

In-Vehicle Speaker Recognition Using Independent Vector Analysis

Toshiro Yamada, Ashish Tawari and Mohan M. Trivedi

Abstract—As part of human-centered driver assist framework for holistic multimodal sensing, we present an evaluation of independent vector analysis for speaker recognition task inside an automotive vehicle. Independent component analysis-based blind source separation algorithms have attracted attentions in recent years in the application of speech separation and enhancement. Compared to the traditional beamforming technique, the blind source separation method may typically require less number of microphones and perform better under reverberant environment. We recorded two speakers in the driver and front-passenger seats talking simultaneously inside a car and used independent vector analysis to separate the two speech signals. In the speaker recognition task, we show that by training the model with the speech signals from the IVA process, our system is able to achieve 95 % accuracy from a 1-second speech segment.

I. INTRODUCTION

Safety in automotive vehicles is a major concern as the in-vehicle technologies and infotainment systems become more feature-rich and complex, and cause drivers to be distracted. In an effort to make vehicles safer, researchers from a wide range of disciplines have come together to develop human-centered driver assistance (HCDA) systems using a holistic multimodal sensing approach [1], [2]. The goal of these systems is to capture dynamic context of the holistic environment, vehicle and driver (EVD), analyze the data using model-based approach and safety metrics, and display or alert the information back to the driver. As part of the HCDA framework, being able to robustly capture the speech and identify the driver and passengers is a critical part of the holistic approach, allowing the HCDA system to be personalized.

Recent automobiles are often equipped with voice communication and speech recognition systems, which allow voice commands as an interface to the system and reduce the risk of accidents caused by drivers being distracted from the touch or control interfaces. These systems often use a microphone array [3], speech enhancement technique, or a combination of both [4], [5] to reduce car noise and acquire clean speech signal. Using a microphone array, it is also possible to separate the driver and passenger speech signals. There are two commonly used microphone array techniques for separating the speeches: beamforming and blind source separation (BSS). The classical beamformer constructs a directional signal reception by filtering or delaying the elements and summing the signals. It requires the knowledge of the source direction in order to be able to direct the array

to a particular source, and the performance is constrained by the array geometry. Independent component analysis (ICA), a BSS approach, on the other hand, is a statistical method that separates statistically independent sources without the knowledge of their directions.

In addition, personalized HCDA system allows the EVD data analysis and model to include the preference of the individual so that the system response and output is customized. Since many voice communication systems in cars are already equipped with a microphone (and often an array of microphones), it is natural to prefer the use of an audio-based recognition system. Speaker recognition is a popular and mature field of research, and has a rigorous annual evaluation of state-of-the-arts systems conducted by the National Institute of Standards and Technology (NIST) since 1996.

In this paper, we evaluate speaker recognition performed inside a car using independent vector analysis (IVA), which is an extension of frequency-domain ICA, as a method for acquiring clean speech signals. In the rest of the paper, we present related work in Section II and describe the overview of the system used in this paper in Section III. We introduce the problem of BSS and the IVA method in Section IV. Section V describes the speaker recognition system that is used in our experiment. Finally, the experiment and the results are presented in Section VI and we conclude in Section VII.

II. RELATED WORK

In HCDA systems, audio and visual are the primary sensors for the human-computer interface (HCI) to the intelligent system. While this paper focuses on the auditory sensor, and particularly speech inputs, it is important to review the benefit of the multimodal audio-visual HCI as the goal is to integrate robust speaker recognition to the HCDA framework. A survey of extensive studies of intelligent systems with audio and visual interfaces can be found in [6]. More recently, a distributed testbed for multisensory signal acquisition has been developed to deal with spontaneous and subtle user behavior [7].

