ROBUST RECOGNITION OF REVERBERANT AND NOISY SPEECH USING COHERENCE-BASED PROCESSING

Anjali Menon[1], *Chanwoo Kim*[2], *Richard M. Stern* [1]

[1]Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA, USA
[2]Samsung Research, Seoul, Republic of Korea

ABSTRACT

This paper describes a combination of techniques for improving speech recognition accuracy using two microphones in reverberant and noisy environments. These techniques include both monaural and binaural processing. The first stage is monaural precedence-based processing that enhances the onsets of the incoming speech signal, and hence suppresses later components that are more affected by reverberation. Onset enhancement has been shown to be useful to the human auditory system in separating the direct field from the reverberant field in reverberant environments. The second stage applies emphasis or suppression to signal components based on an estimation of the inter-microphone coherence of the incoming speech signal. Specifically, portions of the speech signal that are less coherent are suppressed, which is intended to reduce the contributions of components that are dominated by diffuse noise or high degrees of reverberation in the input signal. A combination of these techniques is shown to lead to significant improvements in speech recognition accuracy. A DNN-based automatic speech recognition system was used to evaluate the techniques described in this study over a range of reverberation times and signal-to-interferer ratios.

Index Terms— speech recognition, binaural hearing, onset enhancement, interaural coherence, reverberation

1. INTRODUCTION

Room reverberation is a significant impediment to robust automatic speech recognition (ASR), and the presence of competing talkers inevitably worsens recognition accuracy. Robust speech recognition in challenging environments is especially important because of the widespread use of voice-controlled devices such as smart loud speakers, home assistants etc. that operate at a distance from the speaker. While methods that are based on Interaural Time Differences (ITDs) have enjoyed some success in source separation in the presence of multiple talkers, the presence of reverberation degrades the accuracy with which ITD information can be extracted (*e.g.* [1]).

One proposed solution for the dereverberation problem has been the use of inverse filtering to reduce early reverberant components followed by spectral subtraction for later-arriving components (*e.g.* [2]). A similar solution was proposed in [3] but with the addition of spatio-temporal averaging. With the popularity of deep neural networks, the use of deep autoencoders for dereverberation has also yielded promising results (*e.g.* [4]). One technique used in the present study is based on human auditory processing and emulates some attributes of the "precedence effect" [5, 6, 7]. The precedence effect describes the phenomenon in which directional cues representing the first-arriving wavefront (which correspond to the direct component) are given greater perceptual weighting than those cues that arise as a consequence of subsequent reflected sounds. The algorithm called *Suppression of Slowly-varying components and the Falling edge of the power envelope* (SSF) [8, 9] was motivated by this principle and has been successful in improving ASR accuracy in reverberant environments.

The methods discussed above are all monaural and do not take any spatial information into account. In the presence of multiple microphones where spatial information is available, one approach has been to characterize the extent to which each portion of the speech signal is dominated by coherent versus diffuse energy. For example, the technique proposed in [10] uses spectral subtraction initially for suppression of late reverberations followed by coherence-based processing.

In this paper, we discuss the use of a combination of techniques that use steady-state suppression initially to achieve onset enhancement. This is followed by a second stage introduced in this paper that uses inter-microphone coherence to suppress other residual reverberant and non-coherent components. Section 2 describes these approaches. The results obtained from ASR experiments using these techniques are discussed in Section 3.

2. STEADY-STATE SUPPRESSION AND COHERENCE-BASED PROCESSING

This paper addresses binaural processing in adverse conditions, which include reverberation and interfering talkers. The approaches described assume that recordings are made with two microphones as shown in Figure 1. The two microphones are placed in a reverberant room with the target talker directly in front of them. This does not limit the generality of the ap-

978-1-5386-4658-8/18/$31.00 ©2019 IEEE

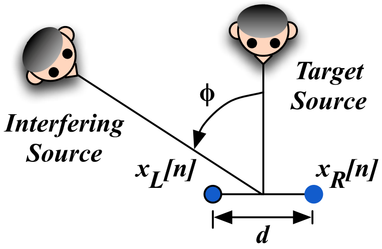


Fig. 1. Two-microphone recording with an on-axis target source and off-axis interfering source used in this study.

proach, as steering delays can easily be inserted to rotate the "look direction" to a different azimuth (*e.g.* [11]). An interfering talker is also present, located at an angle of ϕ with respect to the two microphones.

As noted above, in our most effective form of signal processing, the incoming signal is processed in two stages. First, we invoke SSF processing at the monaural level, which enhances onsets and suppresses steady-state components in each subband. Second, we apply *Coherent-to-Diffuse-Ratio-based Weighting* (CDRW) which emphasizes or suppresses components based on the extent to which they are binaurally coherent. We describe our application of SSF and CDRW in Sections 2.1 and 2.2, respectively.

