BINAURAL PROCESSING FOR ROBUST RECOGNITION OF DEGRADED SPEECH

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

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

ABSTRACT

This paper discusses a new combination of techniques that help in improving the accuracy of speech recognition in adverse conditions using two microphones. Classic approaches toward binaural speech processing use some form of cross-correlation over time across the two sensors to effectively isolate target speech from interferers. Several additional techniques using temporal and spatial masking have been proposed in the past to improve recognition accuracy in the presence of reverberation and interfering talkers. In this paper, we consider the use of cross-correlation across frequency over some limited range of frequency channels in addition to the existing methods of monaural and binaural processing. This has the effect of locating and reinforcing coincident peaks across frequency over the representation of binaural interaction and provides local smoothing over the specified range of frequencies. Combined with the temporal and spatial masking techniques mentioned above, this leads to significant improvements in binaural speech recognition.

Index Terms—

Binaural speech, auditory processing, robust speech recognition, speech enhancement, cross-correlation

1. INTRODUCTION

Speech recognition systems have undergone significant improvements in recent times especially with the advent and widespread use of machine learning techniques [1, 2]. Nevertheless, noise robustness remains problematical. Robustness is especially important with the increasing use of voice-based user interface for cell phones, smart home devices, cars etc. Improving speech recognition accuracy in the presence of non-stationary noise sources and other adverse conditions such as reverberation is still a challenge.

Human beings, on the other hand, are extremely good at localizing and separating simultaneously-presented speech sources in a variety of adverse conditions, the well known "cocktail party problem". Human hearing, even in adverse conditions, remains fairly robust. It is useful to attempt to understand the reason behind the robustness of human perception and to apply techniques based on our understanding of auditory processing 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 others).

Among the models of binaural hearing, one of the earliest was the model of Sayers and Cherry [8], which related the lateralization of binaural signals to their interaural cross-correlation. In terms of binaural speech processing, a popular approach towards separating target sounds in adverse environments is the grouping of sources according to common source location. This usually entails the use interaural time difference (ITD) and interaural intensity difference (IID). ITD is caused by differences in path length between a source and the two ears, producing corresponding differences in the arrival times of that sound to the two ears. (Normally, binaural recordings must be made using an artificial head in order for significant IID cues to be present.) Models that describe how these cues are used to lateralize sound sources are reviewed in [9, 10], among other sources. *Straightness weighting* refers to a hypothesis that greater emphasis is given to the contributions of ITDs that are consistent over a range of frequencies [11, 12, 13]. This was motivated by the fact that real sounds emitted by point sources produced ITDs that were consistent over a range of frequencies. Hence, the existence of a "straight" maximum of the interaural cross-correlation function over a range of frequencies could be used to identify the correct ITD.

Missing-feature techniques attempt to identify the subset of spectro-temporal elements in a spectrogram-like display that are unaffected by sources of distortion such as additive noise, competing talkers, or the effects of reverberation, and reconstruct a signal based only on the undistorted components [14]. These algorithms can provide rather good performance provided that the undistorted components are correctly identified. Several researchers have demonstrated that information based on ITD (or in some cases IID or interaural correlation) can be very useful in estimating binary (or continuous) masks that indicate which components of a signal are close to those of the desired source (*e.g.* [15, 7, 16]. In [17], the Phase Difference Channel Weighting (PDCW) algorithm is used to perform binary mask estimation using interaural phase difference in the frequency domain, leading to considerable improvements in recognition accuracy.

The *precedence effect* describes the phenomenon where

directional cues attributed to the first-arriving wavefront (corresponding to the direct sound) are given greater perceptual weighting than those cues that arise as a consequence of subsequent reflected sounds [18, 19, 20]. While the precedence effect is clearly helpful in maintaining constant localization in reverberant environments, many researchers believe that it also contributes to improved speech intelligibility in the presence of reverberation. The precedence effect is typically modeled as a mechanism that suppresses echoes at either the monaural level [21] or binaural level [22]. A reasonable way to overcome the effects of reverberation would be to boost these onsets or 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, 23] was motivated by this principle and has been successful in improving speech recognition accuracy in reverberant environments. There have been several other techniques developed based on precedence based processing that have also shown promising results ($e.g.$ [24, 25]).

