ADAPTIVE GAIN CONTROL AND TIME WARP FOR ENHANCED SPEECH INTELLIGIBILITY UNDER REVERBERATION

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

[1]Toshiba Research Europe Ltd., Cambridge, UK [2]Department of CS, University of Crete, Greece

ABSTRACT

Moderate and severe reverberation reduce speech intelligibility as a result of the overlap-masking effect, which constitutes the simultaneous observation of multiple delayed and attenuated copies of the speech signal. Recent progress has been made in ameliorating the degradation in intelligibility by adaptively controlling the signal gain as a function of both the signal statistics and the properties of the environment. While the intelligibility gain is at present small, it is significant and serves as a clear indication that reduced masking in non-stationary portions of the signal, under appropriate smoothness constraints, correlates well with an increase in intelligibility. A multi-modal modification framework is expected to improve performance further by i) introducing additional means to reduce masking in designated portions of the speech signal and ii) reducing signal distortion introduced by the use of a single modification modality. In particular we consider the combination of adaptive gain control and local time-warp dependent on the gain modification. A listening test shows that the proposed approach outperforms linear time warp. Objective validation results provide additional insight into the specifics of the proposed multi-modal modification approach.

Index Terms— speech intelligibility, reverberation, speech modification, power dynamics recovery, time warp

1. INTRODUCTION

Normal-hearing listeners exhibit high robustness to mild and moderate reverberation. As reverberation time, a characteristic of the environment describing the decay rate of reverberation power, increases beyond the one second threshold, a measurable intelligibility degradation occurs [1, 2]. In addition, excessive signal energy levels, caused by signal reflections, can become the cause of low listening comfort. High levels of reverberation are more commonly observed in large acoustic environments such as shopping malls, hotel lobbies, airport halls and tunnels. These can degrade the usability of public announcement and disaster prevention systems. In this work we consider the design of a multi-modal signal modification approach for improved speech intelligibility under reverberation.

The impulse response of a room consists of three components: direct path (shortest path if no direct path between speaker and listener exists), early reflections (ER) and late reverberation (LR), e.g., [3]. ER span a short window after the arrival of the direct signal and are not considered to be detrimental to intelligibility due to high correlation with the direct-path signal. LR, on the other hand, is induced by signal portions with larger time separation from the direct path signal. The lower correlation between the two implies that LR is the primary cause of intelligibility degradation. Unlike ER that depend strongly on the hall geometry and the positions of the talker and the listener, LR is diffuse and can be modeled statistically [4].

A number of intelligibility enhancing speech modifications for reverberant environments, which preserve the time scale, can be found in the literature. These range from modulation enhancement filtering [5], to steady state suppression [2] and room impulse response reshaping [6]. In practice, there is limited evidence for the effectiveness of these methods under strong reverberation.

A subset of methods addresses the joint effect of noise and reverberation. A perceptual distortion measure is optimized for a band-based gain modification in [7]. A similar modification strategy optimizing an augmented variant of the speech intelligibility index (SII) [8] is proposed in [9]. SII optimization by spectral shaping and dynamic range compression for noise and reverberation is studied in [10]. The first two methods are validated subjectively in contexts where degradation is largely dominated by the noise.

Time scale modifications improve intelligibility at the cost of reducing the information transfer rate. Zero-padding in the steady state is considered in [11], while linear time-scaling is employed in [12, 13]. Manual adjustment of the pause duration and time scale factors in these methods prevents context awareness and scalability. A multi-modal signal modification approach comprising gain adjustment, time scaling and pause insertion is proposed in [14]. Similarly, the degree of modification in each components is set manually.

Previously we showed that adaptive gain control (AGC) enhances intelligibility by increasing the signal-to-late-reverberation ratio (SLRR) in non-stationary portions of the speech signal under a smoothness constraint [15]. This result is interpreted in terms of the relative importance of signal segments with low predictability to intelligibility. The method increases or decreases signal power adaptively, as a function of the reverberation level and the degree of signal non-stationarity. We now consider the combination of AGC and time warp (TW) as a means to achieving additional SLRR increase in signal regions important for intelligibility.