The problem of blind source separation has attracted a lot of interests in recent years as a microphone array technique for speech enhancement, and speaker localization and tracking. In the application of automotive vehicles, Saruwatari et al. [8] use null-steering beamformer to improve the spatial filter and convergence of the frequency-domain ICA for speech enhancement and speech recognition. The authors show that their proposed method has a higher noise reduction rate (NRR) compared to the conventional ICA

The authors are with the Department of Electrical and Computer Engineering, University of California, San Diego, La Jolla, CA 92093, USA {toyamada, atawari, mtrivedi}@ucsd.edu

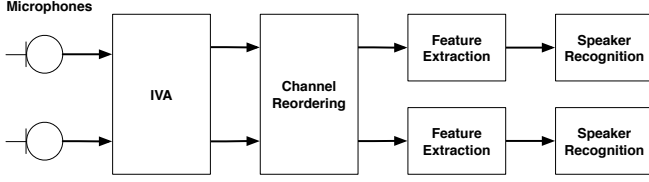


Fig. 1. Overview of the system used in the vehicle.

or null-steering beamformer, and achieves a higher NRR in fewer iterations. Beack et al. [5] proposes a different approach by combining BSS and Kalman filter to enhance speech in a car environment. Their method uses BSS to reduce spatial noise (e.g. background music and passenger's chattering sound) while Kalman filter to reduce temporal noise (e.g. engine vibration and wind).

Deployment of speech recognition systems inside automotive vehicles have lead to recent work in speaker recognition inside a car. Gottlieb et al. [9] argues that the state-of-the-art speaker recognition system based on Gaussian mixture models (GMMs) performs poorly under realistic car noise conditions and that more work in car-specific noise reduction is necessary. Herbig et al. [10] further presents a self-learning speech controlled system that comprises speech recognition, speaker identification and speaker adaptation for a small number of users. The speakers are modeled by GMMs and a standard Wiener filter is used as a front-end noise reduction. Their final combined adaptation approach evaluated on the US-SPEECON database shows a highest accuracy of 94.64 %. Beyond speech and speaker recognition for driver assistance using speech, Tawari and Trivedi have investigated real-world issues and solutions of emotion analysis and classification of the driver to accurately monitor driver's state [11], [12], [13], [14].

While there are a number of studies of BSS and speaker recognition, to our knowledge, there has not been any study that uses a BSS-based speech enhancement for speaker recognition task inside a car nor a study showing multi-interactivity of the car system.

III. SYSTEM OVERVIEW

The overview of the system used in the vehicle is shown in Figure 1. The speech signals are captured through a microphone array consisting of omnidirectional Lavalier microphones mounted on the rear-view mirror. The microphone signals are inputted to a computer through an external audio interface with integrated preamplifiers. The audio signal is first processed through the IVA algorithm. Since IVA does not preserve channel ordering at the output, it may be necessary to reorder the channels so that the first channel contains the driver's speech and the second channel contains the front passenger's speech. Once the driver and front passenger speeches are separated, the mel-frequency cepstral coefficients (MFCCs) are extracted to be used for the speaker recognition task. When the system is set for training, the MFCC features are used to train the Gaussian mixture

model (GMM) for each speaker. When the system is set for evaluation, the system searches for the high probable speaker from the GMM database. All components of the systems were developed at the Laboratory for Intelligent and Safe Automobiles (LISA). Details of each module are provided in the following sections.

IV. BLINDE SOURCE SEPARATION PROBLEM

A. Sound Mixing Model

The signals observed at the microphone array are convolved mixture of the sources, where the sources are convolved by the acoustic transfer function of the vehicle, and the noise term from the environment. Let L be the number of microphones, the observation at the i th microphone is

$$x_i(t) = \sum_{l=1}^L \sum_{\tau=0}^{T-1} h_{il}(\tau) s_l(t-\tau) + n_i(t) \quad (1)$$

where $h_{il}(t)$ is an impulse response of duration T from the l th source to the i th microphone, $s_l(t)$ is the source signal, and $n_i(t)$ is the noise observed. We assume that $n_i(t)$ is zero mean Gaussian noise. We are interested in extracting the driver's or passenger's speech signals $s(t)$ with background noise (e.g. music or other passengers) without any knowledge of the acoustic transfer functions and noise.

