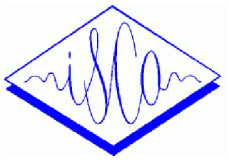

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Binary Mask Estimation for Improved Speech Intelligibility in Reverberant Environments

*Oldooz Hazrati, Jaewook Lee and Philipos Loizou**

Center for Robust Speech Systems (CRSS),
The University of Texas at Dallas, Richardson, TX 75080-3021, USA
{hazrati, jaewook, loizou}@utdallas.edu

10.21437/Interspeech.2012-57

Abstract

A blind (non-ideal) time-frequency (T-F) masking technique is proposed for suppressing reverberation. A binary mask is estimated at each T-F unit by extracting a single variance-based feature from the reverberant signal and comparing its value against an adaptive threshold. The performance of the estimated binary mask is evaluated using intelligibility listening tests with hearing impaired listeners in four moderate to highly reverberant conditions. Results indicated that the proposed T-F masking technique yielded significant improvements in intelligibility even in highly reverberant conditions (T_{60} = 1.0 s). This improvement was attributed to the recovery of the vowel/consonant boundaries which are severely smeared in reverberation.

Index Terms: Binary mask, cochlear implant (CI), dereverberation

1. Introduction

Several speech dereverberation techniques have been proposed, some of which consist of multiple stages treating early and late reverberations differently [1]. Inverse filtering is one of the commonly used techniques for speech dereverberation which removes the reverberation by passing the reverberant signal through a finite impulse response (FIR) filter [2]. With the use of multiple microphones, an exact inverse of the room impulse response (RIR) can be obtained assuming there are no common zeros among the RIRs. Although less effective, single microphone dereverberation algorithms are usually more practical and desirable. A few examples of such algorithms are spectral subtraction [3], and excitation source information based [4] techniques. However, these techniques are still far from perfect, and some do not result in acceptable performance in practice.

In this study, an alternative dereverberation algorithm based on time-frequency (T-F) masking is proposed. T-F masking algorithms decompose the signal into T-F units and select T-F units satisfying a given criterion (e.g., SNR$>$ 0 dB, for noise suppression), while discarding the rest. This is typically achieved by applying a binary mask to the T-F units of the decomposed signal, i.e., a given T-F unit is set to 0 if it does not satisfy a given criterion or is set to 1 if it satisfies the criterion [5]. Binary masks have been widely used for different speech enhancement and sound separation applications resulting in gains in intelligibility and quality of the processed noisy speech [5]-[7]. Only a few studies, however, have evaluated the use of appropriate binary masks for suppressing reverberation (and additive noise) [8]. Use of the binary masks for dereverberation is attractive as it does not rely on the inversion of the RIR. Palomaki et al. [8] has evaluated a reverberant (a priori) binary mask; however, the mask was not clearly defined and was evaluated primarily by ASR systems rather than by human listeners. Mandel et al. [9] evaluated a number of oracle reverberant (binary and soft) masks using source-separation algorithms and human listeners. All masks were constructed based on several combinations of the ratio of the target direct signal energy to either the target late-reverberant signal energy and/or the masker (direct-path and) late-reverberant energy. All masks, however, were based on the decomposition of the reverberant signal to its direct-path, early echo and late reflections components, i.e., they assumed access to the RIR.

Motivated by the intelligibility gains obtained from the ideal reverberant mask (IRM) [10], a blind (non-ideal) T-F masking technique is proposed for improving the intelligibility of reverberant speech. A nonparametric and unsupervised method of automatic threshold selection that has been used originally for image segmentation is adopted as the local criterion in making decisions for each T-F unit. For each T-F unit, a feature based on the ratio of signal variances is computed and the local criterion is constructed using this feature. Intelligibility listening tests were conducted to assess the performance of the proposed dereverberation algorithm. The challenge faced with evaluating the intelligibility of dereverberated speech with normal-hearing listeners is that they generally perform extremely well (near 100%) even in highly reverberant environments. Hearing-impaired listeners, on the other hand, perform poorly, even at moderate T_{60} values, when presented with reverberant speech. For that reason,

