MISSING FEATURE SPEECH RECOGNITION USING DEREVERBERATION AND ECHO SUPPRESSION IN REVERBERANT ENVIRONMENTS

Hyung-Min Park and Richard M. Stern

Language Technologies Institute and Department of Electrical and Computer Engineering
Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15213, U.S.A.
{hmpark, rms}@cs.cmu.edu

ABSTRACT

This paper describes an algorithm that efficiently segregates desired speech features from spatially-separated interfering sources in reverberant environments. Although most binaural segregation techniques successfully remove interference components in the absence of reverberation, source segregation in reverberant environments remains a challenging problem. In order to reduce the effects of reverberation, we present a method that dereverberates input signals before they are segregated. The dereverberation filter is estimated from the auto-correlation of the observations and primarily deals with early reflections, while late reflections are effectively suppressed by an inhibitory mechanism that estimates their relative contribution in each time-frequency segment. Information about the salience of the target in a given time-frequency segment based on source separation is combined with the corresponding information based on reverberation suppression through the use of a continually-variable weighting function or mask. Use of the novel reverberation processing results in a relative decrease in WER of 11.5% to 20.9% and use of the combined approaches reduces relative WER by as much as 65.3%.

Index Terms— Speech recognition, missing feature theory, binaural processing, dereverberation, spatial segregation

1. INTRODUCTION

While humans can understand speech even in very adverse acoustical environments, the performance of automatic speech recognition (ASR) systems is severely degraded by environmental noise and other interfering sources. To make matters worse, signal degradation in natural acoustical environments very frequently includes the effects of room reverberation. While reasonable success has been attained in coping with the effects of many types of quasi-stationary additive noise sources [1], such approaches are largely ineffective in dealing with the effects of reverberation. One popular route to robust recognition in recent years has been the use of physiologically- and perceptually-motivated signal processing schemes (*e.g.* [2]). Processing motivated by human binaural perception, which utilizes spatial cues including interaural time difference (ITD) and interaural intensity difference (IID) [3, 4, 5], has long been thought to be useful for separating sound sources from different directions and for coping with the effects of reverberation, and this approach is now being extended to speech recognition (*e.g.*[6]). Some of the recent applications of binaural processing to ASR include the work of [7, 8, 9]. Nevertheless, these separation systems are much less effective in reverberant environments.

It is been known for decades that the auditory system focuses on the first-arriving direct sound wave and suppresses the effects of later-arriving reflected waves. This effect, which is called the "precedence effect" is reviewed in [10], and has been implemented in various forms in several computational models for sound localization (*e.g.* [11, 12]). The precedence effect is likely to help localize sound sources or segregate signals originated from a target source in reverberant environments. Nevertheless, it is not clear that the precedence effect can be useful for ASR, as reverberant components coming even from the target source degrade recognition accuracy. In order to achieve high recognition accuracy in reverberant environments, it is very important to remove reverberant components to the extent possible.

In this paper we describe a method to dereverberate incoming signals before segregating sound sources using binaural processing. The dereverberation filter is estimated from the auto-correlation of the corresponding observation in a straightforward way. Normally, acoustic reverberation is so complex that it is very difficult and computationally demanding to estimate a dereverberation filter exactly. Moreover, adjacent samples of speech are highly correlated, and this inherent correlation must be differentiated from the reverberation to obtain the filter successfully. In our work, the linear prediction (LP) residual of speech is employed to estimate the filter because the residual is relatively free of the inherent correlation as pre-

This work was supported by the National Science Foundation (Grant IIS-0420866), and by the Information and Telecommunication National Scholarship Program sponsored by the Institute of Information Technology Assessment, Korea.

