Source Separation with Character Matching

Hongfei Wang, Zhong Zhang
Department of Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh, PA 15213
{*hongfeiw,zhongz*}*@andrew.cmu.edu*

Abstract

The cocktail party effect describes human's ability to follow one voice source from a mixture of conversations, and often with the addition of background noises. These conversations may happen simultaneously. However, it is tricky for computers to handle this sort of auditory source separation problems. One relatively successful approach is to use Independent Component Analysis (ICA) to separate voice sources from a mixture of them. Specifically in this project, FastICA, an efficient and popular algorithm for ICA is implemented. We then apply a correlation filter called the Minimum Average Correlation Energy (MACE) filter to match the separated voices with their characters for the purpose of identification. Both the methods have been validated by experiments.

1 Introduction

Imagine you are at a cocktail party. Different conversations may happen simultaneously, in addition to background music and noise. It is easy for human beings to follow the discussion of their neighbors. However, this is a very difficult signal processing problem for computers.

In this project of *Source Separation and Character Matching*, we apply both signal processing and machine learning techniques to help this problem. Our goals are i) separation of different people's speech in conversations, and ii) matching people's identity by their speech.

2 Problem Definition

2.1 Source separation

Assume there are ***M*** speakers and ***N*** microphones in a party to record various conversations. The source signal are represented as

$$S = \begin{bmatrix} S_1 & S_2 & \cdots & S_{M-1} & S_M \end{bmatrix}^T$$

where random variable S_m represents the *m*th speaker's voice. And the observation signal is

$$X = \begin{bmatrix} X_1 & X_2 & \cdots & X_{N-1} & X_N \end{bmatrix}^T$$

where random variable X_n represents the *n*th microphone's record. Ignoring multipath effect and noise, the output of each microphone is

$$X_n = \sum_{m=1}^{N} A_{nm} S_m \ .$$

The outputs of all microphones are

$$X = AS$$

The source separation problem is formulated as: given ***X***, estimate ***A*** and ***S.*** Our task in this project is to obtain ***S***.

2.2 Character matching

Once we obtain the separated sources $\hat{S}$, we need to match it with people's identities. For each candidate person, we build a correlation filter ***H_i*** by

$$H_i = F(S_{training})$$

where ***F*** is the function to construct correlation filter, $\boldsymbol{S}_{\text{training}}$ is the voice recordings from the $\boldsymbol{i}^{\text{th}}$ person to train the filter. Therefore the matching problem can be formulated to maximize the correlation response as

$$\max_i \quad H_i(\hat{S}_i)$$

which means if a separated sources matches a correlation filter, then it will be regarded as the person whose voices are used to build the filter.

3 Methodology

2.1 Independent Component Analysis (ICA)

Some assumptions have been made about sources. Different people have different voice pattern, from common sense. It is thereby reasonable to assume their voices are independent. Independent Component Analysis (ICA) provides a statistical technique for revealing independent sources given mixed observations [1].

Given the independence assumption, original source separation problem is converted to an optimization problem

$$\max_W I(\hat{S}_1, \hat{S}_2, \cdots, \hat{S}_N)$$
$$st.: \ \hat{S} = WX$$

where $W \in R^{N \times M}$ denotes separation matrix and $I(\hat{S}_1, \hat{S}_2, \cdots, \hat{S}_N)$ is the cost function evaluating the independence between estimated source $\hat{S}_1, \hat{S}_2, \cdots, \hat{S}_N$. The formulation of cost function $I(\bullet)$ varies in differently applications. Mutual information[2], Infomax[3] and Renyi information[4] are several widely used criteria. In this project we adopted the negentropy cost function defined in [5]. A random variable $\hat{S}_i$'s negentropy $J(\hat{S}_i)$ is defined as

$$J(\hat{S}_i) = H_G(\hat{S}_i) - H(\hat{S}_i)$$

where $H(\hat{S}_i)$ is the entropy for $\hat{S}_i$ and $H_G(\hat{S}_i)$is the entropy of the Gaussian variable with the variance for $\hat{S}_i$. The cost function in (1) is

$$I(\hat{S}_1, \hat{S}_2, \cdots, \hat{S}_N) = \sum_{n=1}^{N} J(\hat{S}_i).$$

[6] introduces an novel approximation approach of negentropy. In the simplest case, the approximation is of this form:

