Speech Separation in Multi-Channel by ICA and in Monaural by Deep Recurrent Neutral Networks

Wei Zhou & Yufei Zhang
Dept. of Electrical and Computer Engineering, University of Rochester

Introduction

The "cocktail-party" problem, a traditional audio signal processing problem, is to recognize or isolate what is being said by an individual speaker in a mixture of speech from various speakers with noisy background. To achieve better human-human or human-machine communication, such speech separation tasks attracts people's attention.

In this project, we presents three methods to deal with speech-speech separation in multi-channel and in monaural, and speech denoising in monaural conditions, respectively.

Table 1. Part of dataset list.

Filename	Speaker	Gender	Sentence text
sp01.wav	CH	M	The birch canoe slid on the smooth planks.
sp02.wav	CH	M	He knew the skill of the great young actress.
sp03.wav	CH	M	Her purse was full of useless trash.
sp04.wav	CH	M	Read verse out loud for pleasure.
sp05.wav	CH	M	Wipe the grease off his dirty face.
sp06.wav	DE	M	Men strive but seldom get rich.
sp07.wav	DE	M	We find joy in the simplest things.
sp08.wav	DE	M	Hedge apples may stain your hands green.
sp09.wav	DE	M	Hurdle the pit with the aid of a long pole.
sp10.wav	DE	M	The sky that morning was clear and bright blue.
sp11.wav	JE	F	He wrote down a long list of items.
sp12.wav	JE	F	The drip of the rain made a pleasant sound.
sp13.wav	JE	F	Smoke poured out of every crack.
sp14.wav	JE	F	Hats are worn to tea and not to dinner.
sp15.wav	JE	F	The clothes dried on a thin wooden rack.

Datasets

The noisy speech corpus NOIZEUS [1] contains 30 IEEE sentences (produced by three male and three female speakers) corrupted by eight different real-world noises at different SNRs. The noise was taken from the AURORA database and includes suburban train noise, babble, car, exhibition hall, restaurant, street, airport and trainstation noise. A noise segment are appropriately scaled to reach the desired SNR level and finally added to the filtered clean speech signal.

Methods

Multi-channel: ICA (Independent Component Analysis)

- Solving the equation $X=AS$
- FastICA

Centering, whitening, iteration to maximize the contrast function

$$J(\boldsymbol{y}) \propto [E\{G(\boldsymbol{y})\} - E\{G(\boldsymbol{y}_{\text{gauss}})\}]^2$$

J is the *negentropy* function used to measure the *Gaussianity* of signals, G is some non-quadratic function, E denotes the expected value.

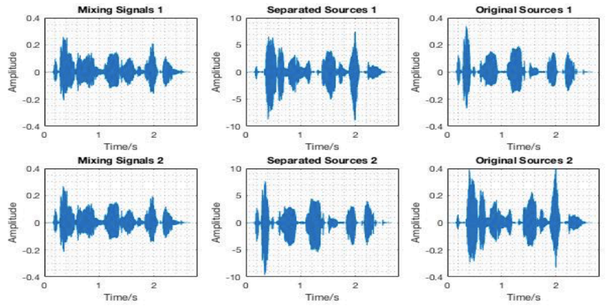
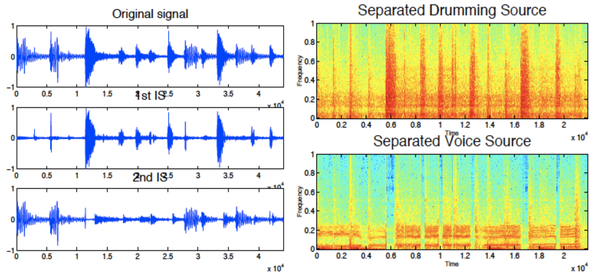


Figure 1. Applying ICA on mixing time signals

Monaural: ISA (Independent Subspace Analysis)

Figure 2. Demonstration of non-stationary ISA on

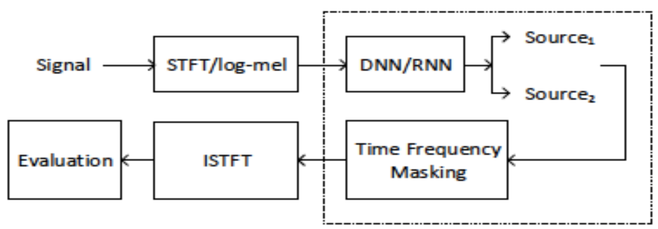


- Analysis in the time-frequency plain
- Decomposing the spectrogram into blocks to deal with non-stationary sound sources
- Building low dimensional subspaces by *Singular Value Decomposition*
- Performing ICA algorithm
- Reconstructing the separated signals combined with phase information