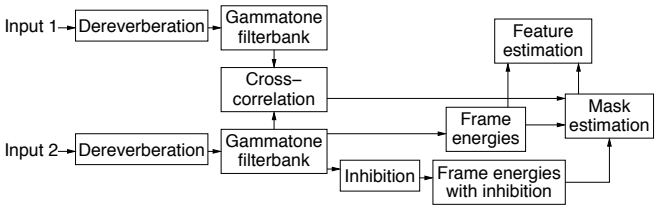

Fig. 1. Overall procedure of the system.

sented in approaches such as [13, 14]. As noted above, the dereverberation filter focuses primarily on early reflections, and it is computed by inverting the truncated auto-correlation corresponding to early reflection. A second inhibitory "echo suppression" mechanism suppresses later reverberant components in a fashion that is motivated by the precedence effect. This mechanism, which is closely based on the algorithms of Martin [12] and Palomäki [15], features an inhibitory signal that is derived from the short-term envelopes of the bandpass-filtered input signals. Finally, we estimate a continuously-variable mask that specifies the estimated relative contribution of the desired target signal to each time-frequency segment.

2. OVERALL PROCEDURE

Fig. 1 describes the overall procedure of the system we described in this conference. In this figure, the two sensors receive reverberant signals from a target source $t(n)$ and possibly several interferences $s_m(n)$ as given by

$$x_1(n) = \sum_{k=0}^{K-1} h_{10}(k)t(n-k) + \sum_{m=1}^{M}\sum_{k=0}^{K-1} h_{1m}(k)s_m(n-k),$$
$$x_2(n) = \sum_{k=0}^{K-1} h_{20}(k)t(n-k) + \sum_{m=1}^{M}\sum_{k=0}^{K-1} h_{2m}(k)s_m(n-k), \quad (1)$$

where $h_{pm}, p = 1, 2$ denotes a transfer fuction from a source to a sensor, and M is the number of interferences.

The signals arriving at the two sensors are first dereverberated to remove early reflection components. In order to avoid distortions in estimating the dereverberation filter by the inherent correlation of speech, the LP residual signal is considered instead of speech itself [13, 14]. The auto-correlation of the LP residual at each sensor is efficiently computed from the inverse Fourier transform of the power spectrum, which estimates the transfer function. Since acoustic reverberation is usually very complicated, a very long filter is normally needed for accurate dereverberation. In the present work we focus only on early reflections of the transfer function, which reduces the length of the estimated reverberation filter. Early reflections are especially problematical for sound source segregation because they affect the cross-correlation within the same frame as the direct sound wave.

The reverberation processing in the present paper was accomplished by truncating the auto-correlation function of the residual of the LPC estimate to ignore the effects of late reflection. Specifically, for nearly exponentially-decaying reverberation like a typical room impulse response, the dereverberation filter can be roughly estimated by

$$h_{derev}(n) \approx \mathrm{IDFT}(1./\mathrm{DFT}([0\cdots 0\, c(0)\cdots c(R)\, 0\cdots 0])), \quad (2)$$

where R is the largest time lag of the auto-correlation $c(\tau)$ after truncation. In our work we used a value of 240 for R, which corresponds to 15 ms at a 16-kHz sampling rate. Although there are many adaptive techniques to train the dereverberation filter (*e.g.* [13, 14]), they may need long adaptation time and lots of computations in advance of dereverberation filtering, which is not acceptable for some ASR systems.

The dereverberated signals are input to a standard model of peripheral auditory processing. Cochlear frequency analysis is performed by a bank of 40 gammatone filters with center frequencies that are linearly spaced in equivalent rectangular bandwidth between 130 Hz and 6.8 kHz. As a crude simulation of auditory nonlinearities, the output of each gammatone filter output is half-wave rectified. Cross-correlation between the resulting signals from the two sensors is computed for all frequency bands by

$$c_i(j,\tau) = \sum_{n=0}^{N-1} x_1(n)x_2(n+\tau)w(n-jT), \quad (3)$$

where i, j, N, T, and $w(n)$ are the frequency index, the frame index, the frame length, the frame period, and the window function. Here, we consider cross-correlation values of the time lag τ that are less than 1 ms in magnitude, and a rectangular window.

