Sound Source Separation Using Phase Difference and Reliable Mask Selection Selection

Chanwoo Kim[1], Anjali Menon[3], Michiel Bacchiani[2], and Richard M. Stern[3]

[1]Samsung Research, [2]Google Speech, [3]Carnegie Mellon University

[1]chanw.com@samsung.com, [2]michiel@google.com, [3]{anjeli, rms}@cs.cmu.edu

Abstract

In this paper, we present an algorithm called Reliable Mask Selection-Phase Difference Channel Weighting (RMS-PDCW) which selects the target source masked by a noise source using the Angle of Arrival (AoA) information calculated using the phase difference information. The RMS-PDCW algorithm selects masks to apply using the information about the localized sound source and the onset detection of speech. We demonstrate that this algorithm shows relatively 5.3 percent improvement over the baseline acoustic model, which was multistyle-trained using 22 million utterances on the simulated test set consisting of real-world and interfering-speaker noise with reverberation time distribution between 0 *ms* and 900 *ms* and SNR distribution between 0 *dB* up to clean.

1 Motivation

1. Spatial masks may be constructed using the Angle of Arrival (AoA) information.
2. AoA is not accurate under reverberation.
3. We selectively apply masks based on their reliability.

2 Entire Structure

1. **Angle Of Arrival(AoA)** information is used for obtaining binary masks.
2. **Reliable Binary Mask Selection(RBMS)** is used to select reliable binary masks.
3. The channel mask for each channel is calculated using Channel Weighting(CW) [1].
4. For each filter bank channel, **Reliable Channel Mask Selection(RBMS)** approach is used.
5. Speech is resynthesized using Over-Lap Addition(OLA) after applying masks.

The entire structure is shown in Fig. 3.

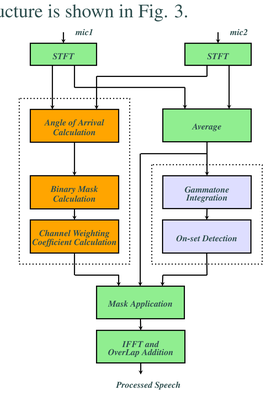

Figure 1: A block diagram showing the structure of the Reliable Mask Selection - Phase Difference Channel Weighting (RMS-PDCW) algorithm.

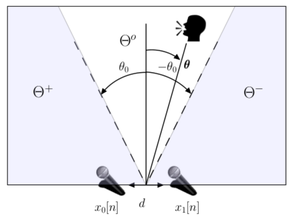

Figure 2: Two microphones and the target sound source. The space inside a room is divided into three regions depending on the azimuth angle θ: Θ^+, Θ^o, and Θ^-. We use θ_0 of 15^o.

2.1 Angle of Arrival(AoA) calculation

The AoA $\theta[m, \omega_k]$ is estimated using the following equation [2]:

$$\theta[m, \omega_k] = \arcsin\left(\frac{c_{air}\Delta\phi[m, \omega_k]}{f_s \omega_k d}\right), \quad 0 \le k \le \frac{K}{2}, \qquad (1)$$

where f_s is the sampling rate of the signal, and c_{air} is the speed of sound in air, $\Delta\phi[m, \omega_k]$ is the phase difference, and d is the distance between two microphones.

2.2 Reliable Binary Mask Selection

- For each spatial region Θ^+, Θ^o, Θ^- shown in Fig. 2, the mean and standard deviation of the AoA is calculated.
- Presence of the target source and noise sources in each region is determined using the mean and standard deviation of AoA calculated from each region.
- We apply binary masks corresponding to Θ^+, and Θ^-, only when noise source is detected in that region.
- If target source is not detected, binary masks are not applied for that frame.

2.3 Reliable Channel Mask Selection

- Channel mask is calculated using the Channel Weighting(CW) algorithm [1].
- Onset portion of spectrum is determined using [3].
- Channel mask is applied only for the onset portion.

