Robust Speech Recognition in Reverberant Environments Using Subband-Based Steady-State Monaural and Binaural Suppression

Hyung-Min Park[1], *Matthew Maciejewski*[2], *Chanwoo Kim*[3], *Richard M. Stern*[2]

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

hpark@sogang.ac.kr, mmacieje@andrew.cmu.edu, chanwcom@gmail.com, rms@cs.cmu.edu

10.21437/Interspeech.2014-153

Abstract

The precedence effect describes the ability of the auditory system to suppress the later-arriving components of sound in a reverberant environment, maintaining the perceived arrival azimuth of a sound in the direction of the actual source, even though later reverberant components may arrive from other directions. It is also widely believed that precedence-like processing can also improve speech intelligibility, as well as the accuracy of speech recognition systems, in reverberant environments. While the mechanisms underlying the precedence effect have traditionally been assumed to be binaural in nature, it is also possible that the suppression of later components may take place monaurally, and that the suppression of the later-arriving components of the spatial image may be a consequence of this more peripheral processing. This paper compares the potential contributions of onset enhancement (and consequent steady-state suppression) of the envelopes of subband components of speech at both the monaural and binaural levels. Experimental results indicate that substantial improvement in recognition accuracy can be obtained in reverberant environments if the feature extraction includes both onset enhancement and binaural interaction. Recognition accuracy appears to be relatively unaffected by the stage in the suppression processing at which the binaural interaction takes place.

Index Terms: robust speech recognition, precedence effect, reverberation, onset enhancement, steady-state suppression, binaural interaction

1. Introduction

An important attribute of human binaural hearing is that binaural localization is dominated by the first-arriving components of a complex sound [1]. This phenomenon, which is referred to as the precedence effect and has been discussed extensively (*e.g.* [2, 3]), is clearly helpful in enabling the perceived location of a source in a reverberant environment to remain constant, as it is dominated by the characteristics of the components of the sound which arrive directly from the sound source while suppressing the potential impact of later-arriving reflected components from other directions. In addition to its role in maintaining perceived constancy of direction of arrival in reverberation, the precedence effect is also believed by some to improve speech intelligibility in reverberant environments, although it is difficult to separate the potential impact of the precedence effect from that of conventional binaural unmasking.

There have been multiple theories concerning the mechanisms that underly the precedence effect. Most of these theories involve the modification of the outputs of a standard model of binaural processing in a fashion that enhances onset components of binaural responses and, correspondingly, suppresses the steady-state components (*e.g.* [2]). (The standard model of binaural processing typically proposes, in effect, a combination of bandpass filtering and nonlinear rectification in the auditory periphery followed by the development of an interaural cross-correlation function on a frequency-by-frequency basis (*e.g.* [4, 5]).) One popular realization of such a scheme is the computational model of Lindemann [6, 7], which proposes a neural mechanism based on a type of lateral suppression along the lag axis of the putative binaural processor. Other investigators have suggested that at least some of the phenomena associated with the precedence effect can be accounted for strictly on the basis of conventional models of auditory processing that consider bandpass filtering and subsequent interaural correlation, without the need for an explicit "precedence mechanism" (*e.g.* [8]).

While theories like the Lindemann model have been able to account for some of the observed psychoacoustical results, it is also possible that the precedence effect may be a simple consequence of binaural processing of acoustical signals that had been subjected to onset enhancement or steady-state suppression *monaurally*. For example, Martin [9] had proposed a monaurally-based suppression mechanism that is applied to the interaural correlation display. Kim later proposed a different model called *suppression of slowly-varying components and the falling edge* (SSF) that also accomplishes a similar type of onset enhancement and steady-state suppression on a band-by-band basis [10]. Kim demonstrated that the SSF approach can provide substantial improvements of speech recognition accuracy in reverberant environments.

