INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Robust Speech Recognition Based on Binaural Auditory Processing

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

[1]Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA
[2]Google, Mountain View, CA

anjalim@cs.cmu.edu, chanwcom@gmail.com, rms@cs.cmu.edu

Abstract

This paper discusses a combination of techniques for improving speech recognition accuracy in the presence of reverberation and spatially-separated interfering sound sources. Interaural Time Delay (ITD), observed as a consequence of the difference in arrival times of a sound to the two ears, is an important feature used by the human auditory system to reliably localize and separate sound sources. In addition, the "precedence effect" helps the auditory system differentiate between the direct sound and its subsequent reflections in reverberant environments. This paper uses a cross-correlation-based measure across the two channels of a binaural signal to isolate the target source by rejecting portions of the signal corresponding to larger ITDs. To overcome the effects of reverberation, the steady-state components of speech are suppressed, effectively boosting the onsets, so as to retain the direct sound and suppress the reflections. Experimental results show a significant improvement in recognition accuracy using both these techniques. Cross-correlation-based processing and steady-state suppression are carried out separately, and the order in which these techniques are applied produces differences in the resulting recognition accuracy.

Index Terms: speech recognition, binaural speech, onset enhancement, Interaural Time Difference, reverberation

1. Introduction

The human auditory system is extremely robust. Listeners can correctly understand speech even in very difficult acoustic environments. This includes the presence of multiple speakers, background noise and reverberation. On the other hand, Automatic Speech Recognition (ASR) systems are much more sensitive to the presence of any type of noise or reverberation. In spite of the many advances seen recently using machine learning techniques ($e.g.$ [1, 2]), recognition in the presence of noise and reverberation is still challenging. This is especially pertinent given the rapid rise in voice based machine interaction in recent times.

It is useful to understand the reason behind the robustness of human perception and to apply auditory processing based techniques to improve recognition in noisy and reverberant environments. There have been several successful techniques born out of this approach ($e.g.$ [3, 4, 5, 6, 7] among other sources).

Human auditory perception in the presence of reverberation is widely attributed to processing based on the "precedence effect" as mentioned in [8, 9, 10]. The precedence effect describes the phenomenon where directional cues due to the first-arriving wavefront (corresponding to the direct sound), is given greater perceptual weighting than those cues that arise as a consequence of subsequent reflected sounds. The precedence effect is thought to have an underlying inhibitory mechanism that suppresses echoes at the binaural level [11], but it could also be a consequence of interactions at the peripheral (monaural) level

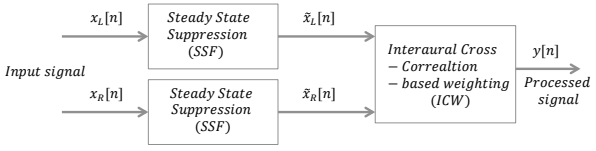


Figure 1: *Overall block diagram of processing using steady-state suppression and interaural cross-correlation based weighting.*

(*e.g.* [12]). Considering the monaural approach, a reasonable way to overcome the effects of reverberation would be to boost these initial wavefronts. This can also be achieved by suppressing the steady-state components of a signal.

The Suppression of Slowly-varying components and the Falling edge of the power envelope (SSF) algorithm [4, 13] was motivated by this principle and has been very successful in improving ASR in reverberant environments. There have been several other techniques developed based on precedence-based processing that have also shown promising results ($e.g.$ [14, 15]).

The human auditory system is also extremely effective in sound source separation, even in very complex acoustical environments. A number of factors affect the spatial aspects of how a sound is perceived. An interaural time difference (ITD) is produced because it takes longer for a sound to arrive at the ear that is farther away from the source. Additionally, an interaural intensity difference (IID) occurs because of a "shadowing" effect of the head causing the sound to be more intense at the ear closer to the source. Spatial separation based on ITD analysis has been very effective in source separation ($e.g.$ [7]).

This study presents a combination of the concepts of precedence-effect-based processing and ITD analysis to improve recognition accuracy in environments containing reverberation and interfering talkers. In this paper we introduce and evaluate the performance of a new method of ITD analysis that utilizes the envelope ITDs.