B. Independent component analysis

A popular approach in BSS problem is independent component analysis (ICA) [15], which is a method for estimating statistically independent non-Gaussian components [16]. Since ICA of convolved mixtures is computationally demanding, it is often performed in the frequency-domain where the mixing model is instantaneous [17]. In frequency-domain ICA, the time-domain signals in (1) are transformed using the short-time Fourier transform (STFT). The observation at the i th microphone becomes

$$X_i^{(k)}[n] = \sum_{l=1}^L H_{il}^{(k)}[n] S_l^{(k)}[n] + N_i^{(k)}[n] \quad (2)$$

where $X_i^{(k)}[n]$ denotes the frequency component at the k th frequency bin of n th time frame. Similarly, $H_{il}^{(k)}[n]$, $S_l^{(k)}[n]$ and $N_i^{(k)}[n]$ are the frequency-domain representations of $h_{il}(t)$, $s_l(t)$ and $n_i(t)$ respectively. More succinctly, we rewrite (2) as

$$\mathbf{X}^{(k)} = \mathbf{H}^{(k)} \cdot \mathbf{S}^{(k)} + \mathbf{N}^{(k)}.$$

In order to find the sources $\mathbf{S}^{(k)}$, The goal is to estimate a unmixing matrix $\mathbf{W}^{(k)}$ that allows us to find signals that are close to the original signals $\mathbf{S}^{(k)}$, such as

$$\mathbf{Y}^{(k)} = \mathbf{W}^{(k)} \mathbf{X}^{(k)}$$

where the components of $\mathbf{Y}^{(k)}$ is $Y_l^{(k)} \approx S_l^{(k)}$.

Since frequency-domain ICA finds the unmixing matrix for each frequency independently, it suffers from ambiguity of bin-wise permutation and additional processing is required to align the bins.

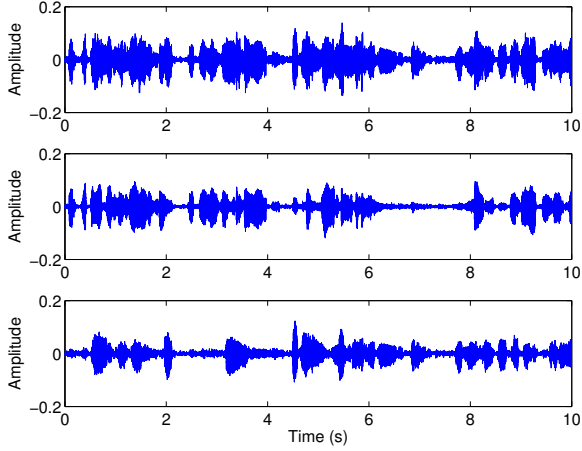


Fig. 2. An example of IVA applied to real speech signals recorded in a car. Top signal is the original recording of one of the microphones, and middle and bottom signals are separated speech signals.

C. Independent vector analysis

Independent vector analysis (IVA) is an extension of the frequency-domain ICA that reduces the permutation ambiguity considerably by treating the entire frequency bins as a vector of multivariate random variable [18], [19]. In this paper, we follow Hiroe’s algorithm [19] and choose a super-Gaussian multivariate probability density function (pdf) of the form

$$p\left(Y_l^{(k)}[n]\right) = \alpha \exp\left(-\beta \sqrt{\sum_{k=1}^K \left(Y_l^{(k)}[n]\right)^2}\right)$$

where $\beta = \sqrt{K}$.

The choice of multivariate pdf, whether it be a sub-Gaussian or super-Gaussian, is dependent on the context of application. In case of speech separation, it is known that the distribution in the frequency-domain is super-Gaussian due to the dynamic nature of speech [20]. Figure 2 shows an example of IVA applied to one of the examples in our experiment.