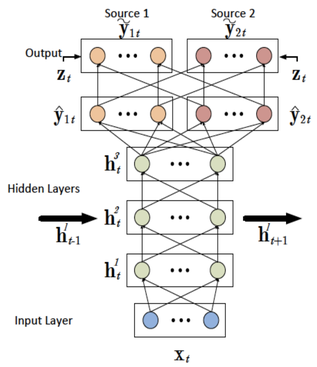
DRNN (Deep Recurrent Neural Network)

Architecture:

- Framework



- Neural Network:



Results:

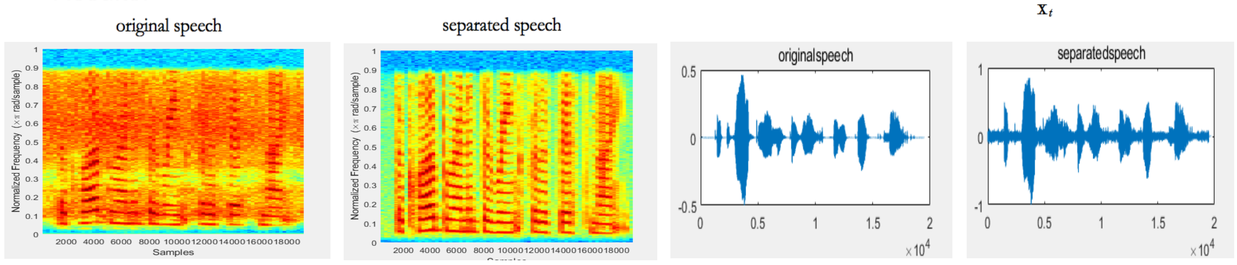


Table 2. Results of performing ICA

SAR	M_1-M_1	F_1-F_1	M_1-M_2	F_1-F_2	M_1-F_1
2	41.6266	39.3103	67.1854	39.3073	55.2952
3	41.5690	39.2454	67.1785	39.3059	55.6942
4	41.5636	38.1494	67.1784	39.3059	55.6942

Evaluation

In Table. 2, the leftmost column indicates the number of channels which are built by adding two scaled time signals of independent speech. M_i and F_i (i=1, 2) represents different male and female speaker. The evaluation criterion is sources to artifacts ratio (SAR).

While for the ISA model, we try to build corresponding algorithm as indicated by M. A. Casey and A. Westner [2]. However, our algorithms didn't achieve good separation results for lacking of effective clustering methods. So we use the experiment results presented by Virtanen [3] to give an overview of this method. His pitched instrument and drum sound separation tasks achieve about 3.6 dB overall signal to noise ratio.

For speech denoising by DRNN, We applied this method to 3 examples. The evaluation results, indicated by SDR, SIR and SAR, are shown in Table 3. The outcomes have achieved same speech separation result level comparing with the state of the art neural network techniques.

Objects	SDR	SIR	SAR
1	4.19	2.89	8.95
2	7.11	12.81	7.40
3	7.27	11.51	7.41

Table 3. Results of DRNN

Conclusions

We first introduce two traditional methods, ICA and ISA for multi-channel and monaural condition speech separation, respectively. The ICA model shows significant clear separation results. Meanwhile, the number of input mixing, that is the number of channel in some degree, won't influence the separation results clearly. And the differences in the timbre of speech won't affect the separation results as well. For ISA model, we introduce the basic theory and give a brief view on the pitched instruments and drum separation performance.

The speech denoising tasks by applying DRNN achieve quite good results. But this method requires many training data to train to extract the features ahead of time. This is not effective compared to other methods.

References

1. P. G. Shivakumar and P. Georgiou, "Perception optimized deep denoising autoencoder for speech enhancement," *Interspeech*, 2016.
2. M. A. Casey and A. Westner, "Separation of mixed audio sources by independent subspace analysis," in *Proc. Int. Comp. Music Conf.*, Berlin, Germany, 2000, pp. 154–161.
3. T. Virtanen, "Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criterion," *IEEE Transactions On Audio, Speech, And Language Processing*, vol. 15, no. 3, 2007.