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

Automated pause insertion for improved intelligibility under reverberation

Petko N. Petkov, Norbert Braunschweiler and Yannis Stylianou

Toshiba Research Europe Ltd., Cambridge Research Laboratory, Cambridge, United Kingdom
{petko.petkov,norbert.braunschweiler,yannis.stylianou}@crl.toshiba.co.uk

Abstract

Speech intelligibility in reverberant environments is reduced because of overlap-masking. Signal modification prior to presentation in such listening environments, e.g., with a public announcement system, can be employed to alleviate this problem. Time-scale modifications are particularly effective in reducing the effect of overlap-masking. A method for introducing linguistically-motivated pauses is proposed in this paper. Given the transcription of a sentence, pause strengths are predicted at word boundaries. Pause duration is obtained by combining the pause strength and the time it takes late reverberation to decay to a level where a target signal-to-late-reverberation ratio criterion is satisfied. Considering a moderate reverberation condition and both binary and continuous pause strengths, a formal listening test was performed. The results show that the proposed methodology offers a significant intelligibility improvement over unmodified speech while continuous pause strengths offer an advantage over binary pause strengths.

Index Terms: speech intelligibility, reverberation, speech modification, pause prediction

1. Introduction

Reverberation degrades speech intelligibility as a result of overlap-masking, i.e., multiple, delayed and attenuated copies of an acoustic signal are observed simultaneously [1, 2]. Reverberation comprises early reflections (ER) and late reverberation (LR) [3]. ER arrive shortly after the direct sound and depend on the hall geometry, and the position of the speaker and the listener. LR, on the other hand, is diffuse and consists of a large number of reflections with diminishing magnitudes. LR is the primary cause for the loss of intelligibility cues whereas ER have been shown to have a positive contribution [4].

A range of intelligibility enhancing speech modifications have been proposed. Inverse filtering aims to cancel the effect of reverberation but introduces dependence on the (loud)speaker and listener locations. In addition, room impulse response (RIR) inversion is not generally feasible and requires approximations, e.g., [5]. Modulation enhancement is proposed in [6]. The method performance suggests lack of robustness, likely caused by insufficient context dependence in the design of the modulation enhancement filters.

More recently, distortion criteria are minimized to derive spectral modifications. Perceptual distortion measures are optimized in [7, 8]. A speech intelligibility index (SII) [9] based measure is optimized in [10]. Local SII optimization by spectral shaping and dynamic range compression is studied in [11] and validated with objective measures. All of the above-listed methods address reverberation in combination with noise only.

Decreasing signal power in portions with high redundancy reduces overlap-masking of transients [12, 13]. While the idea is well-grounded, evaluation results suggest room for improvement. Local and global time-scale modifications are shown to be more effective in this respect. Zero-padding in the steady state is proposed in [14] while fixed time-scaling is evaluated in [15]. These methods do not adapt to the specific conditions and apply a user-defined modification depth.

Pause insertion represents an alternative approach to modifying the time scale in the context of reverberation. Signal-based methods, i.e., methods that do not include linguistic information are proposed in [16, 17] and validated on number sequences and matrix sentences respectively. In both cases pause durations are fixed in advance.

When a sentence transcription is available, linguistically motivated pauses can be inserted to improve intelligibility. The importance of pause locations inside a sentence is revealed, e.g., in [18]. Well-placed pauses facilitate sentence parsing by listeners while maintaining a short overall sentence duration. The proposed method computes pause strengths at word boundaries based on the transcription. Using a model of exponential decay for LR power, the time separation needed to achieve a target signal-to-LR ratio is computed for the first sound transition in the word. The effective pause duration is obtained as the product of the pause strength and the time separation. Segmentation information is obtained by forced alignment using pre-trained HMMs from an automatic speech recognizer.

The remainder of this paper is organized as follows. Pause prediction for intelligibility enhancement under reverberation is discussed in Section 2. A practical system design is given in Section 3. Validation results are provided in Section 4 followed by conclusions.

2. Pause prediction

The proposed method identifies pause strengths at word boundaries and determines pause duration as a function of the strengths and the time it takes LR to decay to a particular level. Pause strengths are discussed in Section 2.1. The pause duration model is given in Section 2.2.