Deriving an optimal, in some sense, degree of TW is an attractive but challenging objective. Joint optimization with the power gain is impractical. More generally, optimizing SLRR for the degree of TW would incur an algorithmic delay due to the need for a look-ahead window. A less sophisticated but more tangible alternative is to induce TW in response to the outcome of the gain optimization. In particular, stretching the time when signal power decreases under strong reverberation (typically observed in stationary portions of the signal) allows for additional decay of the reverberation power. In turn, SLRR in upcoming non-stationary portions of the signal is expected to increase. We validate this strategy objectively and show that the degree of TW adapts smoothly to the LR power level. A listening test verifies that the proposed combination of AGC and TW (AGCTW) outperforms linear time warp (LTW) of the signal.

The remainder of this paper is organized as follows. The methodology is summarized in Section 2. A practical system implementation is described in Section 3. Validation results are given in Section 4 followed by conclusions.

2. METHODOLOGY

Section 2.1 summarizes key aspects of AGC. The TW operation and its interaction with AGC are discussed in Section 2.2.

2.1. Power gain optimization

AGC is based on an input-output power mapping derived as the minimizer to the distortion criterion

$$\eta = \int_{\alpha}^{\beta} \left(\frac{1}{x} \left(y + l - x\frac{dy}{dx} \right)^2 + \lambda l^2 \frac{y}{x} \right) f_X\left(x|b\right) dx, \qquad (1)$$

where x is the (instantaneous) power of the input (natural speech) signal, y is the power of the output (modified) signal, l is the LR power, α and β determine the range of interest and b is the shape parameter in the probability density function of x. The first additive term under the integral is a distortion measure and the second term is a penalty on the signal gain weighted by a simple polynomial function of the LR power [15]. The distortion measure characterizes the deviation of the power dynamics of the signal, in the presence of an additive distortion, from the dynamics of clean speech.

We thus assume that LR can be modeled as uncorrelated additive noise. This is a strong assumption which works well in practice due to the time separation of the current instant from the signal past responsible for the LR signal. The second power over l in the penalty leads to an inversion of the modification direction for sufficiently large l due to the rapidly increasing importance of the penalty.

The optimal solution comes in the polynomial form

$$y\left(x\right) = c_1 x + c_2 x^b + \frac{l}{2b}\left(l\lambda - 2b\right), \qquad (2)$$

where c_1 and c_2 are determined from the boundary conditions:

$$y\left(\alpha\right) = \alpha \qquad (3) \qquad\qquad y'\left(\psi\right) = \rho. \qquad (4)$$

ρ is further parametrized as $\rho = \varsigma^l, \quad \varsigma \in (0,\ 1)$ to ensure that the signal remains unchanged in the absence of reverberation.

The power mapping in (2) is readily calibrated to change modification direction at a particular maximum boosting power (MBP), defined as the crossing point between $y\left(x\right)$ and $y = x$. The short notation for the Lagrange multiplier λ producing the desired behavior is $\tilde{\lambda}$. The reverberation power $\tilde{l}$ at which this MBP is achieved is $\tilde{l} = {}^{b}/_{\tilde{\lambda}}$. The behavior of the mapping is illustrated in Figure 1.

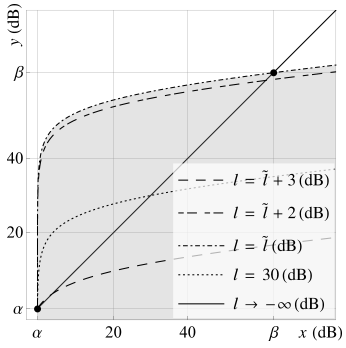

Fig. 1: Power mapping functions for $\lambda = \tilde{\lambda}$ and several values for l. Note the rapid decrease in MBP for $l > \tilde{l}$.