2.1. Steady-state suppression

Steady-state suppression can vastly improve ASR accuracy in the presence of reverberation. It aims at boosting the parts of the input signal that are believed to correspond to the direct sound, which indirectly suppresses reflected sounds. While the use of steady-state suppression was originally motivated by the onset enhancement implied by the precedence effect [7], it can also be applied at the monaural level (*e.g.* [12]). In this paper we use the SSF algorithm as formulated by Kim *et al.* [8, 9] to achieve steady-state suppression.

As described in [8], the SSF algorithm decomposes the input signal into 40 frequency channels. In each channel the frame-level power is computed and then lowpass filtered in nonlinear fashion to ensure that the output remains positive. This lowpass-filtered representation of the short-time power is subtracted from the original power contour to obtain the processed power. A weighting coefficient is then computed by taking the ratio of the processed power to the original power. A set of spectral weighting coefficients are then derived from these weights. The spectral weighting coefficients, in turn, are multiplied by the spectrum of the original input signal to produce the processed signal. This suppression of the falling edge of the power contour is highly effective in improving ASR performance in reverberant environments, as seen in [13].

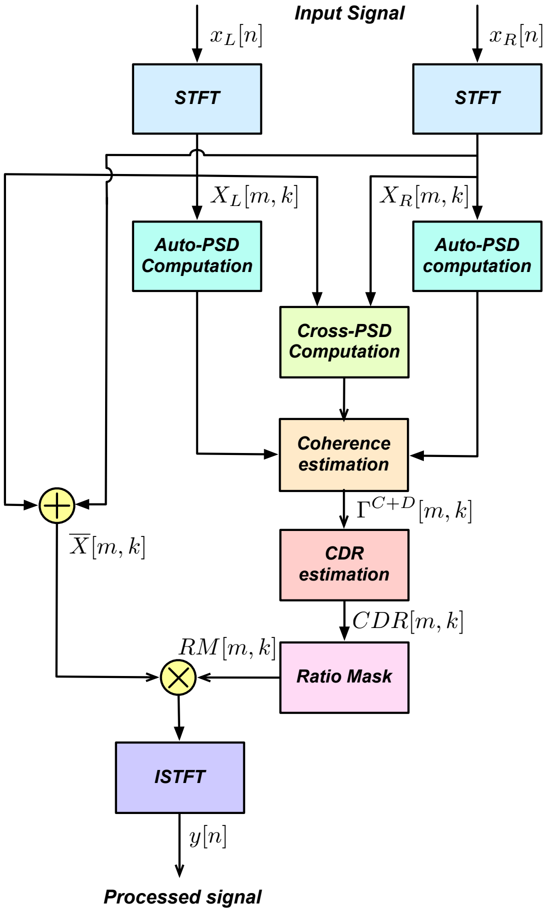


Fig. 2. Block diagram describing the CDRW algorithm.

2.2. Coherent-to-Diffuse Ratio based Weighting (CDRW)

The Coherent-to-Diffuse Ratio-based Weighting algorithm (CDRW) is based on the principle that sounds from the target source as in Figure 1 would be expected to perfectly coherent across the two microphones. The presence of reverberation produces a more diffuse noise field. The technique used in this study determines the degree of coherence between the two microphone signals. With this knowledge, it is possible to apply a mask on the input signal that suppresses regions where the the ratio of coherent-to-diffuse energy is low.

Several approaches that have been proposed to estimate the ratio of coherent energy to diffuse energy in a given acoustic environment [14, 15, 16, 17]. We estimate the Coherent-to-Diffuse Ratio (CDR) using the method proposed by Jeub *et al.* [14]. A block diagram describing the Coherent-to-Diffuse Ratio-based Weighting (CDRW) introduced in this paper is shown in Figure 2. While previous studies have used CDR-based measures for true CDR estimation or speech enhancement, to our knowledge this is the first study that considers the efficacy of a CDR-based algorithm in improving speech recognition accuracy in the peer-reviewed literature.