2.1. Pause strengths

Pause strengths indicate how natural it is to observe a pause at a particular word boundary. In this work pause strength assessment is performed from text using the multi-stage architecture shown in Figure 1. Initial text processing, including i) text analysis, ii) text normalization and iii) prosodic chunk prediction, is based on Toshiba's TTS system ToSpeak [19]. Prosodic chunk boundaries are then combined with other features in a rule-based scoring algorithm, which outputs continuous-valued scores reflecting the suitability for inserting a pause [20]. The scores are re-evaluated, at a given threshold, in the pause-sequence-evaluation module to prevent extreme events such as:

http://dx.doi.org/10.21437/Interspeech.2016-960

i) multiple consecutive pauses and ii) absence of pauses in long word sequences. The output of this re-evaluation is the continuous-valued pause insertion suitability $\mathcal{I} \in [0,\ 1]$. Finally, $\mathcal{I}$ is mapped to a pause strength w.

While the scoring algorithm uses prosodic chunks as input features, the final sequence of pauses is not restricted to prosodic chunk boundaries. The prosodic chunk prediction module, based on [21], uses a decision tree model trained on an American English TTS corpus that is hand-labeled with ToBI break indices [22]. ToBI break levels three and four were merged into a single break level. Consequently, presence or absence of a prosodic break is predicted for each word juncture in a sentence.

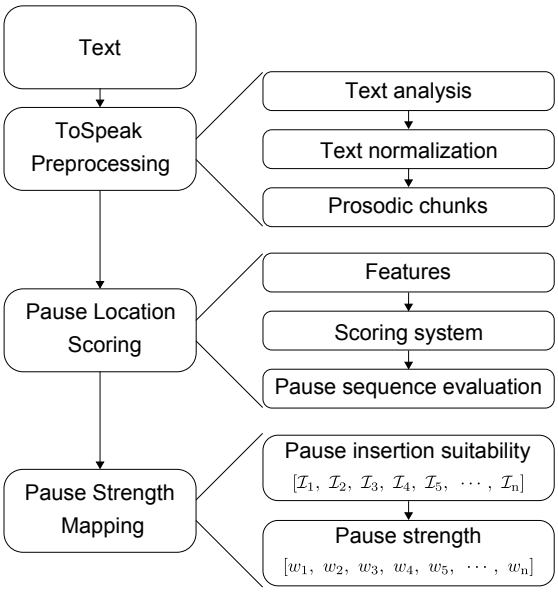

Figure 1: Text-to-pause-strength processing.

Pause strengths, can be obtained in different formats by applying a particular mapping to $\mathcal{I}$. Figure 2 illustrates: i) binary pause strengths based on a threshold $\mathcal{I}_b$ and ii) continuous pause strengths also extending from $\mathcal{I}_b$. In both cases $w \in [0,\ 1]$ holds. Use of a threshold decreases the pause insertion rate and eliminates some of the spurious pauses caused by noise in the training data.

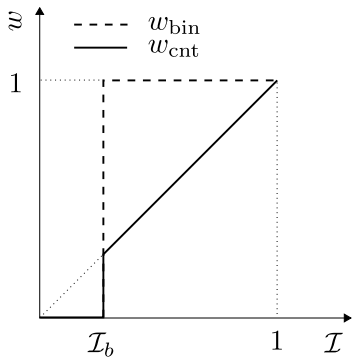

Figure 2: Binary vs. continuous pause strengths.

2.2. Pause durations

The effective pause duration before a word is given by:

$$\tilde{t}_i = w_i\, t_i, \tag{1}$$

where i is the word index, w_i is the pause strength and t_i is the time it takes the late reverberation power, generated by the preceding speech signal, to subside to a particular level. This level is determined from a target signal-to-LR ratio (SLRR) criterion where the first sound transition in word i (in practice the duration of the first two sounds) determines the measurement window.