V. SPEAKER RECOGNITION SYSTEM

Speaker recognition refers to the task of recognizing a person given his or her speech. A speaker recognition system learns the characteristics of the individuals and use the learned model to recognize a person in the database when presented with unlabeled signal. Classical speaker recognition approaches include dynamic time warping (DTW) [21], hidden Markov model (HMM) [22], vector quantization (VQ) [23], Gaussian mixture model (GMM) [24], [25], and support vector machine (SVM) [26]. The first two methods perform text-dependent speaker recognition where the models are trained with specific speech phrases. The last three, on the other hand, can learn text-independent speaker models, where the models learn voice characteristics of the individuals. In this paper, we employ GMM to build



Fig. 3. Microphone setup for data acquisition in LISA-Q. Two microphones are positioned 4 cm apart under the rear-view mirror.

stochastic models of the feature extracted from a speech sequence of a specific speaker. We chose GMM since the number of speakers in the experiment is small and GMM have shown sufficiently high speaker recognition accuracy with our data.

An important part of the speech recognition system is feature extraction. Many successful speaker recognition systems use MFCCs as the primary feature of speech. In the experiment, we use 20 ms segments with 10 ms overlap and find 19 mel-cepstral coefficients excluding the 0th coefficients and coefficients from the first and second derivatives. The final feature is a 57-dimensional vector.

VI. EXPERIMENTAL EVALUATION

This section presents an experimental evaluation of the speaker recognition task performed on the output of the IVA process from speech signals recorded in a real car. We evaluate the performance of the Gaussian mixture model trained on three variation of speech dataset.

A. Dataset descriptions

We recorded the speech dataset at the Laboratory for Intelligent and Safe Automobiles (LISA). The dataset consist of three male and a female speakers. The dataset is comprised of three sets used for training and testing.

The first set is a clean studio recording with minimal amount of noise or reverberation in the signal. For each speaker, we recorded two sessions of two minutes speech in a fairly anechoic studio where the speaker read out loud a technical paper or news article. We call this the Studio set in the rest of the paper.

The second set is based on the same signals as the Studio set, but each signal is convolved by the acoustic impulse response of a car to simulate a recording inside a vehicle. The impulse response measurement was conducted in LISA-Q, an Infiniti Q45 family sedan, with one of the two microphones placed under the rear-view mirror shown in Figure 3, and a full range loudspeaker at the driver-side to generate the log sine sweep. We call this the Simulated set.

TABLE I
NUMBER OF TEST EXAMPLES

Test Data	Test Signal Lengths		
	1 Second	2 Seconds	5 Seconds
Studio set	470	234	92
Car-IVA set	513	256	99

The third set is the real recordings done inside a car. We recorded two speakers sitting in the driver and front passenger seats while they spoke simultaneously for two minutes with the same microphone setup in Figure 3. The car was parked on a street with moderate traffic where a bus passes by every few minutes, and the car remained stationary with all functionality of the car turned off while the recordings were conducted. The notable noise factors were cars and buses passing by in the background. We processed the recorded signals through the IVA to separate the mixed speech signals to create the dataset. The processing was done block-wise with a 15 second block size. Since block-wise IVA has permutation ambiguity of the output channels¹ between blocks, we used a simple algorithm for correcting the permutation based on the direction of arrival of the sources. The algorithm is described in the appendix for the interested readers. We call this the Car-IVA set.

We used the same microphones for all recordings with sampling rate at 16 kHz. No additional processing (e.g. normalization) was performed on the signals before the feature extraction described in Section V.

B. Methodology

The experiment is setup to evaluate training methods for speaker recognition inside a car. We set the baseline method as the GMM trained and tested on the Studio set. For our evaluation, we compare the GMM trained on the above three datasets individually and tested on the Car-IVA set. The training and test data for each set was created by dividing the signals equally in half, resulting in approximately two minutes of speech content for each speaker for both training and testing. Each GMM has 64 mixtures components with diagonal covariance matrices.

