INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Real-Time Modulation Enhancement of Temporal Envelopes for Increasing Speech Intelligibility

Maria Koutsogiannaki, Holly Francois, Kihyun Choo, Eunmi Oh

Samsung Electronics

mkoutsog@csd.uoc.gr, h.francois@samsung.com, khchoo@samsung.com, sait@samsung.com

Abstract

In this paper, a novel approach is introduced for performing real-time speech modulation enhancement to increase speech intelligibility in noise. The proposed modulation enhancement technique operates independently in the frequency and time domains. In the frequency domain, a compression function is used to perform energy reallocation within a frame. This compression function contains novel scaling operations to ensure speech quality. In the time domain, a mathematical equation is introduced to reallocate energy from the louder to the quieter parts of the speech. This proposed mathematical equation ensures that the long-term energy of the speech is preserved independently of the amount of compression, hence gaining full control of the time-energy reallocation in real-time. Evaluations on intelligibility and quality show that the suggested approach increases the intelligibility of speech while maintaining the overall energy and quality of the speech signal.

Index Terms: Modulations, Speech quality, Intelligibility, Energy reallocation, Gain control

1. Introduction

Mobile telephony is a demanding use case for speech intelligibility enhancement since algorithms must operate in real-time with low latency, and be robust to rapidly changing noise environments without perceptibly degrading speech quality. The research community has been developing algorithms that make the speech structure more robust to background noise ([1], [2]), but these do not meet all the industry's needs.

Many successful intelligibility enhancement methods modify the signal's features based on the noise masker ([3], [4], [5]) while others exploit audio and signal properties ([6]), [7]). Another family of algorithms that exploit human-like speech modifications has also been proposed ([8], [9], [10], [11], [12]). These methods analyse casual speech and highly intelligible natural speech (Lombard speech [13], clear speech [14]) and modify the casual speech to reduce the feature differences between the two speaking styles. These human inspired techniques enhance speech regardless of the noise type, which makes them attractive for mobile telephony.

Unfortunately, the majority of these algorithms cannot be directly used in telecommunications. Many of them have been designed to work per sentence (file), making use of advance knowledge of the features of the speech signal ([12], [15], [16], [17]) e.g. the peak per sentence file. However, in real-time applications the speech structure is limited to the current and the past frames or even missing from the current frame due to limitations on the time analysis window (e.g. less than one pitch period), making it more difficult to perform spectral and temporal modifications. In addition, the existing techniques normalize the global energy of the speech signal to its original energy after whole file processing, without reporting the overall energy increase of the input signal directly after modification [15], [16], [18]. For telephony applications, however, the overall output energy must be preserved in order to conserve battery life and prevent loudspeaker distortions, and obviously all processing must be performed in real-time. Equally importantly, the speech quality must be maintained in quiet and fluctuating noise environments, whereas the majority of the previous mentioned highly intelligible algorithms degrade speech quality.

Real-time MODulation enhancement (RMOD) is a new speech intelligibility enhancement algorithm, that addresses these limitations. RMOD is inspired by the DMOD [12], however it performs real-time energy reallocation (approx $2ms$ delay) and uses novel scaling operations to maintain speech quality. The advantage of this algorithm is that it introduces a mathematical approach to link the four important functions for time-energy reallocation, namely; compression function, presentation level, speech quality and maintenance of overall speech power. By describing this problem mathematically, we are able to reallocate the energy in time while preserving the temporal structure of speech and its long-term power. While other algorithms use energy buffers, pre-defined energy input-output curves and threshold operations to estimate the gain of each speech frame to control both the amount of compression and the overall energy increase, RMOD predicts the gains mathematically, based on the desired amount of compression. This avoids additional corrections to the energy reallocation, and hence the addition of distortions and other artefacts.

