-13}$, $p_{MixF} < 10^{-8}$) from OR ($M = 0$; $SD = 0$) in SSN_1. For SSN_2, both DMod ($M = 0.48$; $SD = 0.37$) and SSDRC ($M = 0.62$; $SD = 0.38$) have significantly different means ($p < 10^{-4}$) from OR ($M = 0.18$; $SD = 0.25$) while no statistical significance ($p = 0.028$) was found between the mean of MixF ($M = 0.35, SD = 0.39$) and OR ($M = 0.18$; $SD = 0.25$). No statistical significant difference has been found between MixF, DMod and SSDRC on SSN_1. On SSN_2 condition there is no statistical significant difference between SSDRC and DMod.

In order to investigate possible dependencies of the intelligibility scores on word difficulty, intelligibility scores for each word is computed for all Categories. Fig. 5 shows the normalized scores for each word for the two SSN conditions, SSN_1 (Fig. 5a) and SSN_2 (Fig. 5b). As we can see word difficulty influences the efficiency of the modification algorithms. This variability on the intelligibility scores due to word difficulty justifies the high variance on the intelligibility scores for the modification techniques reported in Fig. 4. In easily predicted words like "zoúla, laliá" the intelligibility scores of {SSDRC, DMod, MixF} are higher than that of more difficult predicted words like "theté, goní". It is worth noticing that original speech, OR, has also higher intelligibility score in easily predicted words for the SSN_2 condition. Finally, 9 out of 16 words have the highest intelligibility scores when modified by SSDRC, while 4 out of 16 words have the highest intelligibility score when modified by DMod.

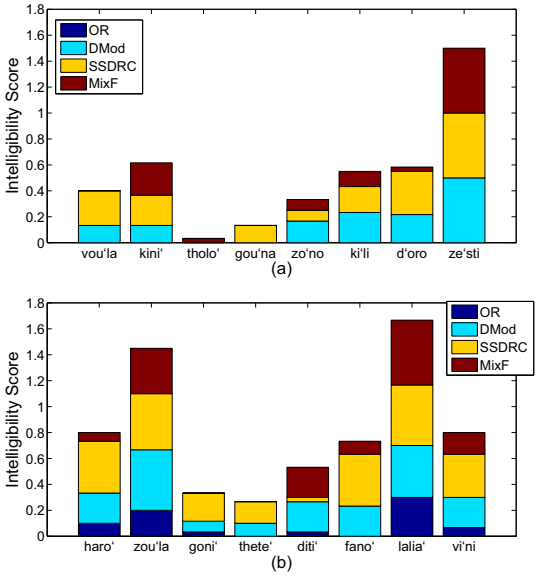

Figure 5: *Intelligibility score of each word per Category (a) SNR=-8dB (b) SNR=-2dB*

4. Discussion and Future Work

Subjective evaluations report that the proposed modification method, DMod, increases speech intelligibility in SSN noise. This intelligibility improvement is based on acoustic differences between clear and casual speech. The transforming function increases the modulation depth of low modulation frequencies (2-8Hz) of the temporal envelopes, as it naturally happens on clear speech. However, unlike other studies the intensity envelope is not extracted using filterbank analysis on frequency bands. On the other hand, the time-varying amplitudes of the quasi-harmonics are extracted using an AM-FM decomposition algorithm. This alleviates possible distortions of the envelope during the carrier and envelope extraction process [10,16] since the eaQHM model, used for analysis-synthesis, is highly adaptive to the signal parameters (amplitudes, frequencies and phases).

Furthermore, DMod decomposes speech into very narrow frequency bands (quasi-harmonics) effectively achieving to enhance the quasi-harmonic structure of speech all over the spectrum. Indeed, as Fig. 6 shows, SSDRC enhances the spectral content of perceptually important frequency regions (1000-3000Hz) which are masked in SSN noise. On the other hand, DMod emphasizes the harmonic structure of speech all over the spectrum. This suggests the intelligibility benefits of DMod can be extended to other types of noise maskers. Future directions aim on testing the proposed modification algorithm on other types of noise (babble noise, reverberation etc.) and for different parameters of γ depending on the noise level, as the evaluation results indicate (Fig. 4). It would be also intererestting to explore combinations of DMod with other intelligibility boosters. Last, we expect that hearing-impaired people will also benefit from DMod by performing modulation enhancement on specific frequency bands to balance the loss of the compressive nonlinearity in the basilar membrane [31].

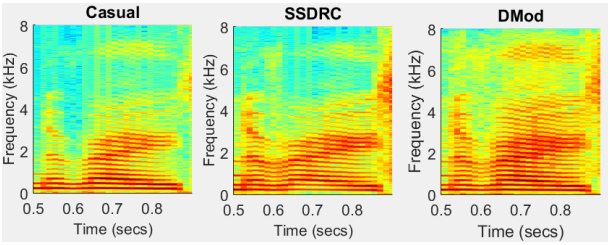

Figure 6: *The emphasized harmonic structure of modified speech by DMod vs. SSDRC and unmodified casual speech*

5. Conclusions

This work proposes a novel method for enhancing speech intelligibility by increasing the modulation depth of the temporal envelopes of speech. The instantaneous amplitudes are extracted from a Sinusoidal model (eaQHM) and are modified using a transforming function that approaches clear speech characteristics. Then, the signal is reconstructed by the eaQHM model using the modified amplitudes and the unmodified instantaneous frequencies and phases. Intelligibility tests from native listeners in low (-8dB) and mid (-2dB) SNR Speech Shaped Noise have shown significant intelligibility improvement of speech using the proposed method, rendering the combination of this method with other spectral transformation techniques promising for further intelligibility enhancement of speech.