The testing is done by dividing the entire speech signal into 1, 2, and 5 second segments and treating each segment as a test example. Table I shows the number of test examples for Studio and Car-IVA sets. For each example, we extract the MFCCs and find the highest probably speaker for each frame. The speaker having the highest number of frame count is identified as the speaker for that test example (i.e. majority rule).

The performance is measured by the percentage of test examples that are identified correctly. For example, out of the 470 samples in the 1-second test signal length in the Studio set, if 376 samples are identified to the correct speaker, the accuracy is 80 %.

¹Not to be confused with bin-wise permutation ambiguity in frequency-domain ICA.

TABLE II
ACCURACY (%) OF SPEAKER RECOGNITION

Training Data / Test Data	Test Signal Lengths		
	1 Second	2 Seconds	5 Seconds
Studio / Studio (Baseline)	92	97	100
Studio / Car-IVA	62	67	78
Simulated / Car-IVA	64	73	77
Car-IVA / Car-IVA	95	99	100

C. Results

Table II shows the result of the baseline speaker recognition accuracy and speech recognition conducted on the output of IVA for the learned models. The result shows that the overall accuracy improves by training the GMM on real data, and training and testing on the Car-IVA set has the highest accuracy, which is better than the baseline rate. Training on Car-IVA set is able to achieve 95 % on the 1-second signal length, and 99 % on the 2-second signal length. On the other hand, training with Simulated set resulted in worse performance, showing that the simulated data does not give a good indicator of performance for this experiment. Furthermore, getting a better performance from the Car-IVA set indicates that the speech signal processed after IVA is consistent and reliable for both training and testing speaker models. We believe that the Car-IVA set performed better than the baseline rate because IVA worked as a speech enhancer and reduced noise and reflections present in the recordings while also achieving a high separation of the speech signals.

VII. CONCLUDING REMARKS

With an increasingly complicated in-vehicle technology and infotainment systems installed in recent automotive vehicles, it is important for an intelligent human-centered driver assistance system to use a holistic multimodal sensing approach to assist drivers to operate the car safely and decrease distractions. As part of the HCDA framework, the paper explores robust speech capturing techniques and an application to speaker recognition to allow HCDA systems to be personalized. We have used IVA as a method for extracting driver and front passenger-side speech signals, and GMM-based speaker recognition. We have shown that, for a small number of subjects, the GMM-based speaker recognition system is able to predict the speaker with 95 % accuracy for a short 1 second segment and perform better than the baseline method with the same amount of training data. The result shows that IVA is a promising front-end processor for a robust HCDA system, and shows promising application in speaker recognition.

As a continuation of our mission for designing intelligent and safe automobiles, our future work includes integration of 3D audio for delivering personalized “sound” alerts such that the HCDA system can deliver information not only auditory but also spatially to alert the driver to certain directions.

APPENDIX

The section describe the method used to to correct channel permutation when using block-wise IVA. Each block is vulnerable to channel permutation since IVA does not guarantee the order of separated speech signals. Empirically, the order of unmixed sources seems to be in the order of loudness of the sources. Thus, a simple solution that uses directivity patterns of the unmixing matrix and energy of unmixed sources is proposed.

The directivity response of the unmixing matrix $\mathbf{W}^{(k)}$ for l th source and k th bin is given as [27]

$$B_l^{(k)}(\theta) = \sum_{i=1}^L W_{li}^{(k)} \exp(j2\pi f_k d_i \sin \theta / c) \quad (3)$$

where f_k is the frequency (Hz) at k th frequency bin, d_i is the i th microphone position and c is the speed of sound. In case where both the driver and passenger are speaking simultaneously, the beam pattern of the driver ideally has a null beam at the direction of the passenger while the beam pattern of the passenger has a null beam toward the driver. The direction of arrival (DOA) of the sources can then be estimated by examining the null beams. This insight leads us to a simple method for detecting speech activity and channel permutation correction, which is similar to bin-wise permutation alignment approach using the directivity pattern [28], [29].