2. Processing based on binaural analysis

The techniques discussed in this paper roughly follow processing in the human auditory system. For this reason, they include components of monaural processing pertaining to the peripheral auditory system as well as binaural processing that is performed higher up in the brainstem. The overall block diagram of the processing described in Section 2.1 and 2.2 is shown in Figure 1. Steady-state suppression, described in Section 2.1, is performed monaurally, and subsequently a weight that is based on interaural cross-correlation is applied to the signal, as described in Section 2.2. Both of these techniques can be applied

http://dx.doi.org/10.21437/Interspeech.2017-1665

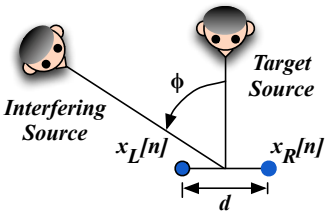


Figure 2: *Two-microphone setup with an on-axis target source and off-axis interfering source used in this study.*

independently of each other.

The processing described in this paper pertains to a two-microphone setup as shown in Figure 2. The two microphones are placed in a reverberant room with a target talker directly in front of them. The target signal thus arrives at both microphones at the same time leading to an ITD of zero. An interfering talker is also present located at an angle of ϕ with respect to the two microphones.

2.1. Steady-State Suppression

The SSF algorithm [4, 13] was used in this study to achieve steady-state suppression. The SSF algorithm is motivated by the precedence effect and by the modulation-frequency characteristics of the human auditory system. A block diagram describing SSF processing is shown in Figure 3. SSF processing was performed separately on each channel of the binaural signal.

After performing pre-emphasis on the input signal, a Short Time Fourier Transform (STFT) of the signal is computed using a 40-channel gammatone filterbank. The center frequencies of the gammatone filterbank are linearly spaced in Equivalent Rectangular Bandwidth (ERB) [16] between 200 Hz and 8 kHz. The STFT was computed with frames of length 50-ms with a 10-ms temporal spacing between frames. These longer-duration window sizes have been shown to be useful for noise compensation [17, 4]. The power $P[m,l]$ corresponding to the m^{th} frame and the l^{th} gammatone channel is given by,

$$P[m,l] = \sum_{k=0}^{N-1} |X[m,k]H_l[k]|^2, 0 \leq l \leq L-1, \quad (1)$$

where $H_l[k]$ is the frequency response of the l^{th} gammatone channel evaluated at the k^{th} frequency index and $X[m,k]$ is the signal spectrum at the m^{th} frame and the k^{th} frequency index. N is FFT size which was 1024.

The power $P[m,l]$ is then lowpass filtered to obtain $M[m,l]$.

$$M[m,l] = \lambda M[m-1,l] + (1-\lambda)P[m,l], \quad (2)$$

Here λ is a forgetting factor that was adjusted for the bandwidth of the filter and experimentally set to 0.4. Since SSF is designed to suppress the slowly-varying portions of the power envelopes, the SSF processed power $\tilde{P}[m,l]$ is given by,

$$\tilde{P}[m,l] = max(P[m,l] - M[m,l], c_0 M[m,l]), \quad (3)$$

where c_0 is a constant introduced to reduce spectral distortion. Since $\tilde{P}[m,l]$ is given by subtracting the slowly varying power envelope from the original power signal, it is essentially a highpass-filtered version of $P[m,l]$, thus achieving steady-state suppression. The value for c_0 was experimentally set to 0.01.

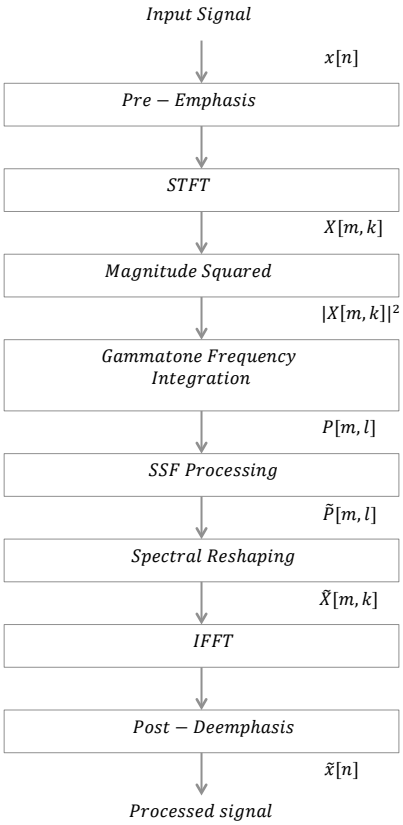


Figure 3: *Block diagram describing the SSF algorithm.*

For every frame in every gammatone filter band, a channel-weighting coefficient $w[m,l]$ is obtained by taking the ratio of the highpass filtered portion of $P[m,l]$ to the original quantity given by