To prevent over-emphasis in stationary regions, and excessive suppression in non-stationary ones, the value of λ is adjusted to reflect the degree of non-stationarity ξ. Finally, the signal gain $g = \sqrt{y/x}$ is smoothed adaptively to produce $\breve{g}$, which prevents intelligibility degradation from rapid gain fluctuations.

2.2. Time warping

When power suppression occurs under the condition $l > \tilde{l}$, an attempt is made to warp the time scale locally. The signal is extended by superimposing (using complementary windows) a segment from the past of the waveform with the latest frame. Similar to waveform similarity overlap and add (WSOLA) [16], the optimal lag is identified as the highest peak in the correlation function $R_{yy}\left(k\right)$ where:

$$R_{yy}\left[k\right] = \sum_{n=1}^{N} \mathrm{y}\left[n - N\right] \mathrm{y}\left[n - k\right], \qquad (5)$$

and N is a suitably chosen window size. The optimal lag is:

$$k^* = \underset{k\in\{K_1,\ K_2\}}{\operatorname{argmax}} R_{yy}\left[k\right], \qquad (6)$$

$$\mathbf{sbj.\ to}\ \bar{r}_1 < {}^{R_{yy}[k^*]}/_{R_{yy}[0]} < \bar{r}_2$$

where K_1 and K_2 bound the search interval and the constraint on the peak value ensures a high degree of smoothness.

An additional constraint is introduced to prevent excessive warping of the speech signal which may result in robotic-sounding speech, i.e., repetition of speech segments with low variability. This objective is achieved by ensuring a minimum time separation $\mathrm{T_{tw}}$ between the current instant, denoted by t and the instant at which the previous time warp, denoted by $\mathrm{t_{ptw}}$, was initiated. The complete set of constraints taken into consideration is illustrated in Figure 2.

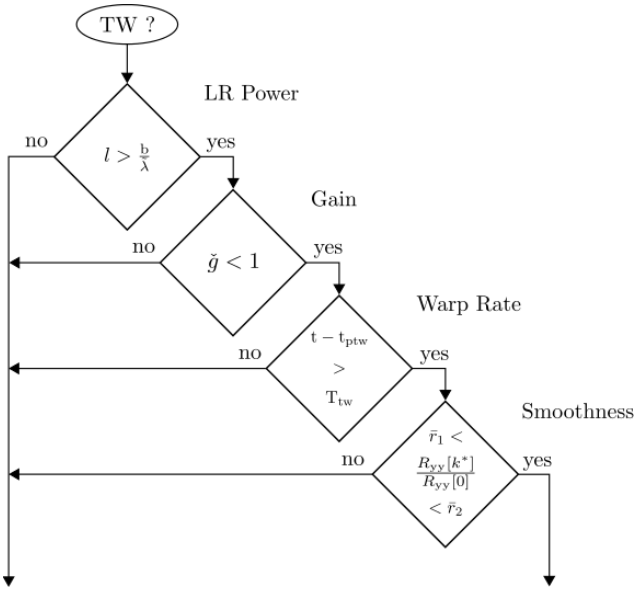

Fig. 2: Complete condition set for warping the local time scale.

To prevent distortion from warping the time scale based on correlation analysis in the domain of the gain-modified signal, the input (natural) signal is used instead. For optimal performance, the signal extension resulting from TW is gain-optimized.

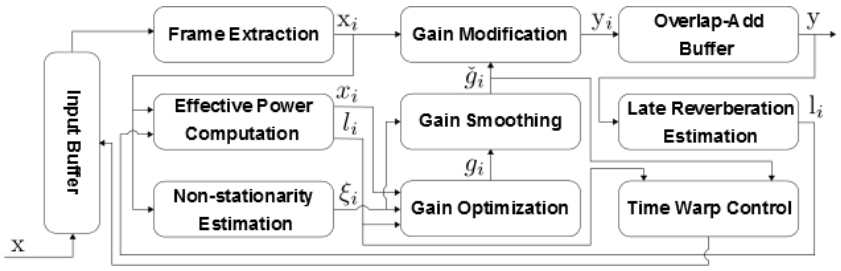

Fig. 3: Operation diagram of AGCTW.