Accurate null directions can not be reliably estimated above the aliasing frequency determined by the microphone spacing and source direction. For a source impinging at the array at an angle θ , the aliasing frequency f_{alias} is given as

$$f_{alias} = \frac{c}{d(1 + |\sin \theta|)}.$$

The aliasing frequency is the highest ($f_{alias} = c/d$) when the source is impinging from broadside ($\theta = 0$), and the lowest $f_{alias} = c/2d$ when the source is impinging from end-fire ($\theta = \pi/2$). Moreover, the directivity response at low frequency is also unreliable. Therefore, it is reasonable to estimate the null directions from a restricted range of frequencies. Setting K_L as the low bound and K_H as the high bound of the frequency bin, the null angle is estimated as

$$\hat{\theta}_l = \frac{1}{K_H - K_L + 1} \sum_{k=K_L}^{K_H} \arg \min_{\theta} |B_l^{(k)}(\theta)|. \quad (4)$$

Given the angles of the driver $\theta^{(d)}$ and passenger $\theta^{(p)}$ and tolerable angle deviation ϕ , we estimate the speaker activity by examining the absolute difference between the estimated nulls and the given angles

$$\begin{aligned} \Delta\theta_l^{(d)} &= |\hat{\theta}_l - \theta^{(d)}|, \\ \Delta\theta_l^{(p)} &= |\hat{\theta}_l - \theta^{(p)}|. \end{aligned}$$

If at least one of the components of $\Delta\theta_l^{(d)}$ is less than ϕ , we determine that the driver is active. Similarly, passenger-side activity can be determined by examining $\Delta\theta_l^{(p)}$.

We also find the energy of the unmixed signal $y_l(t)$ as

$$E_l = \frac{1}{N} \sum_{t=0}^{N-1} |y_l(t)|^2$$

where N is the length of the signal.

Using the null angle estimates, speaker activity information, and source energy level, the channel misalignment is detected for the following three cases: (1) only the driver is active and the energy of the second unmixed channel is louder than the first channel; (2) only the passenger is active and the energy of the first unmixed channel is louder than the second channel; and (3) both the driver and passenger are active and the null beam of the first channel is directed toward the driver-side and vice versa.

ACKNOWLEDGMENT

The authors would like to thank colleagues at LISA for their suggestions and support on data collection. The authors would also like to thank the reviewers for their insightful comments.