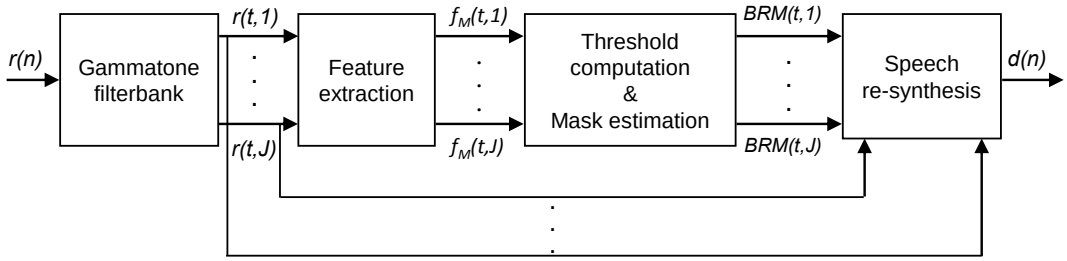

Figure 1: *Block diagram of the proposed blind reverberant mask (BRM) estimation technique for dereverberation.*

in the present study we are using hearing-impaired listeners to evaluate the proposed dereverberation algorithm.

2. Time-frequency masking of reverberant speech

The block diagram of the proposed blind time-frequency masking algorithm is shown in Figure 1. First the reverberant speech, $r(n)$, is passed through a set of gammatone filters yielding $r(t,j)$, where j indicates the band index. In each band, a feature, $f_M(t,j)$, is extracted, and compared to an adaptive threshold. A binary mask is estimated, BRM(t,j), and applied to the subband signal $r(t,j)$. Finally, the signals in each subband are time-reversed, passed through the gammatone filter, time reversed again and summed across all subbands to obtain the dereverberated speech.

2.1. Feature extraction

The proposed feature is used to identify the peaks and valleys in each band. This feature is computed as the ratio of the variance of the signal raised to a power and the variance of the absolute value of the signal. Accordingly, the feature is computed as follows:

$$f_M(t,j) = 10\log_{10}\left(\frac{\sigma^2_{r'}(t,j)}{\sigma^2_{|r|}(t,j)}\right) \quad (1)$$

where $r'(t,j) = (r(t,j))^{\alpha}$, and $|r(t,j)|$ is the absolute value of the L (frame size) dimensional reverberant vector in frame t, and frequency band j. The features are smoothed across time using a 3-point median filter.

2.2. Binary Mask Estimation

In order to make a decision on the features extracted using (1), as to whether they are reverberation-dominant or reverberation-free, they are compared against a threshold. Here, a nonparametric and unsupervised method for automatic threshold estimation is adopted [11]. The input to this histogram-based threshold estimation technique at time frame t and frequency band j is the following feature vector containing features of L_p previous and L_f future frames:

$$f_{hist}(t,j) = \{f_M(t-L_p,j), ..., f_M(t+L_f,j)\} \quad (2)$$

The optimum threshold level tr^* is obtained from the index tr that maximizes $\sigma^2_B(tr)$:

$$\sigma^2_B(tr^*) = \max_{tr=1,...,Tr}(\sigma^2_B(tr)) \quad (3)$$

where Tr is the total number of distinct levels of the histogram of the input feature vector (f_{hist}) and $\sigma^2_B(tr)$ is the between-class variance with distinct intensity level index, tr, which is expressed as follows:

$$\sigma^2_B(tr) = \frac{(m_G.P_s(tr) - m(tr))^2}{P_s(tr)(1-P_s(tr))} \quad (4)$$

where the global intensity mean, m_G, the cumulative mean, $m(tr)$, and the cumulative sum, $P_s(tr)$, are defined as,

$$m_G = \sum_{i=1}^{Tr} i.p_i, \quad m(tr) = \sum_{i=1}^{tr} i.p_i,$$

$$P_s(tr) = \sum_{i=1}^{tr} p_i.$$

where p_i denotes the normalized histogram of the feature vector (Eq. 2).