6. Acknowledgments

The authors would like to thank Prof. Katerina Nicolaidis and Dr. Anna Sfakianaki for providing the database: "SpeakGreek: Developing a biofeedback speech training tool for Greek segmental and suprasegmental features: Application in L2 learning/teaching and clinical intervention", co-financed by the European Union (ESF) and Greek national funds (ARISTEIA II).

7. References

[1] 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*, vol. 29, pp. 434–446, 1986.

[2] 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*, vol. 39, pp. 494–509, 1996.

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

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

[5] S. Liu, E.D. Rio, A.R. Bradlow, and F.G. Zeng, "Clear speech perception in acoustic and electric hearing," *J. Acoust. Soc. Amer.*, vol. 116, no. 4, pp. 2374–2383, 2004.

[6] S.P. Bacon and D.W. Grantham, "Modulation masking: Effects of modulation frequency, depth and phase," *J. Acoust. Soc. Amer.*, vol. 85, pp. 2575–2580, 1989.

[7] R. Drullman, J.M. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," *J. Acoust. Soc. Amer.*, vol. 95, no. 2, pp. 1053–1064, 1994.

[8] K. Wójcicki and P.C. Loizou, "Channel selection in the modulation domain for improved intelligibility in noise," *Journal of the Acoustical Society of America*, vol. 131 (4), pp. 2904–2913, April 2012.

[9] K. Wójcicki K. Paliwal and B. Schwerin, "Single-channel speech enhancement using spectral subtraction in the short-time modulation domain," *Speech Communication*, vol. 52 (5), pp. 450–475, May 2010.

[10] J. H. Won, S. M. Schimmel, W. R. Drennan, P. E. Souza, L. Atlas, and J. T. Rubinstein, "Improving performance in noise for hearing aids and cochlear implants using coherent modulation filtering," *Hearing Research*, vol. 239, pp. 1–11, May 2008.

[11] N. Mesgarani and S.Shamma, "Speech enhancement based on filtering the spectrotemporal modulations," vol. 1, pp. 1105–1108, March 2005.

[12] A. Kusumoto, T. Kinoshita, K. Hodoshima, and N. Vaughan, "Modulation enhancement of speech by a pre-processing algorithm for improving intelligibility in reverberant environments," *Speech Comm.*, vol. 45, pp. 101–113, 2005.

[13] 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.*, vol. 125, no. 5, pp. 3346–3357, 2009.

[14] G.P. Kafentzis, O. Rosec, and Y. Stylianou, "Robust full-band adaptive sinusoidal analysis and synthesis of speech," *ICASSP*, pp. 6260–6264, 4-9 May 2014.

[15] G.P. Kafentzis, Y. Pantazis, O. Rosec, and Y. Stylianou, "An extension of the adaptive quasi-harmonic model," *ICASSP*, pp. 4605–4608, 25-30 March 2012.

[16] L. Atlas and C. Janssen, "Coherent modulation spectral filtering for single-channel music source separation," *ICASSP*, vol. 4, pp. 461–464, 18-23 March 2005.

[17] Y. Pantazis, O. Rosec, and Y. Stylianou, "Adaptive AM-FM signal decomposition with application to speech analysis," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 19, pp. 290–300, 2011.

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

[19] 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.

[20] 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*, 55(4), 2013, pp. 572-585.

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

[22] *http://speakgreek.web.auth.gr/wp/research-goals/.*

[23] K. Nicolaidis, G. Papanikolaou, K. Avdelidis, E. Kainada, A. Sfakianaki, L. Vrisis, K. Konstantoudakis, I. Starchenko, and E. Kelmali, "Speakgreek: Development of an online speech training system," *Proceedings of the 7th Panhellenic Conference "Acoustics 2014"*, October, 20-21, 2014 (in print).

[24] K. Nicolaidis, G. Papanikolaou, E Kainada, and K. Avdelidis, "Speakgreek: An online speech training tool for l2 pedagogy and clinical intervention," *Accepted at the 18th International Congress of Phonetic Sciences*, 2015.

[25] K. Nicolaidis, G. Papanikolaou, A. Sfakianaki, E. Kainada, K. Avdelidis, and K. Konstantoudakis, "Computer assisted teaching of vowel production to learners of greek as an l2 and individuals with speech disorders," *22nd International Symposium on Theoretical and Applied Linguistics*, 2015, Aristotle University of Thessaloniki.

[26] R.B. Monsen, "Toward measuring how well hearing-impaired children speak," *Journal of Speech, Language and Hearing Research*, vol. 21(2), pp. 197–219, 1978.

[27] R.B. Monsen, "The oral speech intelligibility of hearing-impaired talkers," *Journal of Speech, Language and Hearing Research*, vol. 48(3), pp. 286–296, August 1982.

[28] M.A. Picheny, N.I Durlach, and L.D. Braida, "Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech," *Journal of Speech, Language and Hearing Research*, vol. 28(1), pp. 96–103, March 1985.

[29] A. M. Sfakianaki, "An acoustic study of coarticulation in the speech of greek adults with normal hearing and hearing impairment," *PhD thesis in Linguistics, Aristotle University of Thessaloniki*, 2012.

[30] M. Cooke and M.L.G. Lecumberri, "The intelligibility of lombard speech for non-native listeners," *J. Acoust. Soc. Amer., Letters to the Editor*, vol. 132, pp. 1120–1129, 2012.

[31] B.C.J. Moore and A.J. Oxenham, "Psychoacoustic consequences of compression in the peripheral auditory-system," *Psychological review*, 105(1), 1998, pp. 108-124.