In this paper we introduce a new processing procedure, *Cross-Correlation across Frequency* (CCF), which (as the name implies) correlates signals across the analysis channels. We show that although computational intensive, CCF can improve recognition accuracy very substantially in environments that contain both additive interference and reverberation. In Sec. 2 we review some basic binaural phenomena along with some algorithms motivated by aspects of binaural hearing that have been used to improve speech recognition accuracy, and we introduce the CCF algorithm in Sec. 3. We describe our experimental results in Sec. 4 and provide discussion and conclusions in Secs. 5 and 6.

2. BINAURAL PROCESSING

This paper addresses binaural processing in adverse conditions, which include the presence of reverberation and interfering talkers. The techniques 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. An interfering talker is also present located at an angle of ϕ with respect to the two microphones.

The techniques discussed in this paper are largely motivated by knowledge of human monaural and binaural auditory processing. A basic block diagram of the algorithm discussed in this paper is shown in Figure 2. Explanations of each of the blocks are provided below.

2.1. Steady-state suppression

In the presence of reverberation, steady-state suppression has been shown to vastly improve accuracy in automatic speech recognition (ASR). The use of steady-state suppression was originally motivated by the precedence effect and the modulation frequency characteristics of the human auditory system. It aims at boosting the parts of the input signal that are believed to correspond to the direct sound, which indirectly suppresses reflected sounds. In this paper, the SSF algorithm noted above [4, 23] was used to achieve steady-state suppression.

The SSF algorithm in its initial formulation decomposes the input signal into 40 gammatone frequency channels. For each of these channels, the frame-level power is computed and then lowpass filtered. 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

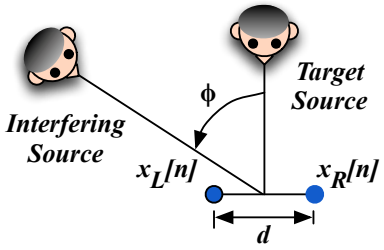

Fig. 1. Two-microphone recording with an on-axis target source and off-axis interfering source used in this study.

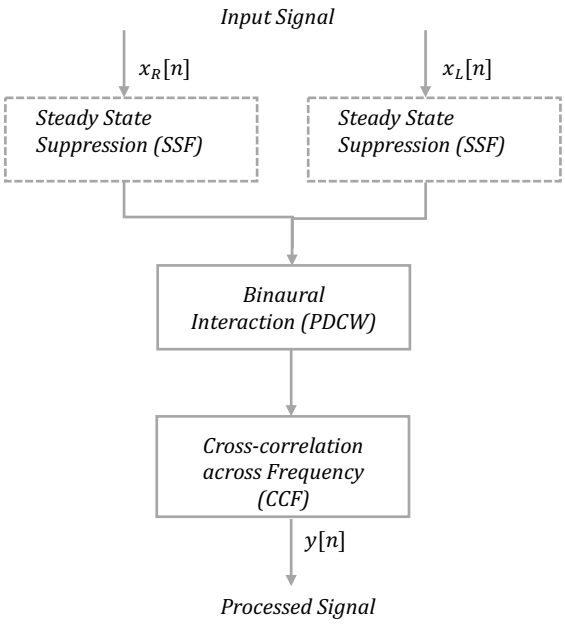

Fig. 2. Block diagram describing the overall algorithm.

weighting coefficients, in turn, are multiplied with the spectrum of the original input signal to produce the processed signal. This suppresses the falling edge of the power contour and is highly effective in reverberant environments to improve ASR performance.