Let ξ denote the target SLRR and x_i denote the signal power of the waveform x_i in the measurement window. The LR power in this window before pause insertion is l_i. The LR power that meets the criterion is:

$$\tilde{l}_i = x_i\, 10^{\frac{-\xi}{10}}. \tag{2}$$

Assuming an exponential decay of LR power with time [23], the time it takes l_i to decrease to $\tilde{l}_i$ is:

$$t_i = \max\left(0, -\frac{\mathrm{RT}_{60}}{6}\log_{10}\left(\frac{\tilde{l}_i}{l_i}\right)\right), \tag{3}$$

where RT_{60} is the reverberation time and $\max(\cdot)$ ensures that the local time scale is preserved for $l < \tilde{l}$.

By considering linguistic aspects, the proposed pause duration model leads to more natural and effective interruptions of the speech signal [18]. The contribution of early reflections to intelligibility [4] is included in ξ.

3. Practical considerations

Pause insertion for recorded natural speech is realized with the system architecture from Figure 3. The transcription of the input speech signal x is denoted by u_{x}. Forced alignment using pre-trained HMMs is used to locate the beginning of each word and the corresponding first sound transition in the signal waveform. A late reverberation model provides an estimate of the LR power $\hat{l}_i$ for the window of the first sound transition prior to pause insertion. Given the pause strength w_i, $\hat{l}_i$ and the signal power x_i for the same time window, the effective pause duration $\tilde{t}_i$ is computed from eq. (1). The assigned pause is inserted in the output signal y followed by the target word. Processing continues with the next word. The LR model and word segmentation are discussed below.

3.1. Late reverberation model

τ seconds after the arrival of the direct sound individual reflections become indistinguishable. This is the boundary between early reflections and LR [3]. A simple model assuming the exponential decay of LR power with time and a constant RT_{60} over frequency is used here [23]. The LR part of the RIR is modeled as a pulse train $\iota[k]$, amplitude modulated by an exponential decay:

$$\tilde{h}[k] = \iota[k]\, 10^{-3\frac{k}{\mathrm{RT}_{60} f_s}}, \tag{4}$$

where f_s is the sampling rate. The energy of the modulated pulse train is equalized to the energy of the LR part of RIR calculated from a measurement. The approximate LR waveform $\hat{\mathrm{l}}$

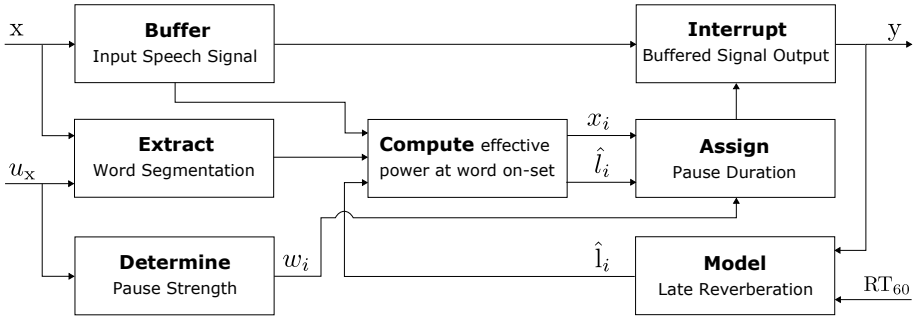

Figure 3: Proposed system architecture.

is given by the convolution of the modulated pulse train $\tilde{h}$ and the output signal y

$$\hat{1}[k] = \sum_{n=1}^{(RT_{60}-\tau)f_s} \tilde{h}[\tau f_s + n]\, y[k - \tau f_s - n], \qquad (5)$$

A sample-based LR power estimate $\hat{l}$ is computed from $\hat{1}$.

3.2. Word segmentation

Phoneme-level segmentation information is obtained by forced alignment using an HTK-based [24] speech recognizer. Training and validation of the recognition engine are summarized in [25]. The phoneme models were context-dependent and the update frame length, which also determines the segmentation resolution of the proposed method, was 10 ms. The measurement window for the signal power x_i in word i includes six states starting from the left-hand-side context of the first phoneme and ending with the right-hand-side context of the second phoneme. For single-phoneme words, only three states are considered.

4. Experimental results

Reverberation is simulated using a source-image method generated RIR [26]. The assumed hall dimensions are $20 \times 30 \times 8$ m, with speaker and listener locations $\{10,\ 5,\ 3\}$ and $\{10,\ 25,\ 1.8\}$ m respectively. For convenience, propagation delay and attenuation are normalized to the direct sound. $\tau = 0.05$ s is assumed based on [3]. A pulse density of $4000\ \mathrm{s}^{-1}$ was used for ι in the late reverberation model.