$$J(\hat{S}_i) \approx [E\{f(\hat{S}_i)\} - E\{f(v)\}]^2$$

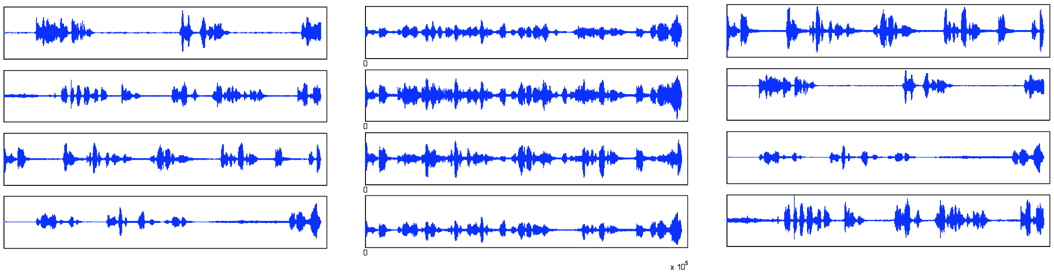
where f is practically any non-quadratic function and ν is a Gaussian variable with the same variance as $\hat{S}_i$. We choose $f(\hat{S}_i) = \hat{S}_i^4$ in implementation.

Table 1: Fast ICA algorithm

Step 1:	Use SVD[1] decomposition to get whitening matrix Q and whitened observation data $Z = QX$ that $E(Z) = 0, \mathrm{cov}(Z) = I$
Step 2:	Let $W \in R^{N \times M}$ represent the separation matrix. Randomly initialize W. Normalize each row of W. Let W_n denotes the nth row of W. Set $n = 1$
Step 3:	$E\{Zf(W_n Z)\}^T - E\{f'[W_n Z]W_n\} \to W_n$
Step 4:	$W_n - \sum_{j=1}^{n-1} < W_j, W_n > W_n \to W_n$
Step 5:	$\frac{W_n}{\|W_n\|_2} \to W_n$

Step 6:	If W_n converges: n++;
	If n<=N: go back to Step 3;
	Else: Finish

[7] proposed a fixed point method to solve the optimization problem. The detail algorithm is shown in Table 1. This algorithm is robust and has outstanding convergence speed. Simulation result is presented in Fig 1.



(a) Speakers's voice (b) Microphones' record (c) Separated voice

Fig. 1: Simulation results of source separation. Each person's voice lasts 30s. Sampling rate is 44000 Hz. The proposed algorithm separate all sources in 0.3734s (Core i5 2.4GHz CPU). Separation result is shown in Fig. 1(c). The difference between separated voice and original voice is negligible.

2.1 Minimum Average Correlation Energy (MACE) Filter

Correlation filter is used to measure how closely two signals matches. Let ***a***, ***b*** denote two signals, the mean square error between them is defined as [8]

$$e = \| a - b \|^2 = a^T a + b^T b - 2a^T b.$$

In the three product terms of the above expression, the first two are the energies of ***a*** and ***b*** respectively. $a^T b$ is the correlation term. Therefore to minimize error equals to maximize the correlation terms.

Matched filters are one kind of correlation filter that does this maximization. They essentially are just the replicas of the patterns that we are trying to find. This leads to the problem of computationally and memory expensive. Moreover, the fact that even small changes of the objects will produce bad results means the test pattern should be in the training data when constructing the filter, making matched filter less useful for many real-life application scenarios.

One improvement from the match filters is the Minimum Average Correlation Energy (MACE) filter. It has better discriminative ability and can handle patterns with small distortions. Figure 2 briefly describes how MACE filter works for voice signal matching [9].