The purpose of this paper is to determine the extent to which the type of steady-state suppression that arises as a consequence of SSF processing can be enhanced by developing a "binaural" implementation of the suppression, in a fashion that is motivated by traditional models of binaural processing. In the next section we briefly review the SSF processing as developed by Kim. In Sec. 3 we describe multiple possible extensions of SSF that involve the combination of inputs from two microphones. The latter sections describe the results of our experiments that measure the extent to which these combinations are effective in reducing word error rate for speech recognition systems in reverberant environments.

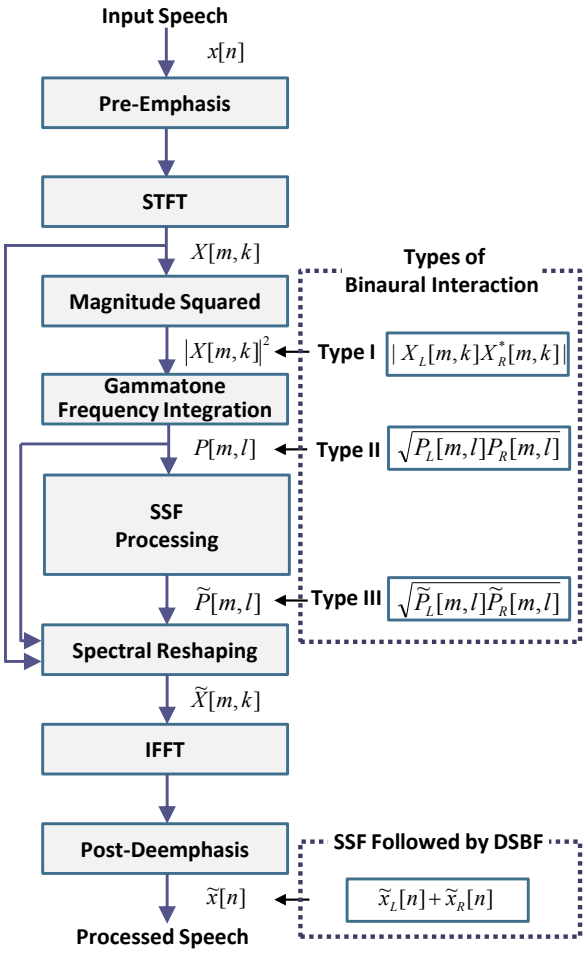

Figure 1: *Left column: Summary of SSF processing for steady-state binaural suppression as described in [10]. Right column: Sites of binaural interaction considered in this paper.*

2. SSF Processing for Steady-State Suppression

The left column of Fig. 1 shows the overall procedure of SSF Type-II processing as described in [10], which we subsequently refer to as SSF processing for simplicity[1]. The input signal is pre-emphasized, and a short-time Fourier transform (STFT) is performed using a 50-ms Hamming window with 10 ms between frames. (We use a longer-duration window because observations in previous work have consistently shown that such "medium-time" processing is more effective for noise estimation or compensation [11, 12, 13].) Magnitude-squared STFT outputs are integrated over frequency to obtain the power $P[m,l]$ at the m^{th} frame and l^{th} Gammatone channel as

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

where $H_l[k]$ is the transfer function of the l^{th} channel evaluated at linear frequency index k, of a gammatone filter bank whose center frequencies are linearly spaced in Equivalent Rectangular Bandwidth (ERB) [14] between 200 Hz and 8 kHz [15]. $X[m,k]$ is the signal spectrum at the m^{th} frame and k^{th} frequency. Here, $N = 1024$ and $L = 40$ represent the FFT size and the number of gammatone channels, respectively.

In each channel, the power $P[m,l]$ is lowpass filtered to obtain $M[m,l]$ using the following equation:

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

where λ is a forgetting factor that is adjusted for the bandwidth of the filter. In SSF processing, the processed power is obtained by

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

where c_0 is a small fixed coefficient to reduce spectral distortion between clean and reverberant signals. Since $M[m,l]$ is subtracted from $P[m,l]$, $\tilde{P}[m,l]$ is essentially a highpass filtered signal with suppression of slowly-varying components and the falling edge of the power contour. (The values of λ and c_0 are experimentally set to 0.4 and 0.01, respectively.)