If the long-term windowed feature vector contains only silence, the algorithm will compute inaccurate threshold levels resulting in incorrect decisions. Therefore, a minimum threshold level (tr_0) is considered to discriminate silence from speech. Use of the long-term windowed feature vectors along with tr_0 results in a robust and effective adaptive threshold level estimation. In the T-F classification stage, if the value of the feature of a T-F unit is greater than the adaptive threshold of that specific T-F unit, the frame is classified as reverberation-free, otherwise it is considered as reverberation-dominant. Frames classified as reverberation-free are retained, while reverberation-dominant frames are zeroed out. This forms a binary reverberant mask which is defined as:

$$BRM\,(t,j) = \begin{cases} 1, & f_M\,(t,j) > max(tr^*(t,j), tr_0) \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

where $f_M(t,j)$ is the feature extracted as in (1). Note that this technique removes the reverberation-dominant T-F units resulting in restoration of the word/syllable boundaries. Our hypothesis is that hearing-impaired listeners

need to have clear access to the location of those boundaries for good speech recognition in reverberant environments.

3. Experiments

3.1. Algorithmic parameters

A total of 64 4-th order gammatone filters are used to divide the signal bandwidth of 50-8000 Hz into quasi-logarithmically spaced bands. In the feature extraction stage, features from 10 previous and 2 future 20-ms frames with a 50% overlap between adjacent frames, were used for the histogram calculation. The parameter α in (1) was set to 2.1. A median filter of order 3 is applied to the features in order to smooth any abrupt changes.

3.2. Speech material and subjective listening tests

The proposed algorithm is evaluated using phonetically-balanced sentences taken from the IEEE database [12]. The sentences were originally sampled at 25 kHz, and downsampled to 16 kHz for our tests.

The reverberant stimuli are generated by convolving the clean signals with real RIRs recorded in a 10.06 m $\times$ 6.65 m $\times$ 3.4 m (length $\times$ width $\times$ height) room [13]. The reverberation time of the room is varied from 0.3 s to 0.6, 0.8, and 1.0 s by adding absorptive panels to the walls and floor carpeting. The direct-to-reverberant ratios (DRR) of the RIRs are 1.5, -1.8, -3.0, and -0.5 dB for T_{60} = 0.3, 0.6, 0.8, and 1.0 s, respectively. The distance between the single-source signal and the microphone is 5.5 m, which is beyond the critical distance.

A total of six hearing-impaired listeners with cochlear implants (Nucleus device) were recruited for testing. Listeners were tested using a PDA-based cochlear implant research platform [14]. The dereverberated signals were streamed off-line via the PDA platform and sent directly to the subject's cochlear implant. The PDA processor was programmed for individual subjects using their threshold and comfortable loudness levels, and coding strategy parameters. All experiments were conducted in a sound proof booth. Twenty IEEE sentences were presented directly to the implant users at a comfortable level. The order of the conditions and sentences presented to the CI users was randomized across subjects.

4. Results and discussion

Table 1 shows the intelligibility scores, averaged across six CI users, in terms of the mean percentage of words identified correctly. The results obtained from the proposed blind reverberant mask (BRM) are compared against those obtained by testing the subjects with the unprocessed reverberant stimuli in four moderate to highly reverberant conditions (T_{60} = 0.3, 0.6, 0.8, and 1.0 s). Scores obtained with the ideal reverberant mask [10] are also given for comparison to provide the upper bound in performance. The average intelligibility score obtained in anechoic quiet condition was 82.6%.