The aim of our algorithm is to improve intelligibility in real-time, while maintaining quality and preserving the long-term energy of the signal. Hence, we evaluate these three areas separately and compare to Spectral Shaping and Dynamic Range Compression ([16],SSDRC). SSDRC outperforms DMOD and was the most successful intelligibility enhancement algorithm in the Hurricane challenge [2]. Objective evaluations of intelligibility show that RMOD produces intelligibility enhancements that are close to the levels of SSDRC. RMOD has equivalent or higher scores than SSDRC when using objective measures that also account for speech distortions and quality. Subjective evaluations of quality verify that RMOD maintains speech quality and is preferred over SSDRC. We also show that independent of the amount of compression RMOD maintains the RMS (Root Mean Square) speech energy whereas SSDRC shows a higher energy increase and variation across different utterances.

2. Algorithm description

The new algorithm, Real-time MODulation enhancement (RMOD), is designed for practical speech intelligibiltiy enhancement in low delay applications, such as hearing aids and telecommunications. It therefore modifies the speech signal in real-time, with an emphasis on ensuring speech quality and controlling the long-term energy. The compression func-

http://dx.doi.org/10.21437/Interspeech.2017-1157

tion used for energy reallocation is inspired by the DMOD, proposed in [12]. DMOD belongs to the family of human-like speech modifications; it enhances the temporal modulations of speech which are a key aspect of speech perception [19],[20],[21] and intelligibility [22], and hence significantly increases word recognition in noise. Rather than using static input-output curves for performing energy reallocation in the frequency [5],[17] and time domain [16],[17],[18] DMOD uses a mathematical equation to increase the energy of quieter parts compared to louder parts of speech; $A'_k[t] = A_k(t)^\alpha$, $0 < \alpha < 1$ is a non-linear function that performs this modification to each decomposed k^{th} speech component. This technique increases the low-frequency temporal modulations of speech and thus significantly improves intelligibility. However, the speech decomposition technique [23] used in DMOD requires whole file processing, so is not applicable to real-time use cases.

RMOD applies the same compression function as DMOD, however it does this independently in the frequency (fRMOD) and time (tRMOD) domains, which allows the time domain module to control the long term energy. In addition, RMOD overcomes the major limitations of DMOD which are (1) a-priori knowledge of the whole waveform (2) speech degradation due to excessive gains at very low amplitudes (3) lack of robustness to changes in the overall energy level of the speech signal (presentation level). The two independent modules are described in detail below.

2.1. Frequency domain RMOD (fRMOD)

The fRMOD enhances speech intelligibility at the frame level, by redistributing energy from the high energy harmonics to the lower energy regions, without affecting the overall frame energy. Let $X = \{X_0, X_1, ...X_{P-1}\}$ be the estimated amplitude spectrum coefficients of the current frame and $\phi = \{\phi_0, \phi_1, ...\phi_{P-1}\}$ be the estimated phase spectrum. From the amplitude spectrum, the power spectrum is estimated and the P energy coefficients are grouped into B frequency bands $Y = \{Y_0, Y_1, Y_{B-1}\} = \{\sum_{k=0}^{N_1} X_k^2, \sum_{k=N_1+1}^{N_2} X_k^2, ..., \sum_{k=N_{B-1}+1}^{P-1} X_k^2\}$. The maximum frequency band energy $Y_M = max\{Y_0, Y_1, ...Y_{B-1}\}$ is estimated and each frequency band is normalized by Y_M, eq(1).

$$N(Y) = \frac{Y}{Y_M} \tag{1}$$

This normalisation makes the energy reallocation independent of the frame energy, so that the energy distribution in the frequency domain depends only on the relative energies between frequency bands. Each normalized frequency band energy is then smoothly bounded by a very small value $\epsilon << 1$ using a scaling function, eq(2). This scaling operation prevents any very low energies from being excessively amplified which could otherwise lead to audible distortions and also removes the need for threshold operations which can distort the speech structure as shown in Fig.1.