Robust ASR systems which include missing-feature processing (*e.g.* [16]) need a mask that indicates which time-frequency segments are reliable and which ones are not. We first construct a continuously-variable mask (or weighting function) that specifies the putative amount of reliability of the target signal in each time-frequency segment. We later set a threshold to discriminate reliable segments from unreliable ones. Each mask value is initialized according to

$$m_s(i,j) = \begin{cases} 0 & \text{if } c_i(j,\tau_t) > c_i(j,\tau_{s_m}) \\ & \text{and } \tau_t - \mu < \arg\max_\tau c_i(j,\tau) < \tau_t + \mu, \\ -L & \text{otherwise,} \end{cases} \quad (4)$$

where τ_t and τ_{s_m} denote the lags corresponding to the target and interference directions. The parameter L should be a sufficiently large number so that the threshold separates the interfering sounds from the desired signal in an appropriate fashion. To segregate target sounds from interfering sounds, the cross-correlation values at τ_t and τ_{s_m} are compared. In addition, if the lag at the maximum cross-correlation does not belong to a neighborhood of τ_t, the segment should be severely corrupted by interference sound or reverberant signal. μ denotes the neighborhood boundary, and we set it to equal 4 or 0.25 ms.

The late reflection components in our system are suppressed by an inhibitory signal that is motivated by the precedence effect. Following the work of Palomäki *et al.*, the inhibitory

signal in each channel, $b_i(n)$, is obtained by low-pass filtering the instantaneous envelope of each gammatone filter output with a time delay [15]. The impulse response of the low-pass filter is

$$h_{lp}(n) = An\exp(\frac{-n}{\kappa})u(n) \quad (5)$$

where the parameters A and κ denote the gain and time constant, respectively, and $u(n)$ is the unit step function. We set A to provide unity gain at DC and κ to provide a 30-ms time constant. Using the inhibitory signal, each gammatone filter output is scaled by

$$r_i(n) = \begin{cases} g_i(n)\frac{(e_i(n)-b_i(n))}{e_i(n)} & \text{if } e_i(n) > b_i(n), \\ 0 & \text{otherwise,} \end{cases} \quad (6)$$

where $g_i(n)$, $e_i(n)$, and $b_i(n)$ are the i^{th} gammatone filter output with half-wave rectification, the corresponding instantaneous envelope, and the inhibitory signal, respectively. We use this scaling approach rather than the subtraction presented in [15] because it depends only on the relative amplitudes of the envelope and the inhibitory signal.

In order to compile information about echo suppression for the mask, energies are computed for both $g_i(n)$ and $r_i(n)$ at each frame. The mask value is then modified by

$$m_c(i,j) = \begin{cases} m_s(i,j) + 10\log_{10}\frac{E_g(i,j)}{E_g(i,j)-E_r(i,j)} \\ \qquad \text{if } 10\log_{10}\frac{E_g(i,j)}{E_g(i,j)-E_r(i,j)} < Q, \\ m_s(i,j) + Q \\ \qquad \text{otherwise,} \end{cases} \quad (7)$$

where $m_c(i,j)$ is the mask that represents the effects of source segregation and echo suppression while $m_s(i,j)$ is a mask that represents the effects of source segregation only. In addition, $E_g(i,j)$ and $E_r(i,j)$ are the frame energies of $g_i(n)$ and $r_i(n)$, respectively. Q also should be sufficiently large but much smaller than L so that Q can not change the label determined by $-L$.

Finally, applying an appropriate threshold to the mask discriminates reliable time-frequency segments from unreliable ones. With this information and the log-spectral frame energies computed from $E_g(i,j)$, we are ready to perform speech recognition with missing features. In this system, the missing features are reconstructed using the cluster-based method [16] because it enables the use of cepstral features which are more effective for ASR than spectral features.