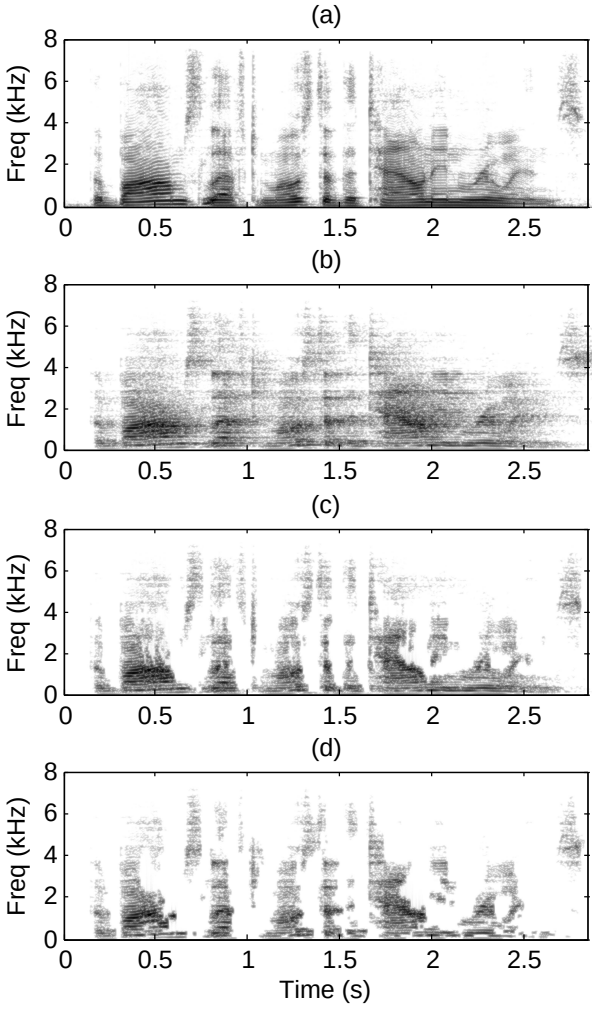

Figure 2: *Spectrograms of the sentence "The stitch will serve but needs to be shortened", for (a) clean, (b) reverberant (T_{60} = 0.6 s), (c) IRM processed, and (d) BRM processed stimuli.*

As results indicate, even with moderate amounts of reverberation (T_{60} = 0.3 s), intelligibility scores drop (relative to the anechoic conditions) by 27.2%. After applying the BRM to the reverberant signals, the intelligibility scores improved by 2.9%, 23.7%, 27.0%, and 27.1% in T_{60} = 0.3, 0.6, 0.8, and 1.0 s conditions, respectively.

Table 1: *Average word recognition scores of six CI users.*

Condition	T_{60} (s)			
	0.3	0.6	0.8	1.0
Reverberant	55.4	26.5	24.7	11.8
IRM	70.1	75.3	75.8	73.3
BRM	58.3	50.2	51.7	38.9

These improvements are found to be statistically significant ($p < 0.05$) at T_{60} = 0.6, 0.8, and 1.0 s. Although the IRM algorithm produces higher intelligibility gains, the BRM method still provides significant intelligibility improvement.

In order to compare the performance of BRM with that of the ideal reverberant mask (IRM) [10], spectrograms of an IEEE sentence in T_{60} = 0.6 s are presented in Figure 2. As can be seen from Figure 2, the vowel/consonant boundaries are obscured and the gaps between vowels and consonants are filled with reverberant energy (see Figure 2b). This is particularly more evident in unvoiced segments of speech, and consequently causes a decrease in intelligibility [10]. As shown in panel (c), after applying the IRM to the reverberant signal, the vowel/consonant boundaries and gaps previously filled with reverberant energy are recovered, resulting in improvement in intelligibility. Comparing the spectrogram of the BRM processed reverberant speech (panel (d)) with that of the IRM processed (panel (c)), it is evident that the vowel/consonants boundaries and gaps are restored to a great extent. The effectiveness of the proposed BRM in the time domain is also demonstrated in Figure 3. The figure shows bandpass filtered signal of the same IEEE sentence (Figure 2) at f_c = 1 kHz for anechoic, reverberant (T_{60} = 0.6 s), and BRM-processed signals. Comparing the BRM processed (panel (c)) with the anechoic and reverberant signals (panels (a) and (b)), we observe that the proposed BRM technique restores the vowel/consonant boundaries.

5. Conclusions

The present study proposed a blind reverberant mask (BRM) algorithm for improving intelligibility of reverberated speech for hearing-impaired listeners. This technique uses the proposed feature (Eq. 1) along with a nonparametric and unsupervised threshold estimation method to classify the T-F units to reverberation-dominant or reverberation-free units. Reverberation was suppressed by retaining only the units that were classified as reverberation-free. Performance of the proposed technique was assessed through listening tests conducted with six hearing impaired listeners with cochlear implants. Listening tests indicated significant improvements in intelligibility in highly reverberant conditions (T_{60} = 0.6, 0.8, and 1.0 s). This improvement was attributed to the recovery of the vowel/consonant boundaries, which are often blurred in reverberation owing to the late reflections.