$$S(Y) = \frac{N(Y) + \epsilon}{1 + \epsilon} \tag{2}$$

After normalization and scaling, the gain $G(Y)$ of each frequency band is estimated, eq(3), based on the compression rule of DMOD, $(.)^\alpha$. Hence the compression per frequency band can be altered depending on the application (e.g. telephony, hearing aids). Each amplitude coefficient is then multiplied by the estimated gain $G(Y)$ of the frequency band that it belongs to. The modified amplitude spectrum X' is normalized to have the same RMS energy as the original spectrum X. The signal

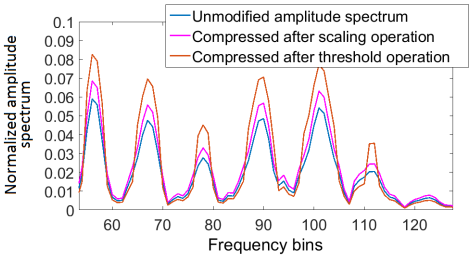

Figure 1: *Importance of the proposed scaling operation compared to state-of-the-art threshold operations. Spectrum of the original speech, compressed speech after scaling and compressed speech after thresholding. Both scaling and thresholding aim to protect the very low amplitudes of the speech from enhancement in order to avoid distortion. While the scaling operation preserves the overall shape of the original waveform, the threshold operation (threshold = 0.015) leads to a distortion at the boundary decision around the 110^{th} frequency bin.*

is reconstructed using the modified amplitude spectrum X' and the original phases ϕ, before being further modified by tRMOD.

$$G(Y) = \frac{S(Y)^{\alpha_f}}{S(Y)} \tag{3}$$

2.2. Time domain RMOD (tRMOD)

The tRMOD enhances speech intelligibility by reallocating energy from the louder to the quieter parts of the speech in real-time. Since this time-energy reallocation is described mathematically, there is full control of the energy of the modified speech, and the final gain for each frame can be predicted before applied to the speech signal. No corrections are made to the signal after applying the gains (e.g energy reduction/increase due to insufficient/sufficient resources), which reduces artefacts in the signal. Furthermore, the estimated gains adapt to the amount of energy reallocation (compression), so the user has the option to preserve the total RMS or increase it by a specific amount. As described previously, the scaling operation removes the need for thresholds that separate low from mid/high amplitudes, thus further ensuring speech quality. tRMOD is designed to adjust dynamically to the presentation level of the speech, making it attractive for real-time applications. Since it is designed to work on $2ms$ frame length, it is also independent of the pitch period and hence the speaker's gender unlike other techniques [16, 18].

Let x_0 be the estimated RMS energy of the current frame and M the maximum peak energy of the past K frames, $M = max\{x_{K-1}, x_{K-2}, ...x_0\}$. The current energy frame is normalized by the peak energy, exactly as in eq.(1), $N(x) = \frac{x}{M}$, where $x = x_0$. This operation disengages the time-energy reallocation from the presentation level of speech. Then, just as in eq.(2) of the frequency module, the estimated normalized energy is smoothly bounded by a very small value $\epsilon << 1$ to protect the quieter parts of speech from being excessively amplified, $S(x) = \frac{N(x)+\epsilon}{1+\epsilon}$. The gain, $C(x)$ of the current frame is then estimated using the compression function. The compression rule increases the louder frames to a lesser extent than the quieter frames ($C(x) > 1$) while it keeps the energy of the peak intact ($C(M) = 1$).

$$C(x) = \frac{S(x)^{\alpha_t}}{S(x)} \tag{4}$$

Finally, the gain is reduced by $\gamma \geq 0$ and applied to the frame:

$$G(x) = C(x) - \gamma(\alpha_t) \tag{5}$$

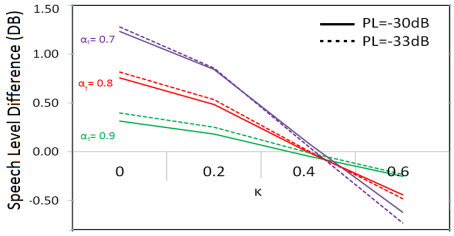

Figure 2: *Energy control of RMOD using the κ parameter. The RMS energy difference between processed and unprocessed speech is depicted with varying α_t and κ values. For $\kappa = 0.45$ the energy of the original speech is maintained independent of the amount of compression α_t and the presentation level (PL).*