$$w[m,l] = \frac{\tilde{P}[m,l]}{P[m,l]}, 0 \leq l \leq L-1 \quad (4)$$

Each channel-weighting coefficient corresponding to the l^{th} gammatone channel is associated with the response $H_l[k]$ and so the spectral weighting coefficient $\mu[m,k]$ is given by

$$\mu[m,k] = \frac{\sum_{l=0}^{L-1} w[m,l]|H_l[k]|}{\sum_{l=0}^{L-1} |H_l[k]|}, 0 \leq l \leq L-1, 0 \leq k \leq N/2 \quad (5)$$

The final processed spectrum is then given as

$$\tilde{X}[m,k] = \mu[m,k]X[m,k], 0 \leq k \leq N/2 \quad (6)$$

Using Hermitian symmetry, the rest of the frequency components are obtained and the processed speech signal $\tilde{x}[n]$ is re-synthesized using the overlap-add method.

2.2. Interaural Cross-correlation-based Weighting

Using SSF processing described in the previous section, steady-state suppression is achieved, effectively leading to enhancement of the acoustic onsets. Interaural Cross-correlation-based Weighting (ICW) is then used to separate the target signal on the basis of ITD analysis.

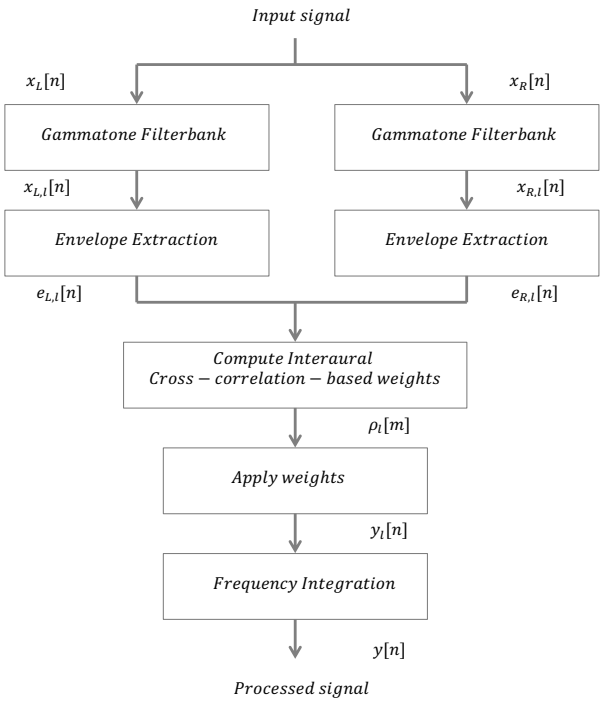


Figure 4: *Block diagram describing the ICW algorithm.*

A crude model of the auditory-nerve response to sounds starts with bandpass filtering of the input signal (modeling the response of the cochlea), followed by half-wave rectification and then by a lowpass filter. The auditory nerve response roughly follows the fine structure of the signal at low frequencies and the envelope of the signal at high frequencies [3, 18, 19]. ITD analysis is based on the cross-correlation of auditory-nerve responses, and the human auditory system is especially sensitive to envelope ITD cues at high frequencies. The ICW algorithm uses this concept to reject components of the input signal that appear to produce greater ITDs of the envelope.

Figure 4 shows a block diagram of the ICW algorithm. As mentioned above, it is assumed that there is no delay in the arrival of the target signal between the right and left channel denoted by $x_R[n]$ and $x_L[n]$ respectively. The signals $x_R[n]$ and $x_L[n]$ are first bandpass filtered by a bank of 40 gammatone filters using a modified implementation of Malcolm Slaney's Auditory Toolbox [20]. The center frequencies of the filters are linearly spaced according to their equivalent rectangular bandwidth (ERB) [16] between 100 Hz and 8 kHz. Zero-phase filtering is performed using forward-backward filtering such that the effective impulse response is given by,

$$h_l(n) = h_{g,l}(n) * h_{g,l}(-n) \tag{7}$$

where $h_{g,l}(n)$ is the impulse response of the original gammatone filter for the l^{th} channel. Since equation (7) leads to an effective reduction in bandwidth, the bandwidths of the original gammatone filters are modified to roughly compensate for this.

