6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

FEATURE VECTOR TRANSFORMATION USING INDEPENDENT COMPONENT ANALYSIS AND ITS APPLICATION TO SPEAKER IDENTIFICATION

Gil-Jin Jang, Seong-Jin Yun, and Yung-Hwan Oh
Department of Computer Science, KAIST
373-1, Kusong-dong, Yusong-gu, Taejon 305-701, Korea
{jangbal, yhoh}@bulsai.kaist.ac.kr
http://bulsai.kaist.ac.kr

10.21437/Eurospeech.1999-186

ABSTRACT

This paper presents a feature parameter transformation method using ICA (independent component analysis) for text independent speaker identification of telephone speech. ICA is a signal processing technique which can separate linearly mixed signals into statistically independent signals. The proposed method transforms them into new vectors using ICA assuming that the cepstrum vectors of the telephone speech collected from various kinds of channel conditions are linear combinations of some characteristic functions with random noise added. The performance of the proposed method was compared to the original cepstrum for the HMM-based speaker identification system. Experiments were done in equal and different channel conditions on SPIDRE, a real telephone speech database for text independent speaker identification. The identification rates increased from about 1∼13% most cases, so it was confirmed that the proposed method is effective for speaker identification systems, and more effective in adverse environments.

Keywords: speaker identification, independent component analysis, feature transformation.

1. INTRODUCTION

Speaker identification is the process of selecting the best-matched speaker among the enrolled speakers, with the personal identity information extracted from speech signals. Many techniques involving statistical or probabilistic approaches have been applied to speaker-specific patterns. However, speaker information and language characteristics are difficult to separate, because the speaker information is a static behavior of utterance dynamics. Most researchers have used a large amount of speaker-specific data to ignore language-specific information. But this kind of approach cannot separate mixed information directly. The distinction between them is a major issue in current speaker identification or verification research [6][7].

There are many techniques for separating of several mixed signals into original source signals, the so-called blind source separation (BSS) [2]. The term “blind” refers to the fact that the method of combination and source signal characteristics are unknown, so BSS permits a wide range of signals as input. Another technique called ICA [2], despite a slight constraint on linearity in mixing structure, tries to express a set of random variables with some noise as linear combinations of components that are statistically independent. Principal component analysis (PCA) [1] is another widely used technique in pattern recognition or in communication theory. Most ICA algorithms use higher-order statistical cumulants to achieve these goal while PCA algorithms use second-order cumulants. Hence, ICA generalizes PCA to produce independent signals rather than simply uncorrelated ones.

In this paper, we propose a cepstrum vector transformation method using ICA for the speaker identification system. PCA and ICA with three types of nonlinear functions are tried and compared. The performance of the speaker identification system has improved with the proposed method.

2. CHARACTERISTICS OF CEPSTRUM VECTOR SPACE

Speech signals are uttered by a speaker, collected by a microphone, and then delivered by a transmission network. It is known that the transmission line acts like a band-limited filter that causes nonlinear distortion in the frequency domain [6]. This filtering effect is multiplied with the characteristic functions of speech production, written as

$$S(\omega) = G(\omega)H(\omega)R(\omega)T(\omega) \tag{1}$$

where the capital symbols refer to the corresponding filter functions; G is the glottal pulse, H the vocal cord's transfer function, R the radiation from the mouth, and T the transmission line distortion. Such filter functions are represented as summation in the cepstrum domain. Because explicit distinction among them is very difficult, it can be assumed that the cepstrum vectors are the summation of certain filter sequences:

$$\begin{aligned} \mathbf{c}[n] &= FFT^{-1}\big(\log S(\omega)\big) \\ &= FFT^{-1}\big(\log G(\omega)H(\omega)R(\omega)T(\omega)\big) \\ &= \mathbf{g}[n] + \mathbf{h}[n] + \mathbf{r}[n] + \mathbf{t}[n] \\ &= \textstyle\sum_{i=1}^{\chi} \mathbf{f}_i[n] \end{aligned} \tag{2}$$

where the $\mathbf{f}_i$'s are vectors representing presumed non-definite filter functions. The spectral difference between two cepstrum vectors is defined by Euclidean distance. Equation 3 shows the q dimensional cepstral distance.