3. SYSTEM DESIGN

To validate the effectiveness of the proposed methodology, a practical system implementation is considered next. A signal flow diagram is given in Figure 3. Generally, instantaneous measures are approximated by frame-based estimates. The key differences from previous work, in the context of AGC, are the presence of an input buffer where the time scale is modified, and an associated time-warp control module. A further difference from [15] is the use of an advanced reverberation model as discussed in Section 3.1.

3.1. Late reverberation model

The experimental validation in [15] relied on simulating reverberation using the source-image method under the assumption of frequency-independent reverberation time RT_{60}. The flatness of the room response results in reverberant speech, which is less intelligible than what can be expected from a real-world environment due to increased masking in the high frequency range.

In this work we similarly simulate reverberation using the source image method but based on a model that introduces frequency-dependent reverberation time and diffuse reflections [17]. The dependence of RT_{60} on frequency is implicitly affected by the choice of frequency-dependent reflection coefficients.

To simulate environments with different levels of reverberation for a hall with fixed dimensions, a realistic choice for the construction material used in all six reflective surfaces is made first. The diffuse reflections coefficient (assuming a single value for all reflective surfaces) is then modified to obtain a range of impulse responses. The direct sound in all cases has the same power based on a constant distance between the source and the receiver but the ER and LR energies change. For brevity of notation we denote the energy of the late part of the impulse response by LIREN. To facilitate the evaluation, the delay and magnitude of each impulse response is normalized to the direct sound.

Regarding LR estimation for the operation of AGCTW, we invoke the diffuseness property and use the LR part of a RIR recording from the same hall based on different source and receiver locations. An alternative approach is to use blind estimation, e.g., [4]. To preserve the source-target distance, and respectively SLRR for given diffusion and reflection coefficients, the new location coordinates are obtained by rotating the reference coordinates (source and receiver) around the center of the floor plane using the rotation matrix:

$$\mathbf{R} = \begin{bmatrix} \cos(\theta) & -\sin(\theta) & 0 \\ \sin(\theta) & \cos(\theta) & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad (7)$$

where θ is the rotation angle in radians.

4. EXPERIMENTS

The experimental set-up considered throughout the evaluation assumes hall dimensions $20\mathrm{m} \times 30\mathrm{m} \times 8\mathrm{m}$, with reference source and receiver locations $\{10\mathrm{m}, 5\mathrm{m}, 3\mathrm{m}\}$ and $\{10\mathrm{m}, 25\mathrm{m}, 1.8\mathrm{m}\}$ respectively. The hall surfaces are: unglazed brick (back), plaster-gypsum (front, left and right), parquet on concrete (floor) and poured concrete (ceiling) [18]. The humidity in the hall is $42\ \%$ and the temperature is 20° C. Both the source and the receiver locations are rotated by $-\ 15^\circ$ to obtain the test positions. The LR model comes from a RIR measurement made for a rotation angle of 30°.

The remaining parameters, not covered by, or different from AGC in [15] are listed below. The search range for the optimal correlation lag k^* corresponds to a pitch range of $50\ -\ 320$ Hz and accommodates a large pool of speakers. $\bar{r}_1\ =\ 1/2$ and $\bar{r}_2\ =\ 3/2$ were found to provide a sufficient degree of smoothness in the TW speech. N (the correlation window length) is set to the duration of the frame. With regard to the system set-up used for generating the material for the listening test, the AGC parameters controlling the degree of gain smoothing when signal power is increased (U) or decreased (D) are set to $\mathrm{U}\ =\ 1.1$ and $\mathrm{D}\ =\ 0.25$. Higher degree of smoothness, compared to [15], was thus allowed in view of the presence of an additional modification modality. Two different values for $\mathrm{T_{tw}}$ are considered to demonstrate the flexibility of the system.