REFERENCES

- [1] M. M. Trivedi and S. Y. Cheng, "Holistic Sensing and Active Displays for Intelligent Driver Support Systems," *IEEE Computer Society*, vol. 40, no. 5, pp. 60–68, 2007.
- [2] A. Doshi and S. Y. Cheng, "A Novel Active Heads-Up Display for Driver Assistance," *IEEE Trans. Systems, Man, and Cybernetics*, vol. 39, no. 1, pp. 85–93, 2009.
- [3] Y. Grenier, "A microphone array for car environments," *Speech Communication*, vol. 12, no. 1, pp. 25–39, 1993.
- [4] J. Meyer and K. Simmer, "Multi-channel speech enhancement in a car environment using Wiener filtering and spectral subtraction," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2. IEEE Comput. Soc. Press, 1997, pp. 1167–1170.
- [5] S. Beack, B. Lee, and M. Hahn, "Blind source separation and Kalman filter-based speech enhancement in a car environment," in *Proc. International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, 2004, pp. 520–523.
- [6] S. T. Shivappa, M. M. Trivedi, and B. D. Rao, "Audio-visual Information Fusion In Human Computer Interfaces and Intelligent Environments : A survey," *Proceedings of the IEEE*, vol. 98, no. 10, pp. 1692–1715, 2010.
- [7] A. Tawari, C. Tran, A. Doshi, T. O. Zander, and M. M. Trivedi, "Distributed multisensory signals acquisition and analysis in dyadic interactions," *Proc. ACM CHI Conference*, 2012.
- [8] H. Saruwatari, K. Sawai, A. Lee, K. Shikano, A. Kaminuma, and M. Sakata, "Speech Enhancement and Recognition in Car Environment Using Blind Source Separation and Subband Elimination Processing," in *Proc. International Symposium on Independent Component Analysis and Blind Signal Separation (ICA)*, no. April, 2003, pp. 367–372.
- [9] L. R. Gottlieb and G. Friedland, "On the Use of Artificial Conversation Data for Speaker Recognition in Cars," in *Proc. IEEE International Conference on Semantic Computing*. Ieee, Sept. 2009, pp. 124–128.
- [10] T. Herbig, F. Gerl, and W. Minker, "Simultaneous speech recognition and speaker identification," in *Proc. IEEE Spoken Language Technology Workshop*, 2010, pp. 218–222.
- [11] A. Tawari and M. M. Trivedi, "Speech Emotion Analysis in Noisy Real-World Environment," in *Proc. International Conference on Pattern Recognition*. Ieee, Aug. 2010, pp. 4605–4608.
- [12] —, "Speech based emotion classification framework for driver assistance system," in *Proc. IEEE Intelligent Vehicles Symposium*, 2010, pp. 174–178.
- [13] —, "Audio-visual Data Association for face expression analysis," in *International Conference on Pattern Recognition*, 2012.
- [14] —, "Speech Emotion Analysis: Exploring the Role of Context," *IEEE Transactions on Multimedia*, vol. 12, no. 6, pp. 502–509, 2010.
- [15] P. Comon, "Independent component analysis , A new concept?" *Signal Processing*, vol. 36, pp. 287–314, 1994.

- [16] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. Wiley-interscience, 2001.
- [17] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, no. 4, 1998.
- [18] T. Kim and T. r. Eltoft, "Independent vector analysis: An extension of ICA to multivariate components," in *Proc. Int. Conf. Independent Component Analysis and Blind Signal Separation*, 2006, pp. 165–172.
- [19] A. Hiroe, "Solution of Permutation Problem in Frequency Domain ICA , Using Multivariate," in *Proc. Int. Conf. Independent Component Analysis and Blind Signal Separation*, 2006, pp. 601–608.
- [20] A. Masnadi-Shirazi and W. Zhang, "Glimpsing IVA: a framework for overcomplete/complete/undercomplete convolutive source separation," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1841–1855, 2010.
- [21] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 29, no. 2, pp. 254 – 272, 1981.
- [22] N. Tisby, "On the application of mixture AR hidden Markov models to text independent speaker recognition," *IEEE Trans. Signal Processing*, vol. 39, no. 3, pp. 563 – 570, 1991.
- [23] F. Soong, A. Rosenberg, L. Rabiner, and B. H. Juang, "A Vector Quantization Approach to Speaker Recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1985, pp. 387–390.
- [24] D. A. Reynolds and R. C. Rose, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models," *IEEE Trans. Speech and Audio Processing*, vol. 3, no. 1, pp. 72 – 83, 1995.
- [25] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, Jan. 2000.
- [26] W. M. Campbell, J. J. Campbell, D. A. Reynolds, E. Singer, and P. A. Torres-Carrasquillo, "Support vector machines for speaker and language recognition," *Computer Speech & Language*, vol. 20, no. 2-3, pp. 210–229, 2006.
- [27] D. H. Johnson and D. E. Dudgeon, *Array signal processing: concepts and techniques*. Prentice-Hall, 1993.
- [28] S. Kuritāt, H. Saruwatari, S. Kajitāt, K. Takedat, and F. Itakura, "Evaluation of blind signal separation method using directivity pattern under reverberant conditions," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 5, 2000, pp. 3140 – 3143.
- [29] M. Z. Ikram and D. R. Morgan, "A beamforming approach to permutation alignment for multichannel frequency-domain blind speech separation," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, 2002, pp. I-881 – I-884.