In this paper, we include results both with and without SSF processing. Steady-state suppression is performed separately on each microphone channel. The application of steady-state suppression monaurally has been seen to be more effective as seen in [6].

2.2. Binaural Interaction

The optional steady-state suppression stage is followed by some sort of binaural interaction between the two microphone channels. The binaural interaction technique used in this paper is the Phase Difference Channel Weighting (PDCW) algorithm that achieves the ITD-based signal separation in the frequency domain. Results from Delay-and-Sum (DS) processing have also been presented in Section 4 as a baseline.

2.2.1. Phase Difference Channel Weighting (PDCW)

The PDCW algorithm separates signals according to ITD, in a crude approximation to human sound separation. PDCW estimates ITD indirectly, computing interaural phase difference (IPD) information in the frequency domain and then dividing by frequency to produce ITDs. Again, it is assumed that there is no delay in the arrival of the target signal between the right and left channel.

The PDCW algorithm applies a Short-Time Fourier Transform (STFT) on the input signals from the two microphones. The phase difference between signals from the two microphones is calculated from the STFT. Components of the STFT are retained if they are within zero ITD in magnitude by a threshold amount. A binary mask $\mu(k, m)$ is derived for the k^{th} time frame and the m^{th} frequency channel using the ITD $d(k, m)$ such that, $\mu(k, m) = 1$ for components with ITD less than the threshold magnitude and 0 otherwise.

While the binary mask provides a degree of signal separation by itself, we have found that recognition accuracy improves when it is smoothed over time and frequency This smoothing along frequency, called "channel weighting" in the original algorithm, is performed using a gammatone weighting function. PDCW provides substantial improvements in ASR accuracy in the presence of interfering talkers, although its performance degrades sharply in the presence of reverberation [6]. The presence of reverberation produces reflections that are added to the direct response in a fashion that leads to unpredictable phase changes, which essentially makes the ITD-estimation processing much less accurate. Since PDCW relies on oracle knowledge of the target location, this might lead to the suppression of what would have been the more viable signal acoustically. Further details about the algorithm are provided in [17].

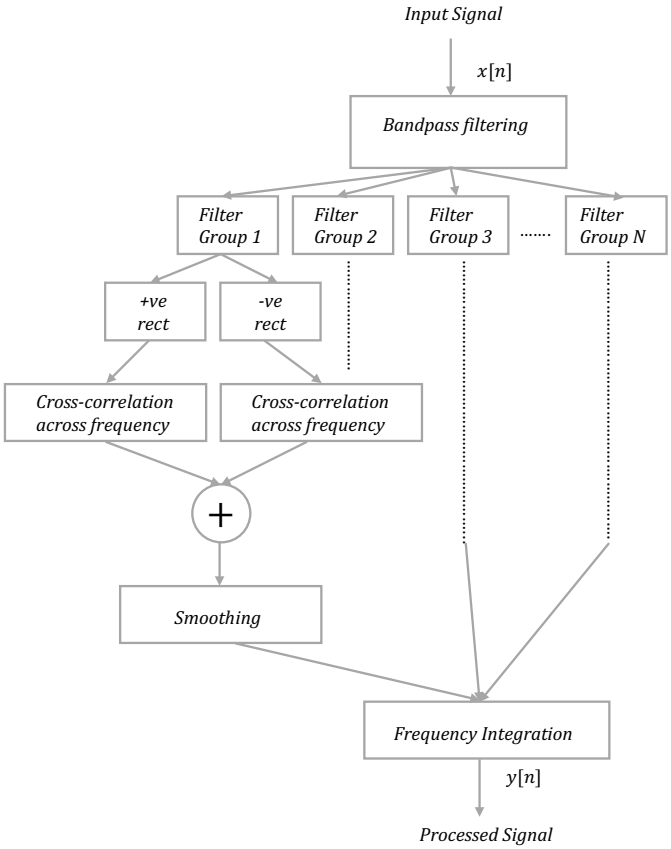

Fig. 3. Block diagram describing the Cross-Correlation across Frequency (CCF) algorithm.