Consider two signals from the microphones $x_R[n]$ and $x_L[n]$ (where n denotes the time index) that have Short-Time Fourier Transforms (STFT) for the m^{th} frame and k^{th} frequency index $X_R[m,k]$ and $X_L[m,k]$ respectively. The complex inter-microphone coherence $\Gamma_{x_Rx_L}[m,k]$ is given by

$$\Gamma_{x_Rx_L}[m,k] = \frac{\Phi_{x_Rx_L}[m,k]}{\sqrt{\Phi_{x_Rx_R}[m,k]\Phi_{x_Lx_L}[m,k]}} \quad (1)$$

where $\Phi_{x_Rx_L}[m,k]$ denotes the Cross-Power Spectral Density (Cross-PSD) of $x_R[n]$ and $x_L[n]$ and $\Phi_{x_Rx_R}[m,k]$ and $\Phi_{x_Lx_L}[m,k]$ denote the Auto-PSD of $x_R[n]$ and $x_L[n]$ respectively. Since longer analysis windows have been shown to be better for estimating the power spectral density of noise [8], a window size of 80 ms with a 50% overlap was used. The Auto-PSD and Cross-PSD functions can be estimated using recursive averaging.

In the case of a diffuse field, as is expected to be caused by reverberation, the spherically isotropic inter-microphone coherence can be calculated by integrating all the plane waves originating from a surface area over the whole surface area of a sphere [18] which results in the expression below.

$$\Gamma^D_{x_Rx_L}[m,k] = sinc\left(\frac{2\pi k f_s d_{mic}}{Nc}\right) \quad (2)$$

where f_s is the sampling frequency, d_{mic} is the distance between the two microphones, N is the total number of frequency channels in the STFT and c is the speed of sound.

In contrast, in the case of a coherent source with the signal arriving at some angle θ, the inter-microphone coherence is

$$\Gamma^C_{x_Rx_L}[m,k] = e^{-\left(\frac{j2\pi k f_s d_{mic} cos(\theta)}{Nc}\right)} \quad (3)$$

since the signals are only a separated by a shift in phase.

In the case of an environment that is a mix of coherent and diffuse sources, the inter-microphone coherence can be derived by summing over the auto- and cross-PSDs of each source separately as seen in [14] which results in,

$$\Gamma^{C+D}_{x_Rx_L}[m,k] = \frac{\Phi^C[m,k] + \Phi^D[m,k] sinc\left(\frac{2\pi k f_s d_{mic}}{Nc}\right)}{\Phi^C[m,k] + \Phi^D[m,k]} \quad (4)$$

where the auto-PSD corresponding to the coherent source is given by $\Phi^C_{x_Rx_R}[m,k] = \Phi^C_{x_Lx_L}[m,k] = \Phi^C[m,k]$. Similarly, the auto-PSD corresponding to a diffuse source is given by $\Phi^D_{x_Rx_R}[m,k] = \Phi^D_{x_Lx_L}[m,k] = \Phi^D[m,k]$. The angle θ was assumed to be $\pi/2$.

As above, CDR is defined as the ratio of the coherent energy to the diffuse energy in a given environment.

$$CDR[m,k] = \frac{\Phi^C[m,k]}{\Phi^D[m,k]} \quad (5)$$

Substituting the expression for CDR into Eq. (4) and rearranging the terms gives us the real-valued CDR [14],

$$CDR[m,k] = max\left(0, \frac{sinc\left(\frac{2\pi k f_s d_{mic}}{Nc}\right) - Re\{\Gamma^{C+D}_{x_Rx_L}[m,k]\}}{Re\{\Gamma^{C+D}_{x_Rx_L}[m,k]\} - 1}\right) \quad (6)$$

Equation (6), as derived in [14], is useful in separating portions of the signal STFT that are dominated by the diffuse noise and therefore need to be suppressed. The quantity of $\Gamma^{C+D}_{x_Rx_L}[m,k]$ is also estimated using recursive smoothing using the smoothing factor $\alpha_C = 0.25$. α_C was determined experimentally.

In this study, our goal is to use a CDR-based weight for improved ASR accuracy. We use the classical Wiener filter to derive a ratio mask from the CDR measure as shown below:

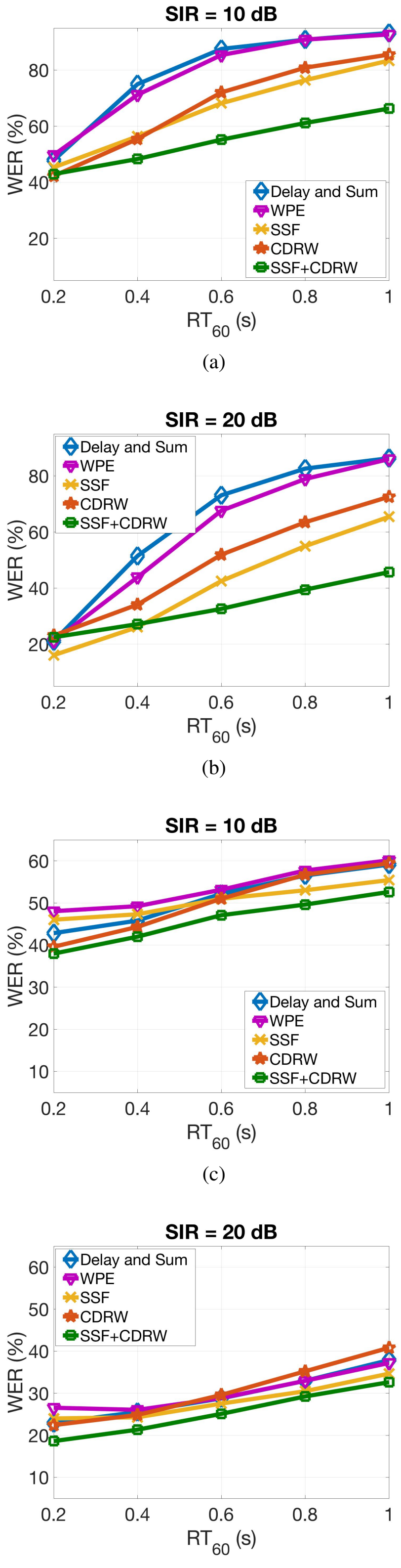
$$RM[m,k] = \frac{CDR[m,k]}{CDR[m,k] + 1} \quad (7)$$

The ratio mask $RM[m,k]$ is applied to the STFT of the mean of the two microphone inputs and an Inverse STFT (ISTFT) is then performed to obtain the processed waveform.

3. EXPERIMENTAL RESULTS

ASR experiments were conducted using the Kaldi speech recognition toolkit [19] and the Wall Street Journal (WSJ) database [20]. We used the WSJ SI84 training set, which consisted of 7138 utterances, along with the WSJ-5K test set, which consisted of 330 utterances. Acoustic models were trained using HMM-DNNs. Each DNN in the HMM-DNN system has 2 hidden layers. The HMM-DNN systems were trained using alignments from an HMM-GMM system trained with the same data. In turn, the HMM-GMM systems were trained by using mel-frequency cepstral coefficients (MFCC) features. The standard lexicon consisting of 5k words and a trigram language model were used for decoding.

We used the RIR simulation package [21] which implements the well-known image method [22] to simulate speech corrupted by reverberation. For the RIR simulations, we used a room of dimensions $5m \times 4m \times 3m$. The distance between the two microphones is 4 cm. The target speaker was located 2 m away from the microphones along the perpendicular bisector of the line connecting the two microphones. An interfering speaker was located at an angle of 45 degrees to one side and 2 m away from the microphones. The microphones and speakers were 1.1 m above the floor. To prevent any artifacts from standing-wave phenomena that create peaks and nulls in response at particular locations, the whole configuration described above was moved around in the room to several randomly-selected locations such that neither the speakers nor the microphones were placed less than 0.5 m from any of the walls. The target and interfering speaker signals were mixed at different levels after simulating reverberation.



(a)

(b)

(c)

(d)

Fig. 3. ASR results for the WSJ database at various reverberation times using a) clean training data and test data at 10 dB SIR b) clean training data and test data at 20 dB SIR c) reverberated training data with and test data at 10 dB SIR e) reverberated training data and test data at 20 dB SIR

Acoustic models were trained using both clean speech and speech with reverberation. When the SSF or CDRW algorithm or their combination was being tested, the training data underwent processing identical to the test data. The training data with reverberation used in this study had roughly equal number of utterances at reverberation times of 0.25, 0.5 and 0.75 s. The location of the microphone setup was randomized for each utterance. For the test data, reverberated speech with interfering talkers mixed in at Signal-to-Interference Ratio (SIR) of 10 dB and 20 dB was used. In the case of the test data, 25 different microphone locations were randomly chosen in the room to simulate speech corrupted by reverberation and an interfering talker. For both the test and training data, the relative positions of the microphones w.r.t the target speaker remained the same.

Results obtained using the (monaural) SSF algorithm alone are compared to results using CDRW and the combination of SSF+CDRW in Figures 3a-3d. Results using the Delay-and-Sum algorithm as well as the Weighted Prediction Error (WPE) algorithm [23] are also reported. The WPE algorithm was applied monaurally. Figures 3a and 3b describe results obtained using clean training data while Figures 3c and 3d used training data with reverberation.

When training using clean speech, the addition of the CDRW algorithm provides very significant improvements, as seen in Figure 3a and 3b, and the SSF+CDRW algorithm is almost always the configuration that provides the lowest word error rate. With the exception of the lowest reverberation time ($RT_{60} = 0.2$ s), the relative improvement provided by the SSF+CDRW algorithm compared to SSF alone increases with increasing reverberation time leading to as much as a 30% relative improvement in WER at $RT_{60} = 1$ s and 20 dB SIR.