This shifting operation compresses the peak ($G(M) = C(M) - \gamma = 1 - \gamma$) and the high energy frames of speech while increasing the frames with low energy. The parameter γ is mathematically derived from eq.(4) and eq.(5) by mapping the κ energy shift from the peak M to the γ gain shift from the gain $C(M)$:

$$\gamma(\alpha_t) = (1-\kappa)^{(\alpha_t - 1)} - 1 \tag{6}$$

In Fig.2, the RMS energy difference between the modified and unmodified speech is depicted in dB with varying α_t and κ values. The unmodified speech file contains 26 seconds of Korean speech from 4 different speakers, the energy was then decreased by 3dB to create a second testing sequence with a different presentation level of speech. tRMOD is then applied to the two input sequences using combinations of 3 different compression values $\alpha_t = \{0.7, 0.8, 0.9\}$ and 4 different $\kappa = \{0, 0.2, 0.4, 0.6\}$ values. Fig.2 shows that for $\kappa = 0.45$ the original speech energy is maintained independent of the amount of compression and presentation level. For higher values of κ, the modified signal has lower energy than the unmodified, while for lower values of κ the RMS energy increases proportionally to the amount of compression α_t. Therefore, by manipulating the α_t and κ parameters, RMOD has full control of the energy increase of speech while operating in real time.

Fig.3 compares the energy distribution in time of the original speech, with that of speech processed by RMOD and SSDRC ([16]). SSDRC uses a static input-output envelope curve to perform energy reallocation in time. When reducing the α_t value of RMOD, the higher energy parts of speech are more compressed and the lower energies are more enhanced. The RMOD maintains the local minima and maxima of the temporal envelope of speech, preserving the overall speech structure compared to SSDRC. Despite the low aggressiveness of the time-energy reallocation algorithm, designed to preserve speech quality, RMOD still increases speech intelligibility.

3. Evaluation

3.1. Evaluation of energy control

It was shown in Fig.2 that the RMS energy difference between the modified and unmodified speech is close to zero, independent of the amount of compression and the presentation level for $\kappa \approx 0.5$. To further support this argument, RMOD was evaluated on 720 sentences of the Harvard corpus [24] for two different compression values $\alpha_t = 0.7$ (RMOD$_1$) and $\alpha_t = 0.9$ (RMOD$_2$) and for $\kappa = 0.5$, $\alpha_f = 0.9$. In Fig.4, the distribution of the energy difference between processed and unprocessed speech across sentences is depicted. RMOD maintains the energy of speech independent of the compression factor α_t, while

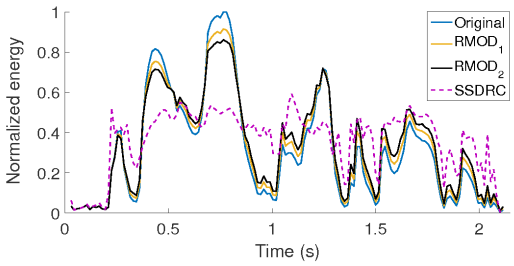

Figure 3: *Energy distribution in time of the original speech, the speech modified by RMOD using two different compression values $\alpha_t = 0.9$ (RMOD$_1$) and $\alpha_t = 0.7$ (RMOD$_2$) and the speech modified by SSDRC. SSDRC over compresses the signal, completely changing the speech structure, while RMOD reallocates energy in time while preserving the local maxima and minima.*