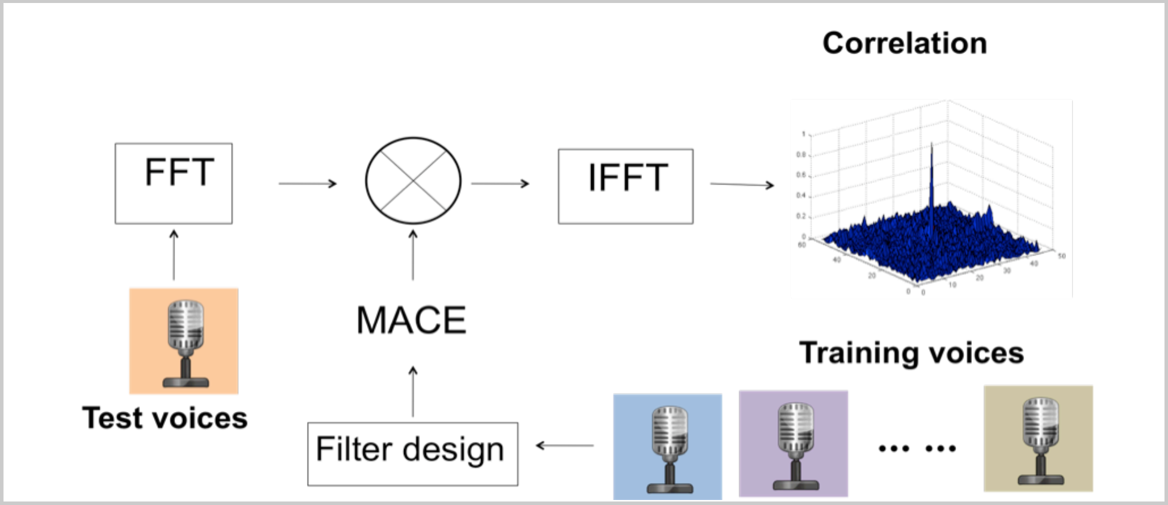
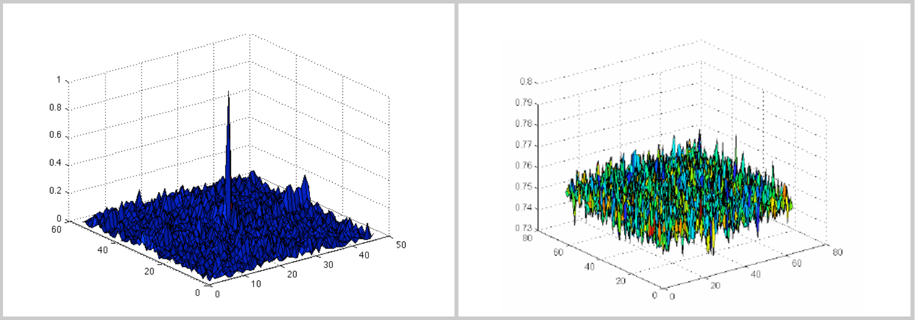


Figure 2: Block diagram showing how MACE filter works for voice matching. Both training and test voices are processed and computed for matching in frequency domain using FFT. Correlation calculation for measuring matching is very similar to convolution. The only difference is that we need to take the conjugate of the filter before doing convolution. The results are transferred back to spatial domain for discriminative analysis using IFFT.

The output response in Fig. 2 has a sharp peak in the center of the response plain, showing that the test signal and the filter model align well. This is further explained in Fig. 3 as the following.



(a) A good match of a signal and a filter

(b) A bad match of the two

Figure 3: Example matching response from a MACE filter for two input signal. A sharp and high peak exists in (a) while no such peak can be found in (b).

The peak in Fig. 3(a) is used to calculate the Peak to Sidelobe Ratio (PSR)

$$PSR = \frac{Peak - mean}{\sigma}$$

which measures how well the signal matches the filter. Peak must not only be high but sidelobes must be small in order to produce a large PSR. In Fig. 3, PSR of (a) is much larger than (b).

Detailed construction of MACE filter can be found in [10].

4 Experiments

A microphone array is best for recording, but it is far too expensive for our project. Mixed audios with original separate sources are also difficult to find. Therefore we will instead download individual sources from the internet. We also use recordings of our selves. We then segment them into small periods to construct MACE filters, one for each person. After that we mix some segments of recording ourselves by a random matrix. Then ICA is applied to separate sources. Last we use MACE filter to match voice with characters. The whole procedure is detailed described in the readme file for the demo.

Matching results is analyzed as follows.