Using the spectral reshaping approach described in [12] and [16], with comparing the processed power $\tilde{P}[m,l]$ to the original gammatone subband power $P[m,l]$, we obtain a processed spectrum $\tilde{X}[m,k]$. Assuming that the phases of the original and the processed spectra are identical, we modify only the magnitude spectrum.

For each time-frequency bin, the channel weighting coefficient $w[m,l]$ is obtained by the ratio of $\tilde{P}[m,l]$ and $P[m,l]$ as follows:

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

Since each channel is associated with $H_l[k]$, the spectral weighting coefficient $\mu[m,k]$ is obtained by

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

For the lower half of the frequency region, the processed spectrum is obtained by

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

After invoking Hermitian symmetry of the processed spectrum to obtain the remaining frequency components, the enhanced speech $\hat{x}[n]$ is re-synthesized using the IFFT and the overlap-add (OLA) method as in [16].

Conventional feature extraction using a window with duration between 20 and 30 ms typical for speech analysis and subsequent speech recognition is performed on the re-synthesized speech.

3. Combining SSF processing with binaural interaction

As discussed in the Introduction, one popular way of thinking about the precedence effect is to posit that a suppression mechanism is applied to the binaural display, as in the Lindemann model [6, 7]. But at least in principle, if suppression were to take place more peripherally using a mechanism such as was proposed by Martin [9], any binaural interaction based on signals that had undergone steady-state suppression should in turn

[1] We consider only SSF Type-II processing because it consistently provides better performance than SSF Type-I processing as shown in [10].

reflect that suppression in a fashion that in practice would resemble precedence-effect processing.

We explored the impact on recognition accuracy of binaural combination of signals undergoing SSF steady-state suppression at three levels of the suppression process: immediately after the initial raw spectral analysis (Type I), after the gammatone frequency integration (Type II), and immediately after the SSF processing (Type III), as well as simple binaural combination of the final reconstructed waveform, which is effectively delay-and-sum processing of the resynthesized waveforms $\hat{x}[n]$ from the two microphones. While in our work we have assumed that the target source is along the perpendicular bisector of the line between the two mics, signals off the midline could easily be processed binaurally by compensating explicitly for the relative time delay caused by the differences between the distances of the direct paths from the source to the microphones. This type of internal delay is entirely consistent with modern theories of human binaural interaction (*e.g.* [4, 5]).

Referring to the right column of Fig. 1, the three types of binaural combination considered are as follows:

$$\text{Type I:}\quad P_{B_1}[m,l] = \sum_{k=0}^{N-1} |X_L[m,k]X_R^*[m,k]||H_l[k]|^2, \quad 0 \le l \le L-1, \tag{7}$$

$$\text{Type II:}\quad P_{B_2}[m,l] = \sqrt{P_L[m,l]P_R[m,l]}, \quad 0 \le l \le L-1, \tag{8}$$

$$\text{Type III:}\quad \tilde{P}_{B_3}[m,l] = \sqrt{\tilde{P}_L[m,l]\tilde{P}_R[m,l]}, \quad 0 \le l \le L-1, \tag{9}$$

where the subscripts L and R denote the results obtained at the corresponding stage of conventional (monaural) SSF Type-II processing for the 'left-' and 'right-ear' signals, respectively. Subscript B refers to binaural interaction, and the associated number represents the type of binaural combination used. The subsequent processing is the same as in the monaural case using the results obtained from the 'left-ear' signal for the frequency-band power $P[m,l]$ in (4) and a frequency spectrum $X[m,k]$ in (6).

It seems that this method can enhance only the speech signal from the plane bisecting the line segment connecting two microphones. However, it can be used for speakers from any directions by applying proper delays to one of the microphone signals to remove the relative time delay between the signals after the time delay is estimated by speaker localization algorithms (e.g. [17]).