3. CROSS-CORRELATION ACROSS FREQUENCY

Cross-Correlation across Frequency (CCF) is a new technique that we introduce in this study to emphasize portions of the input that are consistent across frequency. CCF is motivated by the concept of "straightness" weighting as discussed in [11]. In essence, this technique aims at boosting regions of coherence across frequency, and it also provides smoothing over a limited range of frequencies.

A block diagram describing CCF processing is shown in Figure 3. This technique roughly follows the manner in which speech is processed in the human auditory system. The peripheral auditory system is modeled by a bank of bandpass filters. We use a modified implementation of the gammatone filters in Slaney's Auditory Toolbox [26]. Zero-phase filtering is obtained by computing the autocorrelation function of the original gammatone filters, which are adjusted to roughly compensate for the reduction in bandwidth produced by squaring the magnitude of the frequency response when performing the autocorrelation operation. The center frequencies of the filters were linearly spaced according to the ERB scale [27]. For each of these filters, a secondary set of satel-

lite filters is designed. The total span of these satellite filters determine the range of frequencies over which CCF will be performed.

In other words, a total of N groups of bandpass filters are created, each with one "center" band and $m/2$ satellite bands on either side of the center band in frequency. Here, m represents the total number of satellite bands. Since the satellite bands are symmetric about the center band, m is always even. These N filter groups are denoted by "Filter Group 1", "Filter Group 2" ...."Filter Group N" in Figure 3. Each of these filter groups consists of one center band and the corresponding satellite bands. The center frequency of the l^{th} pair of satellite filters on each side of the filter group center band is given by,

$$CB \pm s \times \alpha^{\frac{m}{2}+1-l}, \quad 1 \leq l \leq m/2 \tag{1}$$

where CB is the center band frequency for a given filter group, s is a parameter that determines the span of the frequencies on either side of the center band frequency and α is a parameter that controls the spacing between the satellite filters. In this study, α was set to 0.7 which produces more closely spaced satellite filters closer to the center band and wider spacing away from the center band. N was set to 20 and m was set to 6. The span parameter s was set to 2500 Hz.

Given the input signal $x[n]$, the filter outputs for a given filter group are given by

$$x_{kp}[n] = x[n] * h_{kp}[n] \tag{2}$$

where $x_{kp}[n]$ is the filter output of the k^{th} band of the p^{th} filter group, with $x[n]$ as input. Here k ranges from 1 to $m+1$ (comprising of m satellite bands and 1 center band) and p ranges from 1 to N.

Bandpass filtering is followed by a rough model of auditory nerve processing, which includes half-wave rectification of the filter outputs. Following our earlier work in "polyaural" processing with multiple microphones [28], the filter outputs are also negated and similarly half-wave rectified. While this component of the processing is non-physiological, it enables the entire signal to be reconstructed, including positive and negative portions. Cross-correlation across frequency is then computed within each individual filter group as shown below,

$$\begin{aligned} X_{fcorr+p}[n] &= \prod_{k=1}^{m+1} x_{+kp}[n] \\ X_{fcorr-p}[n] &= \prod_{k=1}^{m+1} x_{-kp}[n] \end{aligned} \tag{3}$$

where $x_{+kp}[n]$ and $x_{-kp}[n]$ are the positive and negative half-wave-rectified portions of the signals $x_{kp}[n]$ defined above, and $X_{fcorr+p}[n]$ and $X_{fcorr-p}[n]$ denote the cross-correlation across frequency of $x_{+kp}[n]$ and $x_{-kp}[n]$ for the p^{th} filter group. $X_{fcorr+p}[n]$ is combined with $-X_{fcorr-p}[n]$ to produce the complete cross-correlation across frequency for the p^{th} filter group, $X_{fcorr_p}[n]$:

$$X_{fcorr_p}[n] = X_{fcorr+p}[n] + (-X_{fcorr-p}[n]) \tag{4}$$

In order to limit any distortion that may have taken place, the signal is bandpass filtered again to achieve smoothing. The smoothed signal is denoted by $\tilde{X}_{fcorr_p}[n]$. To resynthesize speech, all the filter groups are then combined to produce

$$y[n] = \sum_{p=1}^{N} \tilde{X}_{fcorr_p}[n] \tag{5}$$

The results from ASR experiments using CCF in combination with PDCW and SSF processing are discussed in Sections 4 and 5.

