A novel speech enhancement method based on multiple-microphone arrays

Bo Fu, Yijia Wang, Dan Zou, Wenbo Yang

Lenovo Research

fubo5@lenovo.com

Abstract

In this paper we proposed an approach to enhance the far-field speech by removing the reverberation and the noise in the audio. Firstly the absolute volume and the volume variance of each microphone array are calculated for every minute. The microphone arrays with the largest two volumes are regarded as the arrays that are closest to the speakers during that period. A smooth window is also applied to remove the noise affection. Secondly Weighted Prediction Error (WPE) is adopted using the selected two microphone arrays for every minute. This method minimizes the sum of the squared prediction errors normalized by the source variances so that it can minimize the reverberation of the speech. Then Beamforming provided by the baseline [**?**] is applied on the dereverberated audios. At last, a post-process of spectral subtraction is applied on the audio to remove the noise in the background. In the last step, the audio is sliced into chunks and Short-Time-Fourier-Transform (STFT) are calculated for each chunk. Within each chunk the non-voice regions are detected using VAD and then these regions are averaged to form 'background-noise spectrum'(BNS). Then the chunk is updated by subtracting the BNS. The BNS is updated for each chunk by weighted averaged with previous BNSs. Experiment shows that our approach improves the Signal-to-Noise ratio of the speech very obviously and and improve the ASR result.

This approach uses purely signal processing methods and no external data were adopted.

1. Background

In this approach we contributed to both the single-array track and the multiple-array track. The methodology is based on signal processing and thus no training data is needed. We aim to improve the dereverberation and the noise in the audio. No speech separation (cocktail-party problem) method was performed at this stage.

2. Contributions

2.1. Single-array track

In this approach only one Kinect was selected to perform the speech enhancement. The audio acquired by multiple channels are processed using WPE method. Then Beamforming is applied on the processed audios. At the end a spectral subtraction method is applied to enhance the speech.

1. The following four steps describes the Weighted Prediction Error (WPE) method [**?**].

1.1. The audios from the four channels of the Kinect are cropped into one minute.

1.2. Convert the audios into STFT using 32 ms frame length and 8 ms hop.

1.3. WPE is applied on the 4 one-minute STFTs. The inverse filter is estimated to remove the late reverberation.

1.4. The four new STFTs are then transformed back to audios and concatenated with the previous audios.

2. Beamforming is then applied on the four audios and one new audio is obtained.

3. The following three steps describes the spectral subtraction method.

3.1. Select the earliest quiet frames in the first few seconds and calculate the average as the noise background spectrum.

3.2. Each following frame is compared with the noise frame. If the overall power of the frame is smaller than the noise frame, keep the original frame and update the noise frame with

$$noise_{temp} = G * noise_{\mu}^{expnt} + (1 - G) * sig^{expnt} \quad (1)$$

$$noise_{\mu} = noise_{temp}^{\frac{1}{expnt}} \quad (2)$$

Where *noise(mu)* is the new noise frame, *sig* is the current signal frame, *G* is 0.8 and *expnt* is 2.0. If the overall power of the frame is greater than the noise frame, subtract the noise frame to get the estimated clean frame.

3.3. After the new estimated frames are obtained, convert the STFT frames back to time-series audio.

2.2. Multiple-array track

In this approach multiple Kinects were selected to perform the speech enhancement.

1. The power envelope of each Kinect is calculated for every minute. Then for each minute the Kinects with the top two powers are selected to perform the multiple-array speech enhancement. A median filter is applied on the array selection to get rid of the jumpy result.

2. The Weighted Prediction Error (WPE) method is adopted in this step to do the dereverberation. All 8 channels of the selected arrays are utilized together to perform the WPE.

3. The rest steps are the same as the steps 2 and 3 in single-array track section.

3. Experimental evaluation

The results are generated by the development and evaluation datasets using our approach. No external data was used and the ASR pipeline was not modified. Only speech enhancement was applied on the test data.

Both single-array track and multiple-array track results are reported.

Note that the baseline WER for the single array is 90.97% and our system (90.52%) beats the baseline by 0.4%. But the multiple-array system performs worse on the ASR result(91.00%) though it sounds more clearly objectively. We tested the multiple-array system with the 3rd-party ASR system (Watson, IBM at `https://speech-to-text-demo.ng.bluemix.net/`) and our approach performs better. Thus

Table 1: *Overall WER (%) for the systems tested on the development test set.*

Track	System	WER
Single	System 1	...
Multiple	System 1	91.00

Table 2: *Results for the best system. WER (%) per session and location together with the overall WER.*

Track	Session		Kitchen	Dining	Living	Overall
Single	Dev	S02	93.75	90.46	89.73	90.52
		S09	89.65	90.99	87.55	
	Eval	S01	...	...	...	...
		S21	...	...	...	
Multiple	Dev	S02	93.99	92.25	89.80	91.00
		S09	91.01	90.65	87.46	
	Eval	S01	...	...	...	...
		S21	...	...	...	

we think the worse result was caused by the misaligment of the audio when performing WPE.

4. Acknowledgments

We thank the CHiME challenge organization for providing the dataset and baseline.

5. References

[1] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth 'CHiME' Speech Separation and Recognition Challenge: Dataset, task and baselines," in *Proceedings of the 19th Annual Conference of the International Speech Communication Association (INTERSPEECH 2018)*, Hyderabad, India, Sep. 2018.

[2] T. Nakatania, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation based on variancenormalized delayed linear prediction," *Annalen der Physik*, vol. 322, no. 10, pp. 891–921, 1905.