Considering the newly introduced frequency dependence of RT_{60} we simplify the notation by only listing the value of LIREN instead of all reflection and diffusion coefficients. For reference purposes, $\mathrm{LIREN} = 71.7$ for RIR was used in [15].

4.1. Objective Evaluation

The dependence of the extent of TW on the level of reverberation is illustrated in Figure 4. Two values are considered for the parameter

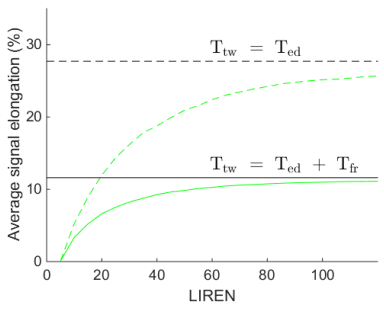

Fig. 4: Average signal elongation as a function of LIREN.

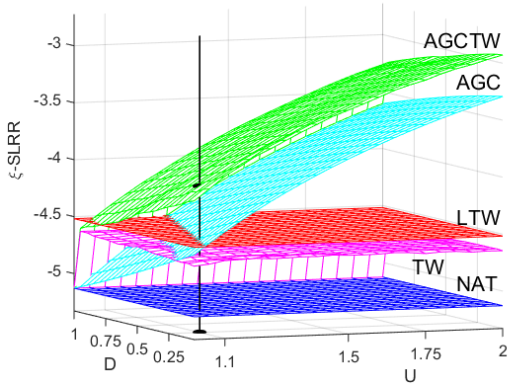

Fig. 5: ξ-SLRR (dB) in the $\mathrm{U} - \mathrm{D}$ space for $\mathrm{LIREN} = 85$.

$\mathrm{T_{tw}}$. In the less restrictive case it is set to the duration of the signal extension resulting from TW at a particular point in time $\mathrm{T_{ed}}$. In the more restrictive case the buffer is increased by the duration of a signal frame $\mathrm{T_{fr}}$. In both cases $\mathrm{U} = 1.1$ and $\mathrm{D} = 0.25$. We note the smooth dependece of the extent of TW on LIREN.

The non-stationarity-weighted SLRR (ξ-SLRR) measurements given in Figure 5 provide an objective perspective of the modification success in reducing masking by late reverberation in non-stationary regions of the signal. This measure correlates well with subjective results when the gain modification is sufficiently smooth [15]. Each point from the grid of values in the U-D space is an average over thirty sentences. Notably AGCTW ($\mathrm{T_{tw}} = \mathrm{T_{ed}} + \mathrm{T_{fr}}$) outperforms AGC justifying the proposed TW criterion. LTW, using WSOLA, outperforms both natural speech (NAT), which is an expected result, and TW - an internal reference based on AGCTW excluding the gain control. The reverberation condition used for this simulation is $\mathrm{LIREN} = 85$.

Figure 6 shows the natural, linearly-time-warped and AGCTW-modified ($\mathrm{T_{tw}} = \mathrm{T_{ed}} + \mathrm{T_{fr}}$) waveforms for one speech utterance and $\mathrm{U} = 1.1$, and $\mathrm{D} = 0.25$. AGCTW visibly decreases the signal energy and modifies the power dynamics in accord with the AGC objective, i.e., the gain is adjusted in response to both the level of late reverberation and the degree of signal non-stationarity.

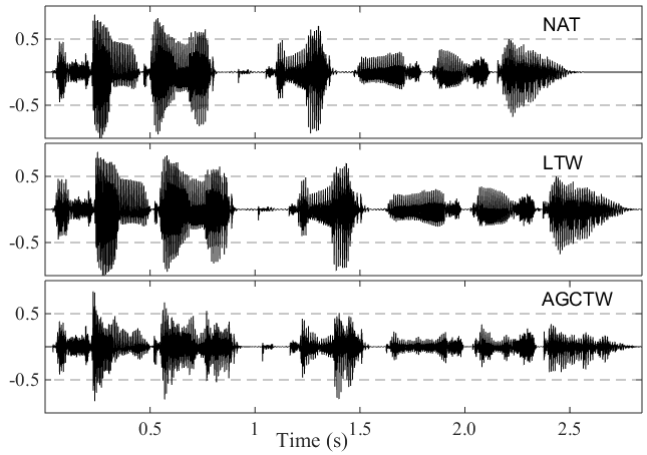

Fig. 6: Natural and modified speech waveforms for $\mathrm{LIREN} = 85$.