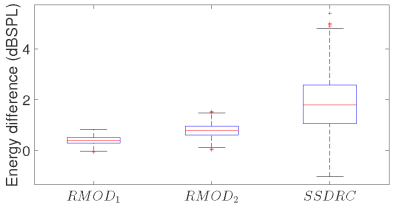

Figure 4: *The $\{$min, 1st quartile, mean, 3rd quartile, and max$\}$ of the energy difference of processed and unprocessed speech across 720 sentences. SSDRC increases the RMS energy of speech and shows much higher variance compared to RMOD. RMOD maintains the RMS energy independent of the amount of compression (RMOD$_1$: $\alpha_t = 0.9$, RMOD$_2$: $\alpha_t = 0.7$)*

SSDRC increases the RMS energy of speech. Furthermore, SSDRC shows high variance across sentences and in some cases it fades the speech signal (negative energy difference), revealing the disadvantage of using static input-output curves for time-energy reallocation, when RMS preservation is a requirement. In contrast, the RMOD algorithm adjusts to the speech characteristics and therefore maintains the RMS of speech.

3.2. Objective evaluations of intelligibility and quality

To assess the intelligibility gain of RMOD, its performance was compared to the original speech and to SSDRC modified speech, the most intelligible algorithm in the Hurricane challenge 2013 [2]. Several objective measures of intelligibility were used, which are highly correlated with subjective evaluations; the extended Speech Intelligibility Index (extSII,[25]), the Glimpse Portion model (GP,[26]), the Distortion Weighted Glimpse Portion model (DWGP,[4]) and the Short-Time Objective Intelligibility measure (STOI,[27]). While GP and extSII assess intelligibility, DWGP and STOI are correlated with both intelligibility and quality. The Perceptual Objective Listening Quality Analysis (POLQA,[28]) was also used, as it is commonly used in the telecommunications industry as an objective measure of quality for speech up to Super Wideband (SWB).

The performance evaluation was designed to simulate real noisy conditions. Three types of noise maskers were used: competing talker (V), real-recorded canteen noise (C) and an artificial combination of competing talker with canteen noise (M). The speech file contains 26s of Korean speech recorded at 32kHz (SWB) uttered by 4 different speakers to simulate a phone call. The speech file was downsampled to 16kHz (Wide-

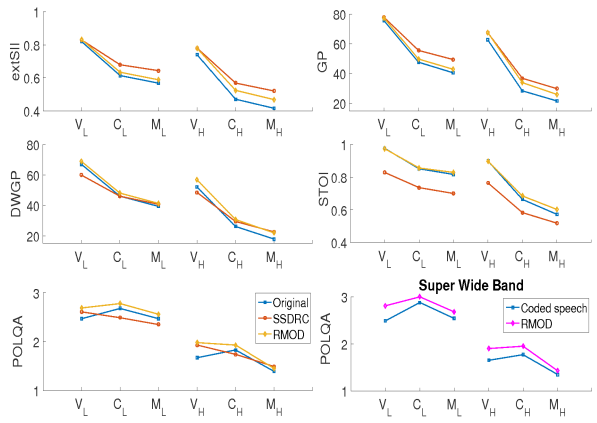

Figure 5: *Objective scores of original unmodified speech, RMOD and SSDRC for three different noise maskers competing speaker (V), canteen noise (C) and their combination (M) in high and low SNR levels.*

band, WB). The objectives scores were calculated for the WB speech, that is the unmodified speech, the RMOD processed speech and the SSDRC modified speech in the presence of each noise masker. Two noise masker levels were used, yielding two different SNR levels to simulate mild $\{V_L, C_L, M_L\}$ = {9, 11 ,8}dB and more severe noisy conditions $\{V_H, C_H, M_H\}$ = {-1, 5, -3}dB. RMOD parameters are set to $\alpha_t = 0.9$, $\alpha_f = 0.9$ and $\alpha_t = 0.7$, $\alpha_f = 0.9$ for the higher and lower SNR respectively. For ensuring speech quality, $\epsilon = 0.01$. For WB speech no frequency band grouping is performed ($B = P$). RMOD performs real-time processing while SSDRC performs whole-file processing. For evaluation purposes, both algorithms were normalized to the RMS of the unmodified speech (the RMS increase for the un-normalised RMOD was less than 0.8dB).