After bandpass filtering, instantaneous Hilbert envelopes $e_{L,l}[n]$ and $e_{R,l}[n]$ of the signals are extracted. Here, l refers to the gammatone filter channel. The normalized cross-correlation of the envelope signals $e_{L,l}[n]$ and $e_{L,l}[n]$ is given by,

$$\rho_l[m] = \frac{\sum_{N_w} e_{L,l}[n;m] e_{R,l}[n;m]}{\sqrt{\sum_{N_w} e_{L,l}[n;m]^2} \sqrt{\sum_{N_w} e_{R,l}[n;m]^2}} \tag{8}$$

where $\rho_l[m]$ refers to the normalized cross-correlation of the m^{th} frame and l^{th} gammatone channel, $e_{L,l}[n;m]$ and $e_{R,l}[n;m]$ are the envelope signals corresponding to the m^{th} frame and l^{th} gammatone channel for the left and right channels respectively. The window size N_w was set to 75 ms and the time between frames for ICW was 10 ms.

Based on $\rho_l[m]$, the weight computation was given by,

$$w_l[m] = \rho_l[m]^a \tag{9}$$

The nonlinearity a is introduced to cause a sharp decay of w_l as a function of ρ_l and it was experimentally set to 3. The weights computed are applied as given below:

$$y_l[n;m] = w_l[m]\bar{x}[n;m] \tag{10}$$

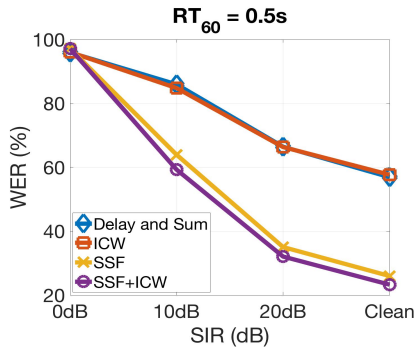
where $y_l[n;m]$ is the short-time signal corresponding to the m^{th} frame and l^{th} gammatone channel and $\bar{x}[n;m]$ is the average of short-time signals $x_{R,l}[n;m]$ and $x_{L,l}[n;m]$ corresponding to the m^{th} frame and l^{th} gammatone channel. To resynthesize speech, all l channels are then combined.

3. Experimental Results

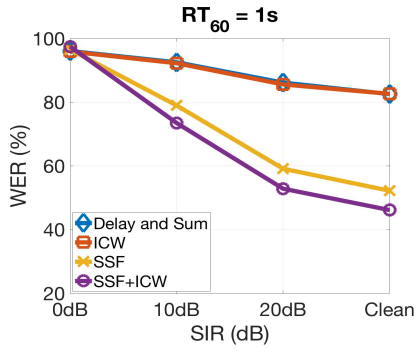
In order to test the SSF+ICW algorithm, ASR experiments were conducted using the DARPA Resource Management (RM1) database [21] and the CMU SPHINX-III speech recognition system. The training set consisted of 1600 utterances and the test set consisted of 600 utterances. Features used were 13^{th} order mel-frequency cepstral coefficients. Acoustic models were trained using clean speech. SSF processing was performed on the training data in cases where SSF was part of the algorithm being tested.

To simulate speech corrupted by reverberation and interfering talkers, a room of dimensions $5m \times 4m \times 3m$ was assumed. The distance between the two microphones is 4 cm. The target speaker is located 2 m away from the microphones along the perpendicular bisector of the line connecting the two microphones. An interfering speaker is located at an angle of 45 degrees to one side and 2 m away from the microphones. This whole setup is 1.1 m above the floor. To prevent any artifacts that may arise from only testing the algorithm at a specific location in the room, the whole configuration described above was moved around in the room to 25 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 using the RIR package [22, 23].