WER for $RT_{60} = 0$	0 dB	10 dB	20 dB	Clean
Delay and Sum	80.78%	32.01%	12.72%	6.54%
PDCW	23.01%	11.48%	8.15%	6.51%
PDCW+CCF	18.19%	11.48%	8.49%	7.48%
PD+CCF	17.86%	10.61%	8.32%	7.48%
SSF	80.34%	31.31%	12.99%	6.82%
SSF+PDCW+CCF	20.98%	12.21%	9.37%	8.51%

WER for $RT_{60} = 0.5s$	0 dB	10 dB	20 dB	Clean
Delay and Sum	95.95%	85.96%	66.44%	56.92%
PDCW	95.36%	86.64%	73.31%	66.63%
PDCW+CCF	94.56%	82.14%	68.53%	63.75%
SSF	97.14%	63.93%	35.03%	25.97%
SSF+PDCW+CCF	84.65%	48.77%	32.53%	26.15%

WER for $RT_{60} = 1s$	0 dB	10 dB	20 dB	Clean
Delay and Sum	96.04%	92.5%	86.12%	82.52%
PDCW	96.08%	93.32%	89.08%	85.54%
PDCW+CCF	96.79%	93.84%	87.27%	84.18%
SSF	96.51%	78.96%	59.1%	52.17%
SSF+PDCW+CCF	92.59%	68.2%	53.27%	46.78%

Table 1. Comparison of algorithms with respect to Word Error Rate as a function of Signal-to-Interferer Ratio for reverberation times of 0, 0.5 and 1 s for the RM1 database (Lowest WER for each condition highlighted)

4. EXPERIMENTAL RESULTS

ASR experiments were conducted using the CMU SPHINX-III speech recognition system and the DARPA Resource Management (RM1) and Wall Street Journal (WSJ) databases [29]. The training set for RM1 consisted of 1600 utterances and the test set consisted of 600 utterances. For WSJ, these

numbers were 7138 and 330 respectively. Features used were 13^{th} order mel-frequency cepstral coefficients. Acoustic models were trained using clean speech that had undergone the same type of processing as the algorithm being tested.

We used the RIR simulation package [30] which implements the well-known image method [31] 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 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. The microphones and speakers are 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 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.

All results from the ASR experiments using the RM1 database are tabulated in Table 1. The lowest Word Error Rate (WER) obtained for each condition is highlighted. We plot a selection of important results from Table 1 in Figure 4. Results using the WSJ database are similarly shown in Figure 5.

Considering first the performance of the older compensation algorithms PDCW and SSF as described in Table 1 and Figs. 4 and 5, we note that PDCW provides excellent compensation for noise in the absence of reverberation, but PDCW becomes less effective as the RT_{60} is increased from 0 to 1 seconds. SSF, in contrast, provides a good improvement in recognition accuracy in the presence of reverberation but its effectiveness is limited by the presence of interfering noise sources. Adding CCF to PDCW and SSF provides an even further drop in WER, especially at low and moderate Signal-to-Interferer Ratios (SIR).

