Robust Speech Recognition Based on Binaural Auditory Processing

Anjali Menon\textsuperscript{1}, Chanwoo Kim\textsuperscript{2}, Richard M. Stern \textsuperscript{1}

\textsuperscript{1}Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA
\textsuperscript{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

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{Fig1}
\caption{Overall block diagram of processing using steady-state suppression and interaural cross-correlation based weighting.}
\end{figure}

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

\begin{align*}
\text{Input signal} & \rightarrow \text{Steady State Suppression (SSF)} \\
\text{SSF} & \rightarrow \text{Interaural Cross Correlation based weighting (ICW)} \\
\text{ICW} & \rightarrow \text{Processed signal}
\end{align*}

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. [15, 16]).

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
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 $\phi$ with respect to the two microphones.

### 2.1. Steady-State Suppression

The SSF algorithm [4, 14] 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) [17] between 200 Hz and 8 kHz. The STFT was computed with frames of length 75-ms with a 10-ms temporal spacing between frames. These longer-duration window sizes have been shown to be useful for noise with a 10-ms temporal spacing between frames. These longer-duration window sizes have been shown to be useful for noise suppression. The SSF algorithm is motivated by the modulation-frequency characteristics of the human auditory system.

The SSF algorithm was used in this study to achieve steady-state suppression. The value for $c_0$ was experimentally set to 0.01. 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

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

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

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

Using Hermitian symmetry, the rest of the frequency components are obtained and the processed speech signal $\hat{x}[n]$ 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, 19, 20]. 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 [21]. The center frequencies of the filters are \(\frac{\text{frequency}}{\text{channel}}\), followed by half-wave rectification and then applied weights. The window size \(N_w\) was set to 50 ms and the time between frames for ICW was 10 ms.

Based on \(\rho_i[m]\), the weight computation was given by,

\[
u_i[m] = \rho_i[m]^{\alpha}
\]

The nonlinearity \(\alpha\) is introduced to cause a sharp decay of \(\nu_i\) as a function of \(\rho_i\) and it was experimentally set to 3. The weights computed are applied as given below:

\[
y_i[n; m] = u_i[n] \tilde{x}[n; m]
\]

where \(y_i[n; m]\) is the short-time signal corresponding to the \(m^{th}\) frame and \(l^{th}\) gammatone channel and \(\tilde{x}[n; m]\) is the average of short-time signals \(x_R[n; m]\) and \(x_L[n; m]\) 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 [22] 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 13th order mel-frequency cepstral coefficients. Acoustic models were trained using clean speech that had undergone SSF processing.

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 [23, 24].

Figure 4 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 shown in Figures 5b-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 ICW to SSF does lead to a reduction in WER compared to performance obtained using SSF alone, at least for reverberation...
times greater than 0.5 s. As seen in Fig. 5a, for the reverberation time of 0.5 s, the WER obtained using ICW+SSF is approximately equal to the WER obtained using SSF alone. For reverberation times of 1 s, 1.5 s and 2 s, however, the improvements provided by ICW are much more significant (Figures 5b-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 around 15%-18% relative. [PLEASE CHECK THIS] The improvements in WER are almost constant 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 interfering noise and reverberation in somewhat equal measure.

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