The target SLLR ξ is set to -10 dB reflecting the 10 dB power advantage of the direct signal and ER over LR for the particular RIR realization. The two underlying assumptions are: i) the complete power advantage of ER over LR translates to an intelligibility gain and ii) a 0 dB level for the ratio of the direct signal and ER to LR is sufficient for accurate parsing of the pause-separated segments in a sentence.

A British English recording of the sentences from [27] was used to conduct objective and subjective experiments. The reference methods included in the evaluation are natural speech (NAT) and linearly time-scaled speech (TSC). The time scale factor reflects the average duration increase from continuous pause strengths and is estimated based on 170 sentences used for testing. The original speech recording is scaled by waveform similarity overlap and add [28]. Both binary ($\mathrm{SP_{bin}}$) and continuous ($\mathrm{SP_{cnt}}$) pause strengths are included. A threshold value $\mathcal{I}_b = 0.25$ was chosen based on subjective judgments made by an expert in linguistics.

4.1. Objective evaluation

Table 1: Examples of pause strengths for two sentences.

Sent. A	*Their*	*eyelids*	*droop*	*for*	*want*	*of*	*sleep*
$\mathrm{SP_{cnt}}$	-	0	.61	0	0	.37	0
$\mathrm{SP_{bin}}$	-	0	1	0	0	1	0
Sent. B	*The*	*marsh*	*will*	*freeze*	*when*	*cold*	*enough*
$\mathrm{SP_{cnt}}$	-	0	.61	0	.57	0	0
$\mathrm{SP_{bin}}$	-	0	1	0	1	0	0

Table 1 presents the continuous and binary pause strengths preceding each word for two test sentences. The first example (Sent. A) illustrates the advantage of $\mathrm{SP_{cnt}}$ over $\mathrm{SP_{bin}}$. The pause before "*droop*" is inappropriately placed and receives full strength when binary mapping is used. The continuous-strength paradigm de-emphasizes this pause. The second example (Sent. B) illustrates appropriate pause placement.

Figure 4 shows the output waveforms from the four methods included in the evaluation for the same test sentence preprocessed for presentation at $\mathrm{RT_{60}} = 1.8$ s. The corresponding reverberant waveforms are presented in Figure 5. In both figures $\mathrm{a} > 0$ is a constant used to illustrate the scale differences.

The average sentence duration increase, as measured over the 170 sentences from set 39 through to set 55 in [27] is 16.9 % for continuous pause strengths and 26 % for binary pause strengths. Less than 2 % of all sentences were not allocated any pauses under the current value for $\mathcal{I}_b$. The effective pause durations increase with $\mathrm{RT_{60}}$. In the absence of reverberation, no pauses will be inserted.

A small number of sentences in the test data were not assigned any pauses under the current $\mathcal{I}_b$. Two examples are i) "*The clothes dried on a thin wooden rack.*" and ii) "*The beetle droned in the hot June sun.*"

4.2. Subjective evaluation

A listening test with twelve naive (no supervised training was performed) native English speakers (average age 24) was conducted to evaluate performance. The subjects did not report any hearing impairments and were paid for their participation. The material was presented diotically, in a soundproof booth using Sennheiser HD 558 headphones.

An initial session comprising ten sentences familiarized the listeners with the task and the test interface. Each method was assigned a macro set of four ten-sentence sets from [27]. The allocation of macro-set to system and the system presentation order were randomly selected for each listener. Upon hearing a sentence once, the listener was prompted to type its content.