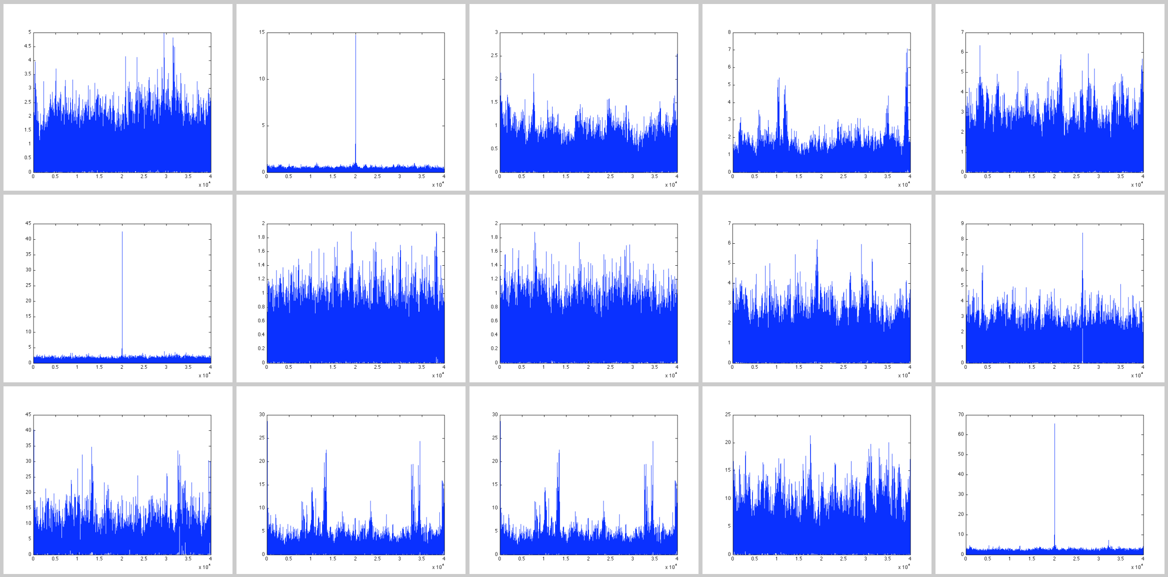


Figure 4: PSR analysis for matching using MACE filters. In this example, we have five MACE filter models, and three inputs. It is trial to find they match 2nd, 1st, and 5th filter respectively, meaning the 1st input voice is from the 2nd person in our training database, etc.

The qualities for separate voice sources are pretty good. Though sometimes we can hear there is a very low voice from other persons added to one separated voice, as well as small noise. **Noticed** the output separated voice source is not the same as original recordings and by no means exists in the training database. We obtain 100% accuracy for matching.

5 Conclusions

In this project, we applied signal processing technique (ICA) and machine learning method (MACE filter) for the problem of source separation and character matching. One possible application is voice authentification, similar to use fingerprints. Anther application is to help crime analysis. For example, the police may record conversations from some bars where drug dealing and/or sex abuse happens. Then by pre-recording the suspects' voices to train the filters, they can use our approach to separate one individual voice and match their identity, or simply to find evidence if a suspect was present on a criminal spot or not.

Future work including real-time training separation and matching.

Acknowledgments

We would like to thank Prof. Bhiksha Raj for his feedback, as well as the course TAs' generous help thru the semester.

References

[1] S. Roberts, and R. Everson, *Independent Component Analysis: Principles and Practice*. Cambridge University Press, 2001.

[2] G. A. Darbellay, and P. Tichavsky, “Independent component analysis through direct estimation of the mutual information,” In *Proceedings of Independent Component Analysis and Blind Signal Separation*, pp. 69-74.

[3] D. Obradovic, and G. Deco, “Information Maximization and Independent Component Analysis: Is There a Difference?” *Neural Computation*, vol. 10, no. 8, pp. 2085-2101, 1998.

[4] K. E. Hild, D. Erdogmus, and J. C. Principe, “Blind source separation using Renyi's Mutual Information (vol 8, pg 174, 2001),” *IEEE Signal Processing Letters*, vol. 10, no. 8, pp. 250-250, Aug, 2003.

[5] T. W. Lee, M. Girolami, A. J. Bell et al., “A unifying information-theoretic framework for independent component analysis,” *Computers & Mathematics with Applications*, vol. 39, no. 11, pp. 1-21, Jun, 2000.

[6] A. Hyvärinen. “New approximations of differential entropy for independent component analysis and projection pursuit,” In *Advances in Neural Information Processing Systems*, volume 10, pp. 273–279. MIT Press, 1998.

[7] A. Hyvärinen. “Fast and Robust Fixed-Point Algorithms for Independent Component Analysis,” *IEEE Trans. on Neural Networks*, 1999, http://www.cs.helsinki.fi/u/ahyvarin/papers

[8] M. Savvides. Lectures from *Pattern Recognition Theory*, ECE Department, Carnegie Mellon University, Spring, 2010.

[9] M. Savvides. B.V.K. Vijaya Kumar and P. Khosla, “Face Verification using Correlation Filters,” in Proceeding of Third IEEE Automatic Identification Advanced Technologies, pp. 56-61, 2002.

[10] A. Mahalanobis, B.V.K. Vijaya Kumar and D. Casasent, “Minimum Average Correlation Energy Filters,” *Applied Optics*, Vol. 26, No. 17, pp. 3633-3640, September, 1987.