Figure 4 (a) depicts the performance of some of the algorithms that provided the lowest WER in the absence of reverberation for RM1. Let us consider for the moment the performance of the algorithms PDCW, PD (which is PDCW without the smoothing along the frequency axis provided by convolving with a kernel in the shape of a gammatone response) and the CCF algorithm, which also provides smoothing over frequency. As was mentioned in Sec. 2.2.1, the use of the binary mask alone in the PDCW and PD algorithms provides signal separation. The PD+CCF method shown in Figure 4 (a) replaces the smoothing in PDCW provided by channel weighting (CW) with the smoothing provided by CCF. The use of PD+CCF leads to a 22% relative drop in WER at 0 dB and an 8% relative drop at 10 dB compared to the use of PDCW alone. At higher SIRs, the opportunity for improvement reduces drastically and the WER for PD+CCF provide slightly worse accuracy than using PDCW alone. For the WSJ database, as seen in Figure 5 (a), the improvement provided by CCF is clear at low SIRs in the absence of reverberation,

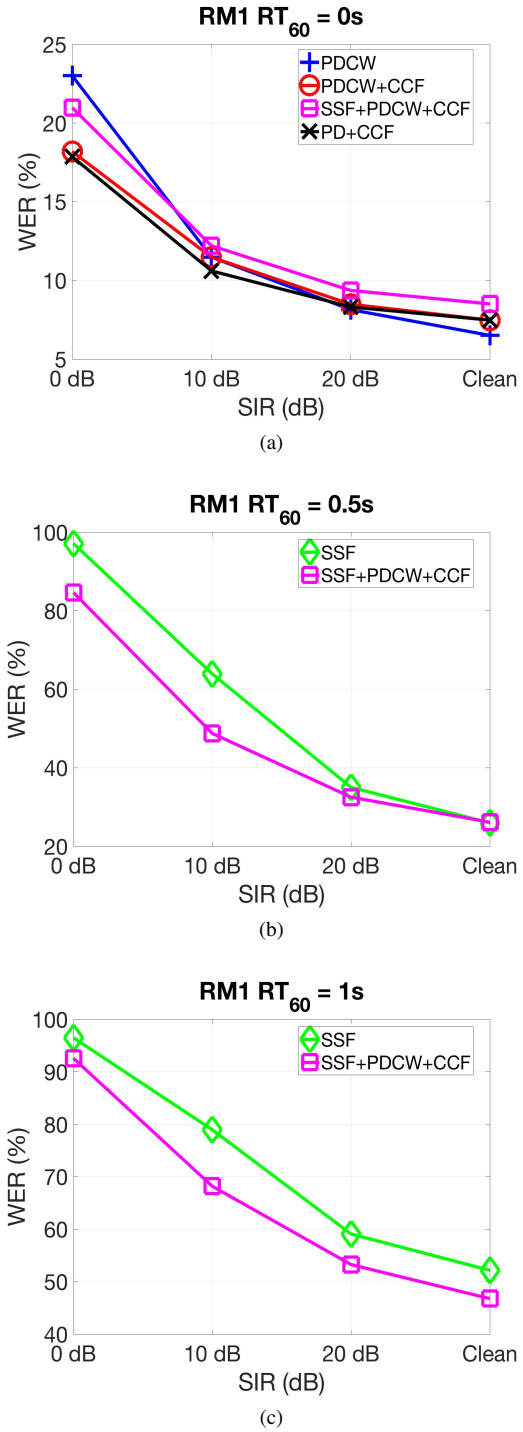

Fig. 4. Word Error Rate for the RM1 database as a function of Signal-to-Interferer Ratio for an interfering signal located 45 degrees off axis at various reverberation times: (a) 0 s (b) 0.5 s (c) 1 s.