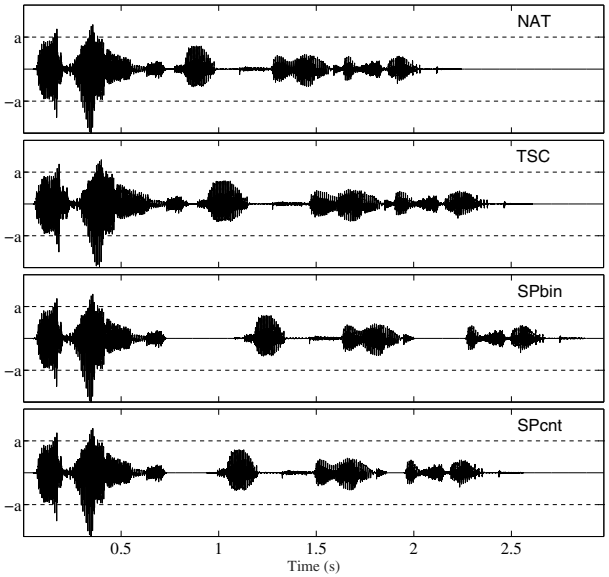

Figure 4: Signal waveforms for sentence A from Table 1 pre-processed for presentation at $\mathrm{RT}_{60} = 1.8$ s.

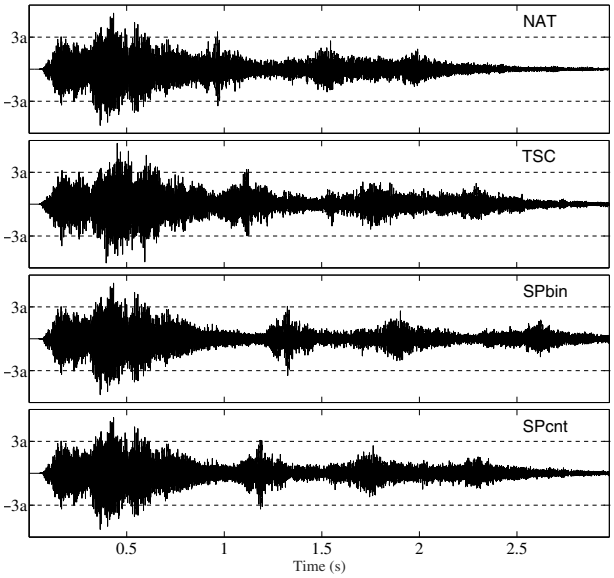

Figure 5: Processed reverberant waveforms for sentence A from Table 1 at $\mathrm{RT}_{60} = 1.8$ s.

A word recognition rate (WRR) was computed as the ratio of correctly-identified to the total number of key words in a sentence [29]. Per-subject WRRs over macro sets are given in Table 2. Mean WRRs, over all test participants, and standard errors for each method are shown in Figure 6.

The results show that both pause insertion and linear time scaling improve intelligibility significantly ($p < 0.01$, Student's t test) compared to un-modified speech. Continuous pause strengths achieve small intelligibility gain over binary pause strengths (at a significantly shorter average sentence duration). The insignificant advantage of $\mathrm{SP}_{\mathrm{cnt}}$ over TSC may increase following improvements in the pause predictor. We note that TSC cannot determine the degree of time-scaling independently but relies on $\mathrm{SP}_{\mathrm{cnt}}$ instead. In addition, quality issues related to scaling transient portions of the signal, e.g., position 0.2 s for TSC in Figure 4, were observed.

Two more factors were identified that limit the intelligibility gain from pause insertion for recorded speech in particular. Pause insertion without prosody modification leads to a deviation from natural speech and may confuse the listener. Word segmentation by forced alignment is not perfect and may introduce artifacts. The effect of these factors can be overcome by use of speech synthesis.

Table 2: Word recognition rates at $\mathrm{RT}_{60} = 1.8$ s.

Subject	NAT	TSC	$\mathrm{SP}_{\mathrm{bin}}$	$\mathrm{SP}_{\mathrm{cnt}}$
1	0.44	0.59	0.54	0.65
2	0.38	0.34	0.37	0.41
3	0.58	0.60	0.66	0.58
4	0.38	0.57	0.49	0.65
5	0.61	0.81	0.67	0.64
6	0.43	0.55	0.39	0.46
7	0.51	0.61	0.76	0.64
8	0.47	0.63	0.62	0.73
9	0.17	0.26	0.16	0.23
10	0.42	0.36	0.40	0.40
11	0.47	0.62	0.65	0.56
12	0.30	0.42	0.38	0.57
mean	**0.43**	**0.53**	**0.51**	**0.54**
std	**0.12**	**0.16**	**0.17**	**0.14**

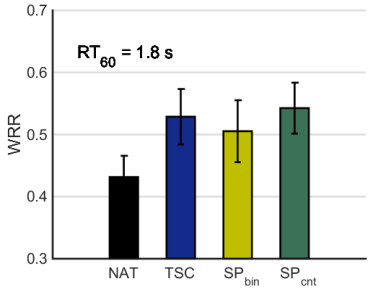

Figure 6: Mean WRRs and standard errors.