When the system is trained on reverberated speech, the overall performance of the baseline configurations improves significantly, which reduces somewhat the relative improvement provided by the SSF+CDRW algorithm. SSF+CDRW again consistently provides the lowest WER. The average drop in WER for both 10 and 20 dB SIR, is on average close to 10%.

4. SUMMARY

We demonstrate in this paper that the combination of steady-state suppression (SSF) and coherence-based weighting (CDRW) provides improved ASR accuracy compared to the use of SSF or CDRW (or WPE) alone. These approaches to robustness based on traditional signal processing provide improvements that are significant and consistent.

5. ACKNOWLEDGEMENTS

This research was supported by the Jack and Mildred Bowers Scholarship in Engineering.

6. REFERENCES

[1] C. Kim, K. Kumar, B. Raj, and R. M. Stern, "Signal separation for robust speech recognition based on phase difference information obtained in the frequency domain." in *INTERSPEECH*. Citeseer, 2009, pp. 2495–2498.

[2] M. Wu and D. Wang, "A two-stage algorithm for one-microphone reverberant speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 774–784, 2006.

[3] N. D. Gaubitch, E. A. Habets, and P. A. Naylor, "Multimicrophone speech dereverberation using spatiotemporal and spectral processing," in *Circuits and Systems, 2008. ISCAS 2008. IEEE International Symposium on*. IEEE, 2008, pp. 3222–3225.

[4] X. Feng, Y. Zhang, and J. Glass, "Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 1759–1763.

[5] H. Wallach, E. B. Newman, and M. R. Rosenzweig, "The precedence effect in sound localization (tutorial reprint)," *Journal of the Audio Engineering Society*, vol. 21, no. 10, pp. 817–826, 1973.

[6] R. Y. Litovsky, H. S. Colburn, W. A. Yost, and S. J. Guzman, "The precedence effect," *The Journal of the Acoustical Society of America*, vol. 106, no. 4, pp. 1633–1654, 1999.

[7] P. M. Zurek, "The precedence effect," in *Directional hearing*. Springer, 1987, pp. 85–105.

[8] C. Kim and R. M. Stern, "Nonlinear enhancement of onset for robust speech recognition." in *INTERSPEECH*, 2010, pp. 2058–2061.

[9] C. Kim, "Signal processing for robust speech recognition motivated by auditory processing," Ph.D. dissertation, Carnegie Mellon University, 2010.

[10] M. Jeub, M. Schafer, T. Esch, and P. Vary, "Model-based dereverberation preserving binaural cues," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1732–1745, 2010.

[11] B. Widrow and P. N. Stearns, *Adaptive Signal Processing*. Prentice-Hall, 1985.

[12] K. D. Martin, "Echo suppression in a computational model of the precedence effect," in *Applications of Signal Processing to Audio and Acoustics, 1997. 1997 IEEE ASSP Workshop on*. IEEE, 1997, pp. 4–pp.

[13] R. M. Stern, C. Kim, A. Moghimi, and A. Menon, "Binaural technology and automatic speech recognition," in *International Congress on Acoustics*, 2016.

[14] M. Jeub, C. Nelke, C. Beaugeant, and P. Vary, "Blind estimation of the coherent-to-diffuse energy ratio from noisy speech signals," in *Signal Processing Conference, 2011 19th European*. IEEE, 2011, pp. 1347–1351.

[15] J. Allen, D. Berkley, and J. Blauert, "Multimicrophone signal-processing technique to remove room reverberation from speech signals," *The Journal of the Acoustical Society of America*, vol. 62, no. 4, pp. 912–915, 1977.

[16] O. Thiergart, G. Del Galdo, and E. A. Habets, "Signal-to-reverberant ratio estimation based on the complex spatial coherence between omnidirectional microphones," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 309–312.

[17] A. Westermann, J. M. Buchholz, and T. Dau, "Binaural dereverberation based on interaural coherence histograms a," *The Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. 2767–2777, 2013.

[18] E. A. Habets and S. Gannot, "Generating sensor signals in isotropic noise fields," *The Journal of the Acoustical Society of America*, vol. 122, no. 6, pp. 3464–3470, 2007.

[19] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[20] D. B. Paul and J. M. Baker, "The design for the wall street journal-based csr corpus," in *Proceedings of the workshop on Speech and Natural Language*. Association for Computational Linguistics, 1992, pp. 357–362.

[21] S. G. McGovern, "A model for room acoustics," 2003.

[22] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[23] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1717–1731, 2010.