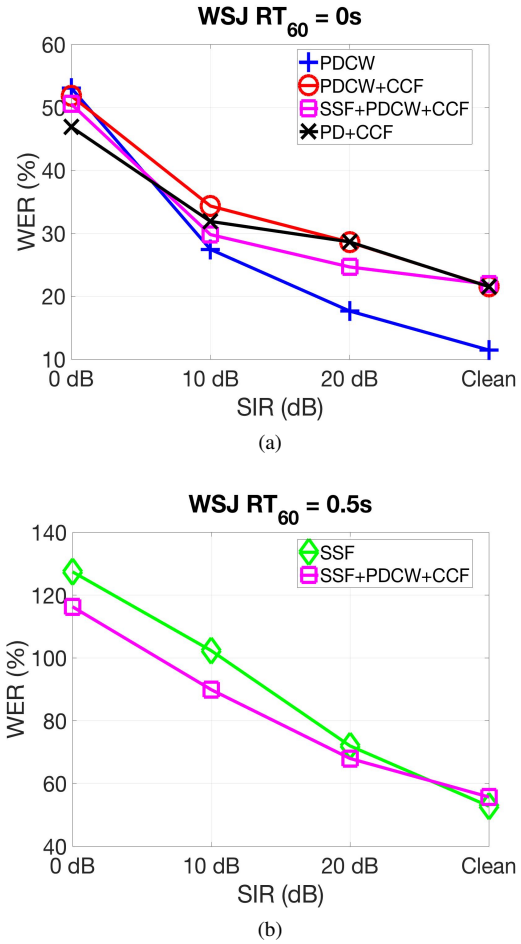

Fig. 5. Word Error Rate for the WSJ database as a function of Signal-to-Interferer Ratio for an interfering signal located 45 degrees off axis at various reverberation times: (a) 0 s (b) 0.5 s

but PDCW alone performs better than the other algorithms for higher SIRs.

Some form of steady state suppression such as the SSF algorithm is required to achieve improvements in ASR in reverberant environments, as seen Table 1 and Figures 4 and 5. As seen in Figure 4 (b) and (c) and Figure 5 (b), combining CCF with SSF and PDCW gives significant gains over using SSF alone. In the presence of reverberation, the contribution of PDCW to ASR improvement is limited. However, in combination with SSF and CCF, the improvements are significant. This is especially the case at moderate SIRs. The use of SSF+PDCW+CCF gives a relative improvement of nearly 24% at 10 dB compared to using SSF alone for the 0.5 s reverberation time case for RM1 as seen in Figure 4 (b). For WSJ, these improvements are slightly lower (close to 12% at 10 dB). These trends, however are quite consistent and hold even at reverberation time of 1 s as seen in Figure 4 (c).

5. DISCUSSION

Reviewing the results described above, we observe that PDCW works best in the absence of reverberation and gives considerable improvements at low SIRs. The CCF algorithm can be thought of as a method to enhance this binaural interaction by both reinforcing coherence and providing local smoothing across frequencies. This is why combining the CCF algorithm with any form of binaural interaction usually leads to significant improvements compared to using binaural interaction alone.

In the presence of reverberation, it becomes necessary to employ some form of steady-state suppression (SSF, in this case) to obtain better recognition accuracy. With the help of SSF in dealing with reverberation, PDCW+CCF could then be used to isolate the target speaker from the interfering talkers. This is consistent with the results wherein SSF+PDCW+CCF outperformed SSF for both reverberation time of 0.5 s and 1 s. Needless to say, all of these algorithms outperformed the Delay and Sum baselines by a huge margin.

It is interesting to note that the combination with CCF provides the most significant gains at low SIRs in the absence of reverberation and at moderate SIRs in the presence of reverberation. We believe that this has to do with the interaction of SSF and PDCW. In the absence of reverberation, PDCW is most helpful at low SIR, with and without the combination with CCF. SSF, on the other hand, helps the most at high SIRs in the presence of reverberation while PDCW performs worse at high SIRs in reverberation. For these reasons, we believe that the combination of SSF+PDCW+CCF gives the most significant gains in WER at moderate SIRs in the presence of reverberation. As seen in Section 4, the best overall gains in reverberation were at 10 dB.

6. SUMMARY AND CONCLUSIONS

In this paper, we discuss a new technique for improved recognition of binaural speech. This technique exploits coherence in frequency for monaural and binaural signals. Combined with steady-state suppression, this technique significantly improves recognition in the presence of reverberation and masking noise.