5. Conclusions

Pause insertion based on text analysis is an effective method for intelligibility enhancement of speech under reverberation. Advanced methodology allows for a significant intelligibility gain with modest duration increase. Use of a criterion that takes into account both the linguistic context and the specifics of the environment provides flexibility. Further sophistication of the method will focus on introducing dependence of the target signal-to-late-reverberation ratio on the phoneme type and improving the text analysis. The proposed method combines easily with other intelligibility-enhancing modifications.

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] E. M. Picou, J. Gordon, and T. A. Ricketts, "The Effects of Noise and Reverberation on Listening Effort in Adults with Normal Hearing," *Ear and Hearing*, vol. 37, no. 1, pp. 1–13, 2015.

[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] J. S. Bradley and H. Sato, "On the importance of early reflections for speech in rooms," *J. Acoust. Soc. Am*, vol. 6, no. 113, pp. 3233–3244, 2003.

[5] 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.

[6] 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.

[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] ——, "Speech Reinforcement with a Globally Optimized Perceptual Distortion Measure for Noisy Reverberant Channels," in *Proc. IWAENC*, 2014, pp. 89–93.

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

[10] 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.*, 2015.

[11] 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.

[12] 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.

[13] M. Tsuji, T. Arai, and K. Yasu, "Preprocessing using consonat emphasis and vowel suppression for improving speech intelligibility in reveberant environments," *J. Acoust. Soc. Japan*, vol. 69, no. 4, pp. 179–183, 2013.

[14] 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.

[15] 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," The Institute of Electr., Inf. & Comm. Eng., Tech. Rep., 2006.

[16] F. Sattar, M. Nilsson, and I. Claesson, "Segmentation and its real-world applications in speech processing," in *Intern. Symposium on Signal Processing and its Applications*, 2008.

[17] 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.

[18] P. J. Scharpff and V. van Heuven, "Effects of pause insertion on the intelligibility of low quality speech," in *Proc. FASE/Speech-88 Symposium*, 1988.

[19] S. Buchholz, N. Braunschweiler, M. Morita, and G. Webster, "The Toshiba entry for the Blizzard Challenge 2007," in *6th ISCA Sp. Synt. Workshop*, 2007, pp. 264–269.

[20] N. Braunschweiler and R. Maia, "Pause prediction from text for speech synthesis with user-definable pause insertion likelihood threshold," in *Proc. Interspeech*, 2016.

[21] T. Burrows, P. Jackson, K. Knill, and D. Sityaev, "Combining models of prosodic phrasing and pausing," in *Proceedings INTERSPEECH 2005 – 9th Annual Conference of the International Speech Communication Association, 4-8 September, Lisboa, Portugal*, 2005, pp. 1829–1832.

[22] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "ToBI: a standard for labelling English prosody," in *ICSLP*, 1992, pp. 12–16.

[23] M. Karjalainen and H. Järveläinen, "Reverberation Modeling Using Velvet Noise," in *Proc. AES 30th Int. Conf.*, 2007.

[24] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 3.4)*. Cambridge University Engineering Department, 2009.

[25] P. N. Petkov, G. E. Henter, and W. B. Kleijn, "Maximizing Phoneme Recognition Accuracy for Enhanced Speech Intelligibility in Noise," *IEEE Trans. Audio, Speech and Lang. Proc.*, vol. 21, no. 5, pp. 1035–1045, 2013.

[26] J. B. Allen and D. A. Berkley, "Image method for effciently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, 1979.

[27] "IEEE Recommended Practice for Speech Quality Measurements," *IEEE Trans. Audio and Electroacoustics*, vol. 17, pp. 225–246, 1969.

[28] 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.

[29] 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.