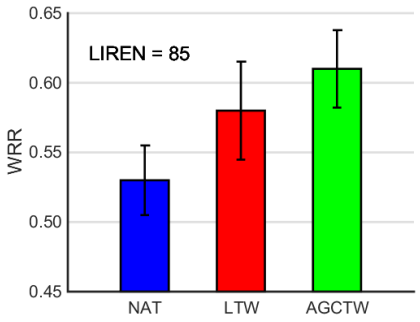

Fig. 7: Word recognition rates (WRR) from a listening test.

4.2. Subjective Evaluation

A listening test was conducted to supplement the objective validation results. The intention was to reproduce the level of difficulty achieved in the listening test from [15]. The difference between the two reverberation models, namely frequency-independent vs. frequency-dependent reverberation time, poses a challenge in this respect. The energy of the late part of the impulse response, in the current scenario, was increased to $\mathrm{LIREN} = 85$ to offset the intelligibility increase due to lower reverberation times in the higher frequency range.

AGCTW used the restrictive constraint $\mathrm{T_{tw}} = \mathrm{T_{ed}} + \mathrm{T_{fr}}$ resulting in an average (over all test sentences) elongation of $10.8\,\%$. The time scaling factor in LTW was fixed accordingly. The objective of the test was, thus, not maximum intelligibility gain but rather an evaluation in a test point where the improvement of AGCTW over LTW could be measured with a small number of subjects.

Nine native (British) English listeners took part in the evaluation. The subjects were paid for their participation. All sentences were power equalized to facilitate comparison. The material was presented diotically, in a silent room, using a pair of Audio-technica ATH-M50x headphones. A training session with 20 sentences familiarized the participants with the task and provided a possibility for volume adjustment. The test material contained 150 Harvard sentences (50 sentences per method) from [19]. Sentence-set-to-method allocation and method presentation order were randomized.

Word recognition rates (WRR) were computed for each sentence based on a single presentation and counting key words only [20]. Mean WRRs over all subjects and standard errors are shown in Figure 7. Both LTW and AGCTW outperform significantly NAT ($p < 0.01$, Student's t test), while AGCTW outperforms significantly LTW ($p < 0.05$).

5. CONCLUSIONS

A bi-modal speech modification approach based on the combination of adaptive gain control and time warp enhances intelligibility significantly under strong reverberation at a modest signal duration increase. The method outperforms linear time warp for the same average duration and, unlike its counterpart, determines the extent of time warp autonomously. Both the gain and the time-scale adjustment adapt continuously to the acoustic environment resulting in scalability. Further significant improvement is expected from extending the operation of the gain controller to multiple bands.

6. REFERENCES

[1] R. H. Bolt and A. D. MacDonald, "Theory of Speech Masking by Reverberation," *J. Acoust. Soc. Am*, vol. 21, no. 6, pp. 577–580, 1949.

[2] N. Hodoshima, T. Arai, A Kusumoto, and K. Kinoshita, "Improving Syllable Identification by a Preprocessing Method Reducing Overlap-Masking in Reverberant Environments.," *J. Acoust. Soc. Am.*, vol. 119, pp. 4055–4064, 2006.

[3] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellerman, "Making Machines Understand Us in Reverberant Rooms," *IEEE Sig. Proc. Mag.*, vol. 29, no. 6, pp. 114–126, 2012.