3. SPEECH RECOGNITION EXPERIMENTS

We evaluated the proposed method by speech recognition experiments using the DARPA Resource Management (RM1) database [17] and the CMU SPHINX-III speech recognition system, which is based on fully-continuous hidden Markov models. Using 13^{th}-order mel-frequency cepstral coefficients computed from reconstructed log-spectral energies, 2,880 RM1 sentences recorded in a quiet environment were used to train the recognition system, and 600 sentences were decoded to give a word error rate (WER). The frame size was 25.6 ms, and the frame rate was 10 ms. Assuming there is one interfering speech source, each test utterance was obtained by combining the clean target and interfering speech signals with a reverberation-simulating filter that was obtained using the image method [18]. Although we obtained results with only a single interfering source in the present paper, the algorithm in principle can cope with multiple interferers without modification. Fig. 2 describes the configuration of sources and sensors used in these experiments. The reflection coefficient was chosen to provide a reverberation time RT_{60} of 0.3 s or 0.5 s. We assumed that the locations of the target and interference directions were known *a priori*.

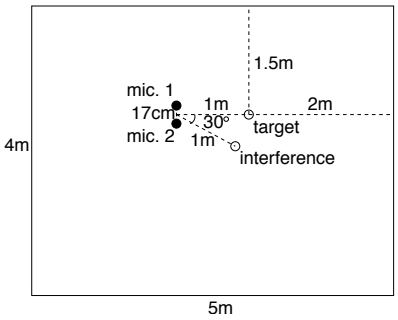

Fig. 2. Configuration of sources and sensors in a rectangular room to simulate room impulse responses from sources to microphones. The simulated height of the room was 3 m, and the height of all sources and microphones was 1.1 m.

Table 1. Comparison of the percentage WER obtained for the CMU SPHINX-III speech recognition system using masks estimated with source segregation plus the inhibitory mechanism or dereverberation at two different signal-to-interference ratios (SIRs) and reverberation times. See text for description.

SIR (dB)	RT_{60}	no proc	ideal masks	seg + inhib	dereverb + seg + inhib
∞	0.0	6.9	—	—	8.5
10	0.3	57.1	16.8	24.6	19.8
	0.5	71.6	18.4	45.6	36.9
0	0.3	109.7	23.6	62.6	49.2
	0.5	107.5	29.0	84.1	74.4

Table 1 presents the WERs obtained using test data as described above. The columns indicate the WER obtained for two SNRs and for two degrees of reverberation. The data columns describe the WER obtained using standard NIST metrics with no processing for signal separation, processing using ideal masks constructed using Oracle knowledge of the signal, and then processing as described in this paper using source segregation with inhibition and in combination with the dereverberation mechanism. For the present task, clean

speech with no interfering signals or reverberation produces a WER of 6.9%. This WER increases to 8.5% if clean speech is processed by the dereverberation and inhibitory mechanisms before recognition.

The ideal masks in Table 1 suggest an upper limit of performance that could be attained from mask-based environmental compensation. These ideal masks were obtained by comparing frame energies of clean target speech and test signal at each time-frequency segment according to

$$m_{ideal}(i,j) = \begin{cases} 10\log_{10}\frac{E_c(i,j)}{E_g(i,j)-E_c(i,j)} & \text{if } E_g(i,j) > E_c(i,j), \\ L & \text{otherwise,} \end{cases} \quad (8)$$

where $E_c(i,j)$ denotes the frame energy of clean target speech in the i^{th} band and j^{th} frame.

It is seen that even without the dereverberation processing, the proposed system provides much greater recognition accuracy than that observed for the baseline system without any processing for environmental robustness for all conditions considered. Furthermore, the rather simple dereverberation processing proposed provided a relative reduction in WER of 11.5% to 20.9%, and the combined processing for reverberation and interference provided relative improvements of up to 65.3% compared to no processing at all.