4. Experimental results

We conducted recognition experiments using the DARPA Resource Management (RM) database [18] and the CMU SPHINX-III speech recognition system. The recognition system was based on fully-continuous hidden Markov models, which are trained on 1,600 utterances recorded in a quiet environment. The test set consists of 600 utterances. We used 13th-order mel-frequency cepstral coefficients with a frame size of 25.6 ms and a frame rate of 10 ms developed in the conventional fashion.

To simulate the effects of room reverberation, impulse responses were generated by the image method (using the software package Room Impulse Response [19]), which simulates acoustics between two points in a rectangular room [20]. Figure 2 depicts the geometry of the virtual room that was used to

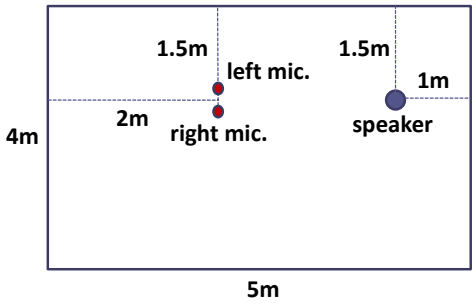

Figure 2: *Virtual room to simulate impulse responses. The room is 3-m high, and the source and microphones are 1.1-m off the floor.*

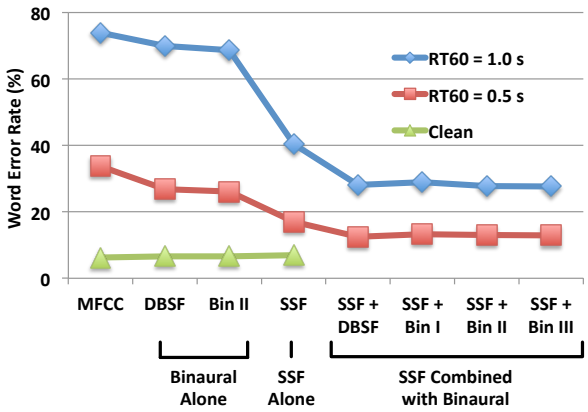

Figure 3: *Comparison of the impact of SSF suppression at the monaural and binaural levels for two reverberation times simulated using the image method.*

simulate the impulse responses, with the virtual microphones placed 110 cm above the floor, and a ceiling height of 3 meters. Various microphone spacings were considered, as discussed below and for 'monaural' processing, a microphone was placed at the center between the two microphones pictured. the systems were trained on clean speech, but with the compensation configuration (*i.e.* SSF and/or binaural processing) that matched the testing condition.

Our major results are summarized in Fig. 3, which compares results obtained using no SSF processing or binaural interaction (*MFCC*), two types of binaural processing without SSF processing, at the stage just after the gammatone frequency integration (*Bin II*) and conventional delay-and-sum waveform combination (*DSBF*), SSF processing without binaural interaction (*SSF*), and SSF processing with four different types of binaural interaction as indicated in the figure. The results in Fig. 3 were obtained with a spacing of 17 cm between the two microphones and two reverberation times, 500 ms and 1 s.

The results shown in Fig. 3 demonstrate clearly that while binaural processing and SSF processing are beneficial in the reverberant environments considered (although the effectiveness of binaural processing with only two microphones is limited), there is a clear advantage of combining binaural and SSF processing, with improvements in relative WER of 30% and 59% in comparison to SSF alone and binaural processing alone, respectively. Interestingly, the exact locus of the binaural inter-

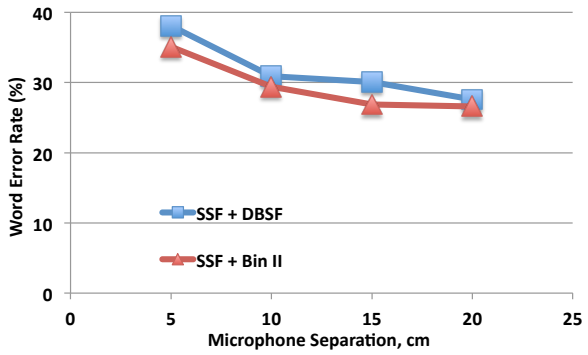

Figure 4: *Comparison of the effect of microphone separation on results.*

action appears to have very little impact on the results. SSF processing produces a relative WER increase of about 13% for clean speech which is a relatively small degradation, as has been noted previously [10].