[4] E. A. P. Habets, S. Gannot, and I. Cohen, "Late Reverberant Spectral Variance Estimation Based on a Statistical Model," *IEEE Sig. Proc. Letters*, vol. 16, no. 9, pp. 770–773, 2009.

[5] A. Kusumoto, T. Arai, K. Kinoshita, N. Hodoshima, and N. Vaughan, "Modulation Enhancement of Speech as a Preprocessing for Reverberant Chambers with the Hearing-Impaired," *Speech Communication*, vol. 45, pp. 101–113, 2005.

[6] A. Mertins, T. Mei, and M. Kallinger, "Room Impulse Response Shortening/Reshaping with Infinity- and p-Norm Optimization," *IEEE Trans. Audio, Speech and Lang. Proc.*, vol. 18, no. 2, pp. 249–259, 2010.

[7] J. B. Crespo and R. C. Hendriks, "Speech Reinforcement in Noisy Reverberant Environments Using a Perceptual Distortion Measure," in *Proc. ICASSP*, 2014, pp. 910–914.

[8] American National Standard, "Methods for the Calculation of the Speech Intelligibility Index," 1997.

[9] R. C. Hendriks, J. B. Crespo, J. Jensen, and C. H. Taal, "Optimal Near-End Speech Intelligibility Improvement Incorporating Additive Noise and Late Reverberation under an Approximation of the Short-Time SII," *IEEE Trans. Audio, Speech and Lang. Proc.*, vol. 23, no. 5, pp. 851–862, 2015.

[10] H. Schepker, D. Hülsmeier, J. Rennies, and S. Doclo, "Model-based integration of reverberation for noise-adaptive near-end listening enhancement," in *Proc. Interspeech*, 2015, pp. 75–79.

[11] T. Arai, "Padding zeros into steady-state portions of speech as a preprocess for improving intelligibility in reverberant environments," *Acoust. Sc. & Tech.*, vol. 26, no. 5, pp. 459–461, 2005.

[12] Y. Nakata, Y. Murakami, N. Hodoshima, N. Hayashi, Y. Miyauchi, T. Arai, and K. Kurisu, "The Effects of Speech-Rate Slowing for Improving Speech Intelligibility in Reverberant Environments," Tech. Rep., The Institute of Electr., Inf. & Comm. Eng., 2006.

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

[14] F. Fuhrmann, K. Dobbler, F. Pokorny, and F. Graf, "A modular system for improving speech intelligibility under extreme acoustic conditions: Subjective evaluation of parameter influence," in *Proc. Forum Acusticum*, 2014.

[15] P. N. Petkov and Y. Stylianou, "Adaptive Gain Control for Enhanced Speech Intelligibility under Reverberation," *IEEE Sig. Proc. Letters*, vol. 23, no. 10, pp. 1434–1438, 2016.

[16] W. Verhelst and M. Roelands, "An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech," in *Proc. ICASSP*, 1993, pp. 554–557.

[17] S. M. Schimmel, M. F. Müller, and N. Dillier, "A fast and accurate "shoebox" room acoustics simulator," in *Proc. ICASSP*, 2009, pp. 241–244.

[18] F.J. Fahy, *Foundations of Engineering Acoustics*, Elsevier Science, 2000.

[19] M. Cooke, C. Mayo, and C. Valentini-Botinhao, "Hurricane natural speech corpus, [sound]. LISTA Consortium: (i) Language and Speech Laboratory, Universidad del Pais Vasco, Spain and Ikerbasque, Spain; (ii) Centre for Speech Technology Research, University of Edinburgh, UK; (iii) KTH Royal Institute of Technology, Sweden; (iv) Institute of Computer Science, FORTH, Greece, http://dx.doi.org/10.7488/ds/140, 2013," Tech. Rep.

[20] M. Cooke, C. Mayo, C. V. Botinhao, Y. Stylianou, B. Sauert, and Y. Tang, "Evaluating the Intelligibility Benefit of Speech Modifications in Known Noise Conditions," *Speech Communication*, vol. 55, pp. 572–585, 2013.