7. ACKNOWLEDGMENTS

This research was supported by the Prabhu and Poonam Goel Graduate Fellowship Fund.

8. REFERENCES

[1] Michael L Seltzer, Dong Yu, and Yongqiang 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] Xue Feng, Yaodong Zhang, and James 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] Jens Blauert, *Spatial hearing: the psychophysics of human sound localization*, MIT press, 1997.

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

[5] Chanwoo Kim, Kshitiz Kumar, and Richard 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] Richard M Stern, Chanwoo Kim, Amir Moghimi, and Anjali Menon, "Binaural technology and automatic speech recognition," in *International Congress on Acoustics*, 2016.

[7] Kalle J Palomäki, Guy J Brown, and DeLiang 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] Bruce McA Sayers and E Colin Cherry, "Mechanism of binaural fusion in the hearing of speech," *The Journal of the Acoustical Society of America*, vol. 29, no. 9, pp. 973–987, 1957.

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

[10] H Steven Colburn and Abhijit Kulkarni, "Models of sound localization," in *Sound source localization*, pp. 272–316. Springer, 2005.

[11] R. M. Stern, A. S. Zeiberg, and C. Trahiotis, "Lateralization of complex binaural stimuli: a weighted image model," *Journal of the Acoustical Society of America*, vol. 84, pp. 156–165, 1988.

[12] R. M. Stern and C. Trahiotis, "The role of consistency of interuaral timing over frequency in binaural lateralization," in *Auditory physiology and perception*, Y. Cazals, K. Horner, and L. Demany, Eds., pp. 547–554. Pergamon Press, Oxford, 1992.

[13] R. M. Stern and C. Trahiotis, "Binaural mechanisms that emphasize consistent interaural timing information over frequency," in *Proceedings of the XI International Symposium on Hearing*, A. R. Palmer, A. Rees, A. Q. Summerfield, and R. Meddis, Eds. 1998, Whurr Publishers, London.

[14] B. Raj and R. M. Stern, "Missing-feature approaches in speech recognition," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 101–115, 2005.

[15] Nicoleta Roman, DeLiang Wang, and Guy J Brown, "Speech segregation based on sound localization," *The Journal of the Acoustical Society of America*, vol. 114, no. 4, pp. 2236–2252, 2003.

[16] S. Srinivasan, N. Roman, and DeL. Wang, "Binary and ratio time-frequency masks for robust speech recognition," *Speech Comm.*, vol. 48, pp. 1486–1501, 2006.

[17] Chanwoo Kim, Kshitiz Kumar, Bhiksha Raj, and Richard 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.

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

[19] Ruth Y Litovsky, H Steven Colburn, William A Yost, and Sandra J Guzman, "The precedence effect," *The Journal of the Acoustical Society of America*, vol. 106, no. 4, pp. 1633–1654, 1999.

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

[21] Keith 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.

[22] 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.

[23] Chanwoo Kim, *Signal processing for robust speech recognition motivated by auditory processing*, Ph.D. thesis, Carnegie Mellon University, 2010.

[24] Chanwoo Kim, Kean K Chin, Michiel Bacchiani, and Richard M Stern, "Robust speech recognition using temporal masking and thresholding algorithm.," in *INTERSPEECH*, 2014, pp. 2734–2738.

[25] Byung Joon Cho, Haeyong Kwon, Ji-Won Cho, Chanwoo Kim, Richard M Stern, and Hyung-Min 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.

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

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

[28] Richard M Stern, Evandro B Gouvêa, and Govindarajan Thattai, "Polyaural array processing for automatic speech recognition in degraded environments," in *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[29] Patti Price, William M Fisher, Jared Bernstein, and David 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.

[30] Stephen G McGovern, "A model for room acoustics," 2003.

[31] Jont B Allen and David 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.