6. Acknowledgements

The authors would like to thank the CI users for their time, and also Dr. Neuman of the NYU Langone Medical Center for providing the RIRs used in this study. Research supported by NIDCD/NIH.

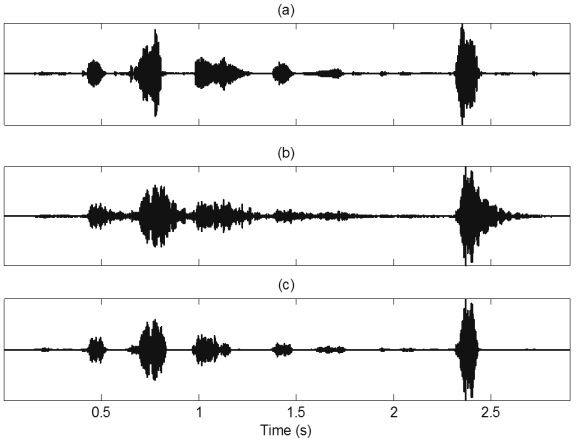

Figure 3: *Band-pass filtered (f_c = 1 kHz) waveforms of the sentence: "The stitch will serve but needs to be shortened", for: (a) clean, (b) reverberant (T_{60} = 0.6 s), and (c) BRM processed.*

7. References

[1] Furuya, K. and Kataoka, A., "Robust speech dereverberation using multichannel blind deconvolution with spectral subtraction", *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, pp. 1579-1591, 2007.

[2] Miyoshi, M. and Kaneda, Y., "Inverse filtering of room acoustics", *IEEE Trans. Speech, Audio. Process.*, vol. 36, pp. 145-152, 1988.

[3] Lebart, K., Boucher, J.M., and Denbigh, P.N., "A new method based on spectral subtraction for speech dereverberation", *Acta Acoust.*, vol. 83, 359-366, 1997.

[4] Yegnarayana, B., Murthy, P.S., "Enhancement of reverberant speech using LP residual signal", *IEEE Trans. Speech, Audio. Process*. vol. 8, pp. 267-281, 2000.

[5] Wang, D.L., Brown, G.J., Computational Auditory Scene Analysis: Principles, Algorithms, and Applications. Wiley, New York, 2006.

[6] Kim, G., Lu, Y., Hu, Y., and Loizou, P.C., "An algorithm that improves speech intelligibility in noise for normal-hearing listeners", *J. Acoust. Soc. Am.*, vol. 126, pp. 3387-3405, 2009.

[7] Li, N., and Loizou, P.C., "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction", *J. Acoust. Soc. Am.*, vol. 123, pp. 1673-1682, 2008.

[8] Palomaki, K.J., Brown, G.J. and Wang, D.L.,"Robust speech dereverberation using multichannel blind deconvolution with spectral subtraction", *Speech Commun*., vol. 43, pp. 361-378, 2004.

[9] Mandel, M.I., Bressler, S., Shinn-Cunningham, B., and Ellis, D.P.W., "Evaluating source separation algorithms with reverberant speech", *IEEE Trans. Audio, Speech, Lang. Process.* vol. 18, pp. 1872-1883, 2010.

[10] Kokkinakis, K., Hazrati, O., and Loizou, P.C., "A channel-selection criterion for suppressing reverberation in cochlear implants", *J. Acoust. Soc. Am.*, vol. 129, pp. 3221-3232, 2011.

[11] Otsu, N., " A Threshold Selection Method from Gray-Level Histograms", *IEEE Trans. Systems, Man, Cybernetics*., Vol. 9, pp. 62-66, 1979.

[12] IEEE, "IEEE recommended practice for speech quality measurements", *IEEE Trans. Audio Electroacoust.*, AU-17, pp. 225-246, 1969.

[13] Neuman, A. C., Wroblewski, M., Hajicek, J., and Rubinstein, A.,"Combined effects of noise and reverberation on speech recognition performance of normal-hearing children and adults", *Ear Hear.*, vol. 31, pp. 336-344, 2010.

[14] Ali, H., Lobo, A. P., and Loizou, P. C., "A PDA platform for offline processing and streaming of stimuli for cochlear implant research", in *Proc. IEEE EMBS.*, Boston, MA, Aug. 2011, pp. 1045-1048.