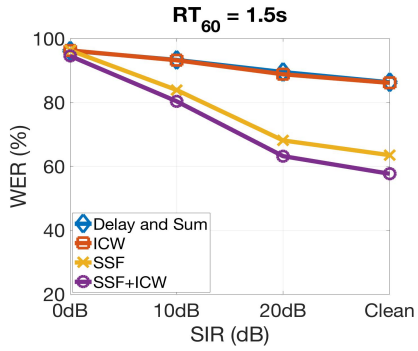
Figure 5 shows the results obtained using baseline Delay and Sum processing, the SSF algorithm alone, the ICW algorithm alone and the combination of the SSF and ICW algorithms. Figures 5a-5d show the Word Error Rate (WER) as a function of Signal-to-Interference Ratio (SIR) for four different values of reverberation time. The performance of the SSF+ICW algorithm is compared to that of SSF alone and ICW alone. The results of the Delay and Sum algorithm serve as baseline. As seen in Figures 5a-5d, the ICW algorithm applied by itself does not provide any improvement in performance compared to baseline Delay-and-Sum processing. Nevertheless, the addition of



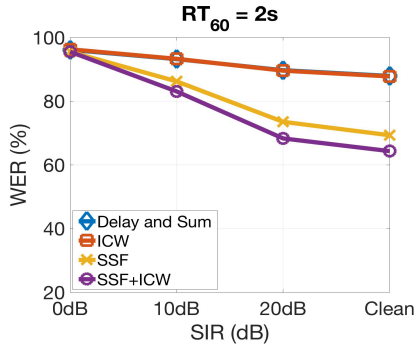
(a)



(b)



(c)



(d)

Figure 5: *Word Error Rate as a function of Signal to Interference Ratio for an interfering signal located 45 degrees off axis at various reverberation times: (a) 0.5 s (b) 1 s (c) 1.5 s (d) 2 s.*

ICW to SSF does lead to a reduction in WER compared to performance obtained using SSF alone as seen in Figures 5a-5d . While the WER remains the same for 0 dB SIR, for all the other conditions the addition of ICW to SSF decreases the WER by upto 12% relative. There is a consistent improvement in WER for 10 dB and 20 dB SIR and in the absence of an interfering talker. The inclusion of envelope ITD cues and their coherence across binaural signals therefore, help with reducing both interfering noise and reverberation.

4. Conclusion

In this paper, a new method of utilizing ITD cues extracted from the signal envelopes is discussed. By looking at the cross-correlation between the high frequency signal envelopes of the two channels of a binaural signal, an ITD based weight is computed that rejects portions of the signal corresponding to longer ITDs. Combining this information with precedence-based processing that emphasizes acoustic onsets leads to improved recognition in the presence of reverberation and interfering talkers.

5. References

[1] M. L. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7398–7402.

[2] 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.

[3] J. Blauert, *Spatial hearing: the psychophysics of human sound localization*. MIT press, 1997.

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

[5] C. Kim, K. Kumar, and R. M. Stern, "Binaural sound source separation motivated by auditory processing," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 5072–5075.

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

[7] 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 Communication*, vol. 43, no. 4, pp. 361–378, 2004.

[8] 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.

[9] 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.

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

[11] W. Lindemann, "Extension of a binaural cross-correlation model by contralateral inhibition. I. simulation of lateralization for stationary signals," *Journal of the Acoustical Society of America*, vol. 80, pp. 1608–1622, 1986.

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

[14] C. Kim, K. K. Chin, M. Bacchiani, and R. M. Stern, "Robust speech recognition using temporal masking and thresholding algorithm." in *INTERSPEECH*, 2014, pp. 2734–2738.

[15] B. J. Cho, H. Kwon, J.-W. Cho, C. Kim, R. M. Stern, and H.-M. Park, "A subband-based stationary-component suppression method using harmonics and power ratio for reverberant speech recognition," *IEEE Signal Processing Letters*, vol. 23, no. 6, pp. 780–784, 2016.

[16] B. C. Moore and B. R. Glasberg, "A revision of zwicker's loudness model," *Acta Acustica united with Acustica*, vol. 82, no. 2, pp. 335–345, 1996.

[17] C. Kim and R. M. Stern, "Power function-based power distribution normalization algorithm for robust speech recognition," in *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on*. IEEE, 2009, pp. 188–193.

[18] R. M. Stern, G. J. Brown, D. Wang, D. Wang, and G. Brown, "Binaural sound localization," *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, pp. 147–185, 2006.

[19] R. M. Stern and C. Trahiotis, "Models of binaural interaction," *Handbook of perception and cognition*, vol. 6, pp. 347–386, 1995.

[20] M. Slaney, "Auditory toolbox version 2," *University of Purdue, https://engineering. purdue. edu/˜ malcolm/interval/1998-010*, 1998.

[21] P. Price, W. M. Fisher, J. Bernstein, and D. S. Pallett, "The darpa 1000-word resource management database for continuous speech recognition," in *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*. IEEE, 1988, pp. 651–654.

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

[23] 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.