Figure 5 summarizes the objective intelligibility scores. Firstly, the extSII and GP models report that the intelligibility score of RMOD is higher than that of the original speech. Especially for the lower SNR, the intelligibility gains of RMOD approach those of SSDRC for both the competing speaker and the canteen noise, although the SSDRC has higher intelligibility gains in all cases. Secondly, the DWGP and STOI that account for both for intelligibility and quality, score the RMOD higher or the same as original speech while SSDRC scores lower than original speech in the majority of the cases. Finally, the POLQA score for speech quality reports increased scores of RMOD for all noise levels and maskers while the performance of SSDRC varies according to the masker type and level.

Results suggest that the intelligibility gains for both algorithms are quite low when tested with real noises. This contrasts with previous results obtained for SSDRC on SSN and demonstrates the importance of testing speech enhancement algorithms in wide range of realistic conditions. The final set of results show the good performance of RMOD for SWB coded speech. Our original speech file recorded at 32kHz was processed by the Enhanced Voice Services speech codec [29] currently used in mobile telephony. POLQA measures the speech quality of the coded speech before and after RMOD modification (Fig.5). Most speech enhancement algorithms have only been designed for Narrow Band (NB) and WB speech, whereas RMOD is applicable for all bandwidths. The development of speech processing algorithms for higher bandwidths is likely to become increasingly important now that SWB speech codecs are commercially deployed.

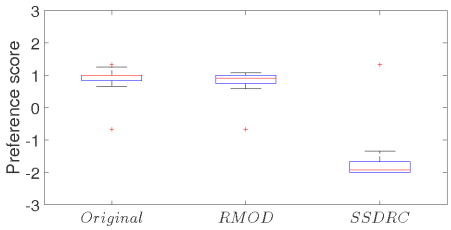

Figure 6: *Subjective quality evaluation of original, RMOD ($\alpha = 0.7$) and SSDRC speech. The {minimum, 1^{st} quartile, mean, 3^{rd} quartile, maximum} of the preference scores across 22 listeners.*

3.3. Subjective evaluations of quality

The quality of the original speech, RMOD and SSDRC modified speech were compared using preference listening tests. These tests were carried out in quiet conditions since for telecommunications purposes it is important to ensure that no speech degradation occurs. Four Harvard sentences, sampled at 16kHz were randomly selected and processed with RMOD ($\alpha_t = 0.7$, $\alpha_f = 0.9$) and SSDRC. They were then presented with the original speech to 22 native or L2 listeners. Listeners had to select from -3 to 3 to indicate the degree of preference between pairs in terms of quality, with 0 corresponding to the same quality and 3 (-3) to much better (worse) quality of the one signal compared to the other. Fig. 6 summarizes the preference scores, RMOD and unmodified speech appear to have similar preference scores, whereas SSDRC gives negative quality scores. This demonstrates that RMOD preserves the quality of speech. Similar results were also obtained in Korean, using the same methodology with a smaller set of listeners.

4. Conclusions

This work introduces a new algorithm that enhances speech intelligibility for mobile telephony. RMOD operates in real-time, while maintaining speech quality and preserving the long-term energy of the speech. RMOD is inspired by the compression rule of DMOD which has been previously shown to increase the temporal modulations of speech and emphasize its harmonic structure, and hence gives high intelligibility gains. Rather than the full file speech decomposition used in DMOD, the compression scheme of RMOD is applied separately in the frequency and time domains, and it is this approach that enables the algorithm to perform in real-time.

Furthermore, in the authors' opinion there are four important facets that must be considered when using time-energy reallocation for speech intelligibility enhancement; namely the compression/expansion function, the presentation level, the speech quality and the maintenance of overall speech power. According to the authors' knowledge, these have not previously been described and simultaneously controlled by a mathematical equation. Our use of this novel equation to enhance intelligibility ensures that the speech quality is maintained and artefacts avoided, whilst also giving control of the long-term RMS energy of the output speech.