3 Acoustic Model Training

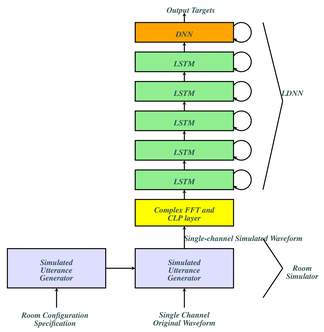

Figure 3: The architecture for acoustic model training using the room simulator and LSTMs and a DNN (LDNN) [4, 5].

4 Experimental Results

- For training, we used an anonymized 22-million English utterances (18,000-hr), which are hand-transcribed.
- For evaluation, we used around 15-hour of utterances (13,795 utterances) obtained from anonymized voice search data. We also generate noisy evaluation sets from this relatively clean voice search data.
- The "room simulator" in [5] was used to generate noisy utterances.
- For reverberation time, we used a uniform distribution from 0 seconds to 900 *ms*. For the SNR distribution, we used 0 dB, 5 dB, 10 dB, 15 dB, 20 dB, and the clean utterance in equal proportions.
- For Multistyle TRaining, we used the same configuration used in training the Google Home system [5, 6].

Table 1: Word Error Rates (WERs) obtained with multi-microphone approaches with Multistyle TRaining (MTR) using the room simulator [7].

	Clean	Simulated noisy set	Relative improvement over the baseline with MTR (%)
Baseline	11.3 %	51.7 %	-
Baseline with MTR	11.7 %	35.1 %	-
Delay and sum with MTR	11.7 %	34.9 %	0.6 %
PPDCW with MTR	11.8 %	34.4 %	3.2 %
PDCW + RCMS with MTR	11.8 %	33.6 %	4.2 %
PDCW + RBMS with MTR	11.8 %	33.3 %	5.0 %
RMS-PDCW with MTR	11.8 %	**33.2** %	**5.3** %

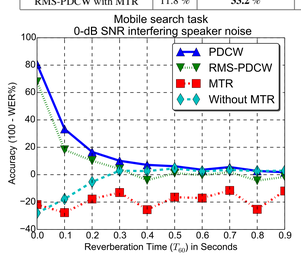

Figure 4: Word Error Rates (WERs) for the voice search test set at different reverberation time corrupted by interfering speech.

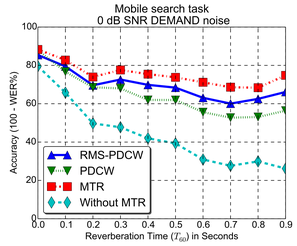

Figure 5: Word Error Rates (WERs) for the voice search test set at different reverberation time corrupted by the DEMAND noise database.

RMS-PDCW system shows relatively 5.3 percent improvement over the MTR-baseline.

5 Conclusions

In this paper, we described the RMS-PDCW algorithm which selects more reliable masks and applies them to utterances corrupted by noise and reverberation. Our experimental results show that the this algorithm shows relatively 5.3 % WER reduction over the single-channel baseline trained using the room simulator .

References

[1] 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 *INTERSPEECH-2009*, Sept. 2009, pp. 2495–2498.

[2] C. Kim, C. Khawand, and R. M. Stern, "Two-microphone source separation algorithm based on statistical modeling of angle distributions," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, March 2012, pp. 4629–4632.

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

[4] T. Sainath, R. Weiss, K. Wilson, A. Narayanan, and M. Bacchiani, "Factored spatial and spectral multichannel raw waveform CLDNNs," in *IEEE Int. Conf. Acoust., Speech, Signal Processing*, March 2016, pp. 5075–5079.

[5] C. Kim, A. Misra, K.K. Chin, T. Hughes, A. Narayanan, T. Sainath, and M. Bacchiani, "Generation of simulated utterances in virtual rooms to train deep-neural networks for far-field speech recognition in Google Home," in *INTERSPEECH-2017*, Aug. 2017, pp. 379–383.

[6] B. Li, T. Sainath, A. Narayanan, J. Caroselli, M. Bacchiani, A. Misra, I. Shafran, H. Sak, G. Pundak, K. Chin, K-C Sim, R. Weiss, K. Wilson, E. Variani, C. Kim, O. Siohan, M. Weintraub, E. McDermott, R. Rose, and M. Shannon, "Acoustic modeling for Google Home," in *INTERSPEECH-2017*, Aug. 2017, pp. 399–403.