4. CONCLUSIONS AND FURTHER WORK

In this paper we have described a system that estimates masks for missing feature recognition of speech corrupted by interfering sound and reverberant signal. Our system employed dereverberation to remove early reflection components so that subsequent segregation based on cross-correlation could identify highly noise-contaminated time-frequency segments successfully. In addition, late reflection components were effectively suppressed by comparing spectral energies with and without an inhibitory mechanism. The segregation mechanism that was used to suppress interfering sources was combined with the reverberation-suppression mechanisms through the use of a continuously-variable masking function. The use of dereverberation and echo suppression provided a very substantial improvement in recognition accuracy for speech in reverberant environments in the presence of interfering sounds. While the described algorithm provided much better recognition accuracy than the baseline, results of ideal masks based on Oracle knowledge show that there is much room for improving the mask estimation especially in very adverse environments.

5. REFERENCES

[1] R. Singh, R. M. Stern, and B. Raj, "Signal and feature compensation methods for robust speech recognition," in *Noise Reduction in Speech Applications*, G. Davis, Ed., chapter 9, pp. 221–246. CRC Press, 2002.

[2] H. Hermansky, "Should recognizers have ears?," *Speech Comm.*, vol. 25, no. 1-3, pp. 3–27, 1998.

[3] J. Braasch, "Modelling of binaural hearing," in *Communication Acoustics*, J. Blauert, Ed., chapter 4, pp. 75–108. Springer-Verlag, Berlin, Germany, 2005.

[4] H. S. Colburn and A. Kulkarni, "Models of sound localization," in *Sound Source Localization*, R. Fay and T. Popper, Eds., Springer Handbook of Auditory Research. Springer-Verlag, 2005.

[5] R. M. Stern and C. Trahiotis, "Models of binaural perception," in *Binaural and Spatial Hearing in Real and Virtual Environments*, R. Gilkey and T. R. Anderson, Eds., chapter 24, pp. 499–531. Lawrence Erlbaum Associates, 1996.

[6] R. M. Stern, B. Brown, and DeL. Wang, "Binaural sound localization," in *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, G. Brown and DeL. Wang, Eds., chapter 5. IEEE Press/Wiley, 2006.

[7] N. Roman, D. Wang, and G. J. Brown, "Speech segregation based on sound localization," *J. Acoust. Soc. Am.*, vol. 114, no. 4, pp. 2236–2252, 2003.

[8] Y.-I. Kim, S. J. An, R. M. Kil, and H.-M. Park, "Sound segregation based on binaural zero-crossings," in *INTERSPEECH*, Lisbon, Portugal, Sept. 2005, pp. 2325–2328.

[9] H.-M. Park and R. M. Stern, "Spatial separation of speech signals using continuously-variable masks estimated from comparisons of zero crossings," in *ICASSP*, Toulouse, France, May 2006, pp. 1165–1168.

[10] P. M. Zurek, "The precedence effect," in *Directional Hearing*, W. A. Yost and G. Gourevitch, Eds., pp. 85–105. Springer-Verlag, New York, NY, 1987.

[11] W. Lindemann, "Extension of a binaural cross-correlation model by contralateral inhibition. II. The law of the first wave front," *J. Acoust. Soc. Am.*, vol. 80, no. 6, pp. 1623–1630, 1986.

[12] K. D. Martin, "Echo suppression in a computational model of the precedence effect," in *Proc. IEEE Workshop on Applications of Signal Processing to Acoustics and Audio*, New Paltz, NY, Oct. 1997, pp. 19–22.

[13] 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.

[14] B. W. Gillespie, H. S. Malvar, and D. A. F. Florêncio, "Speech dereverberation via maximum-kurtosis subband adaptive filtering," in *ICASSP*, Salt Lake City, UT, May 2001, pp. 3701–3704.

[15] K. J. Palomäki, G. J. Brown, and D. Wang, "A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation," *Speech Comm.*, vol. 43, pp. 361–378, 2004.

[16] B. Raj, M. L. Seltzer, and R. M. Stern, "Reconstruction of missing features for robust speech recognition," *Speech Comm.*, vol. 43, no. 4, pp. 275–296, 2004.

[17] P. Price, W. M. Fisher, J. Bernstein, and D. S. Pallet, "The DARPA 1000-word resource management database for continuous speech recognition," in *ICASSP*, New York, NY, Apr. 1988, pp. 651–654.

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