RMOD has been consistently scored positively by the most common objective evaluation measures; its intelligibility improvements approach that of SSDRC. It scores equivalently or better for objective measures that correlate with speech quality, and this is supported by listening preference tests. In contrast to SSDRC, RMOD operates in real-time and preserves both the energy and the quality of the original speech.

5. References

[1] W. B. Kleijn, J. B. Crespo, R. C. Hendriks, P. N. Petkov, B. Sauert, and P. Vary, "Optimizing speech intelligibility in a noisy environment." *IEEE Signal Processing Magazine*, pp. 43–54, 2015.

[2] 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*, vol. 55(4), pp. 572–585, 2013.

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

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

[5] H. Schepker, J. Rennies, and S.Doclo, "Improving speech intelligibility in noise by SII-dependent preprocessing using frequency-dependent amplification and dynamic range compression," *Interspeech*, pp. 3577–3581, Lion France, 2013.

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.

[11] M. Koutsogiannaki, P. Petkov, and Y. Stylianou, "Intelligibility enhancement of casual speech for reverberant environments inspired by clear speech properties," *Interspeech*, pp. 65–69, 2015.

[12] M. Koutsogiannaki and Y. Stylianou, "Modulation enhancement of temporal envelopes for increasing speech intelligibility in noise," *Interspeech*, pp. 2508–2512, 2016.

[13] W. Summers, D. Pisoni, R. Bernacki, R. Pedlow, and M. Stokes, "Effects of noise on speech production: Acoustic and perceptual analysis," *J. Acoust. Soc. Amer.*, vol. 84, pp. 917–928, 1988.

[14] 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.

[15] V. Tsiaras, T. Zorila, Y. Stylianou, and M. Akamine, "Real-time speech-in-noise intelligibility enhancement based on spectral shaping and dynamic range compression," *IEEE Int. Conf. Acoust., Speech, Signal Process.*, Florence, Italy 2014.

[16] 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.

[17] T. Zorila and Y. Stylianou, "On spectral and time domain energy reallocation for speech-in-noise intelligibility enhancement," *Interspeech*, pp. 2050–2054, 2014.

[18] T. C. Zorila, Y. Stylianou, T. Ishihara, and M. Akamine, "Near and far field speech-in-noise intelligibility improvements based on a time-frequency energy reallocation approach," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 10, pp. 1808–1818, 2016.

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

[20] S. Sheft and W. Yost, "Temporal integration in amplitude modulation detection." *J. Acoust. Soc. Amer.*, vol. 88, pp. 496–805, 1990.

[21] S. Shamma, "Auditory cortical representation of complex acoustic spectra as inferred from the ripple analysis method." *Network Comput.Neural Syst*, pp. 489–476, 1996.

[22] R. Drullman, J. Festen, and R. Plomp, "Effect of reducing slow temporal modulations on speech reception," *J. Acoust. Soc. Amer.*, vol. 95, no. 5, pp. 2070–2680, 1994.

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

[24] E. Rothauser, W. Chapman, N. Guttman, H. Silbiger, M. Hecker, G. Urbanek, K. Nordby, and M. Weinstock, "Recommended practice for speech quality measurements," *IEEE Trans.Audio Electroacoust.*, vol. 17, pp. 225–246, 1969.

[25] ANSI S3.5-1997, "American national standard methods for calculation of the speech intelligibility index," American National Standards Institute, New York, Tech. Rep., ANSI (1997).

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

[27] C.H.Taal, R.C.Hendriks, R.Heusdens, and J.Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," *ICASSP 2010, Texas, Dallas*, 2010.

[28] Recommendation ITU-T P.862, "Perceptual objective listening quality assessment (POLQA)."

[29] 3GPP TS 26.441, V13.0.0 (2015-12), "Codec for enhanced voice services (EVS)," General Overview.