Figure 4 summarizes results for two combinations of binaural processing with SSF suppression as a function of distance between the two microphones. While microphones are frequently closely spaced on the order of 2 to 5 cm to avoid the potential effects of spatial aliasing, we observed that substantial improvements in performance are observed when the spacing is increased to about 17 cm. Although not shown in the figure, performance is stable for wider microphone separations up to at least 1 m. In comparison, the actual maximum path length difference incurred in human binaural processing is approximately 22 cm, based on anatomical considerations.

5. Conclusions

We demonstrate in this paper that speech recognition accuracy in reverberant environments can be substantially improved through the use of a combination of suppression of the steady-state components of speech signals combined with binaural-type combination of the suppressed signals. While either of these two approaches can provide benefit when applied in isolation, there is substantial advantage to be had by the use of these two methods in combination. The recognition accuracy observed is surprisingly unaffected by the stage within the processing at which the binaural interaction is imposed.

6. Acknowledgements

This work was supported by LG Yonam Foundation and the Cisco Corpopration (Grant 570877).

7. References

[1] H. W. Wallach, E. B. Newman, and M. R. Rosenzweig, "The precedence effect in sound localization," *American Journal of Psychology*, vol. 62, pp. 315–337, 1949.

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

[3] R. Y. Litovsky, S. H. Colburn, W. A. Yost, and S. J. Guzman, "The precedence effect," *Journal of the Acoustical Society of America*, vol. 106, pp. 1633–1654, 1999.

[4] R. M. Stern and C. Trahiotis, "Models of binaural interaction," in *Hearing*, 2nd ed., ser. Handbook of Perception and Cognition, B. C. J. Moore, Ed. Academic (New York), 1995, ch. 10, pp. 347–386.

[5] R. M. Stern, D. Wang, and G. J. Brown, "Binaural sound localization," in *Computational Auditory Scene Analysis*, D. Wang and G. J. Brown, Eds. Wiley-IEEE Press, 2006, ch. 5.

[6] 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.

[7] ——, "Extension of a binaural cross-correlation model by contralateral inhibition. II. the law of the first wavefront," *Journal of the Acoustical Society of America*, vol. 80, pp. 1623–1630, 1986.

[8] K. Hartung and C. Trahiotis, "Peripheral auditory processing and investigations of the "precedence effect" which utilize successive transient stimuli," *Journal of the Acoustical Society of America*, vol. 110, no. 3, pp. 1505–1513, September 2001.

[9] K. D. Martin, "Echo suppression in a computational model of the precedence effect," in *Proc. IEEE Mohonk Workshop on Applications of Signal Processing to Acoustics and Audio*, 1997.

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

[11] ——, "Feature extraction for robust speech recognition using a power-law nonlinearity and power-bias subtraction," in *Proc. INTERSPEECH-2009*, Sep. 2009, pp. 28–31.

[12] 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 *Proc. INTERSPEECH-2009*, Sep. 2009, pp. 2495–2498.

[13] C. Kim and R. M. Stern, "Feature extraction for robust speech recognition based on maximizing the sharpness of the power distribution and on power flooring," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Mar. 2010, pp. 4574–4577.

[14] B. C. J. Moore and B. R. Glasberg, "A revision of Zwicker's loudness model," *Acustica - Acta Acustica*, vol. 82, no. 2, pp. 335–345, 1996.

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

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

[17] J.-W. Cho and H.-M. Park, "Imposition of sparse priors in adaptive time delay estimation for speaker localization in reverberant environments," *IEEE Signal Processing Letters*, vol. 16, no. 3, pp. 180–183, 2009.

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

[19] S. G. McGovern. A model for room acoustics. [Online]. Available: http://2pi.us/rir.html

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