$$\begin{aligned} \delta(\mathbf{c}, \mathbf{c}') &= \|\mathbf{c} - \mathbf{c}'\|_{\varepsilon} \\ &= \Big\| \textstyle\sum_{i=1}^{\chi} (\mathbf{f}_i - \mathbf{f}'_i) \Big\|_{\varepsilon} \end{aligned}$$

$$= \sqrt{\sum_{j=1}^{q} \left\{ \sum_{i=1}^{\chi} (f_{ij} - f'_{ij}) \right\}^2} \quad (3)$$

In the above equation, large components or some outliers possibly dominate the whole distance. PCA has been used to solve this problem. It transforms the original vector space, resulting decorrelation and amplitude normalization [1]. The transformation is represented by the multiplication of a vector with a matrix:

$$\mathbf{y} = \mathbf{B}^T \mathbf{x}. \quad (4)$$

The transformed vector $\mathbf{y}$ has no correlation between components and each component's variance is normalized. In the second-order statistical sense, the measure of the importance is the contribution to the whole variance which is equal to the magnitude of its eigenvalues. The purpose of this transformation is to find the orthogonal axis in the order of variance, and equalize the amplitude of each axis. The separation of the added filter functions in the cepstrum domain is not guaranteed in PCA.

3. SPEAKER IDENTIFICATION WITH ICA-TRANSFORMED CEPSTRUM

ICA or other BSS techniques can be used to separate mixed signals. The required assumption in applying ICA is that the observed signals are linearly mixed. As general feature vectors for pattern recognition are non-linearly mixed, distinguishing among characteristics is intractable. However, the hormomrophic process of cepstrum parameter extraction, which transforms convolutively mixed filter functions into additive ones, motivates our investigation. ICA is applied to this problem domain and the transformed cepstrum vectors are used for the training and testing of a speaker identification system.

3.1. Independent Component Analysis

It is assumed that the observed q scalar random variables, x_1,x_2,$\ldots x_q$, are linear combinations of p unknown independent components, s_1,s_2,$\ldots s_p$, with $q \geq p$, that are mutually independent and zero-mean. These random variables construct the observed vector $\mathbf{x}$=$[x_1\ x_2\ \ldots\ x_q]^T$ and, respectively, a source vector $\mathbf{s}$. By the assumption of linear combinations, $\mathbf{x}$ is denoted as the multiplication of $\mathbf{s}$ and the corresponding mixing matrix $\mathbf{A}$ with some noise and a constant bias added.

$$\mathbf{x} = \mathbf{A}\mathbf{s} + \mathbf{e} + \mathbf{b} \quad (5)$$

where $\mathbf{e}$ is a noise and $\mathbf{b}$ is a constant bias. The aim of ICA is to find the estimate of source signal $\mathbf{s}$ and the corresponding linear combination matrix $\mathbf{A}$. Writing $\mathbf{y}$ for the estimate of $\mathbf{s}$, and $\mathbf{W}$ for the estimate of $\mathbf{A}^{-1}$, equation 5 is equivalent to

$$\mathbf{y} = \mathbf{W}(\mathbf{x} - \mathbf{b}). \quad (6)$$

The ICA problem is reduced to finding a linear combination $\mathbf{W}$ which transforms the observed signals into independent signals. Comon [2] defined statistical independence as the difference between joint entropy and marginal entropy of the estimated independent components, called mutual information. He also defined negentropy $J(\cdot)$, differential entropy normalized by the entropy of the standardized Gaussian distribution ν:

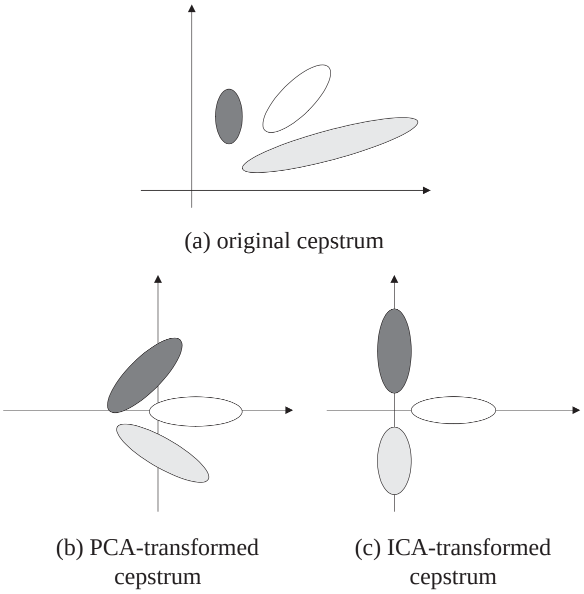

Figure 1: PCA and ICA transformation in a vector space

$$\begin{aligned} J(\mathbf{y}) &= S(\mathbf{V_y}\nu) - S(\mathbf{y}) && (7) \\ I_m(\mathbf{y}) &= \textstyle\sum_{i=1}^{p} S(y_i) - S(\mathbf{y}) \\ &= J(\mathbf{y}) - \textstyle\sum_{i=1}^{p} J(y_i) + \alpha(\mathbf{V_y}) && (8) \end{aligned}$$

where $\mathbf{V_y}$ is a covariance matrix of $\mathbf{y}$, α a function of it, and $S(\cdot)$ Shannon's differential entropy. The solution of ICA is the weight matrix $\mathbf{W}$ which minimizes the mutual information in the sense that small entropy is related to the small value of the joint probability of all random variables. Hyvärinen [3][4] derived the estimation of negentropy through higher-order cumulants by some adequate nonlinear functions G_k:

$$\begin{aligned} J_{G_k}(\mathbf{y}) &= E[G_k(\mathbf{y})] - E[G_k(\mathbf{V_y}\nu)] \\ G_1(u) &= \log\cosh(u) \\ G_2(u) &= -\exp(-\tfrac{1}{2}u^2) \text{ (exponential)} \\ G_3(u) &= \tfrac{1}{4}u^4 \text{ (kurtosis)}\,. && (9) \end{aligned}$$

G_k is called a contrast function [2]. He proposed a fast iterative algorithm using these contrast functions. The characteristics of independent signals are related to the choice of contrast function.

The effects of PCA and ICA transformations are compared in figure 1. ICA considers all possible linear transformations, while PCA can only rotate on axis to keep the original angles. In figure 1-(b), the components are just arranged along the magnitude of distribution because PCA is reflected by the size of the distribution. In (c), the shapes of the distributions are arranged to maximize the projections on each axis, for the difference between distributions is reflected by the higher-order cumulants.

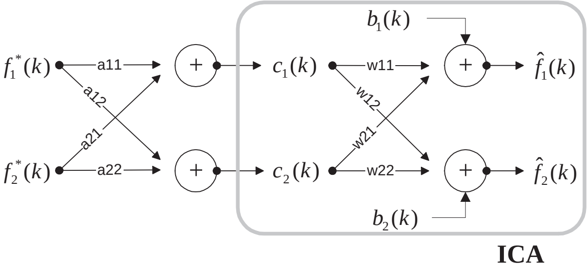

Figure 2: Analysis structure for cepstrum vectors

3.2. Cepstrum Vector Transformation

The cepstrum vectors have the property of summation; they can be regarded as linear combinations of several characteristic signals, each representing a time-varying filter. Assuming the filter functions $\mathbf{f}^*$ to be source signals of ICA, the cepstrum vectors are represented by the multiplication of a combination matrix $\mathbf{A}$ and the source signal $\mathbf{f}^*$.

$$\mathbf{c}[n] = \mathbf{A}\mathbf{f}^*[n] + \mathbf{b} \qquad (10)$$

where $\mathbf{b}$ is a constant bias. The left half of figure 2 corresponds to the above equation. Each source signal f_i^* contributes to many components of the observed cepstrum vectors with the amount of scalar values $\{a_{ij} | j = 1, \ldots, q\}$. It is difficult to measure each function's distance in original cepstrum space if it is calculated with a Euclidean norm, like equation 3. To measure each function's distance independently, ICA can be applied to transform the mixed cepstrum vector into another independent feature vector $\hat{\mathbf{f}}$:

$$\hat{\mathbf{f}}[n] = \mathbf{W}(\mathbf{c}[n] - \mathbf{b}) \qquad (11)$$

where $\hat{\mathbf{f}}$ is a vector constructed from estimated independent signals by ICA. The right half of figure 2 shows this separation network, which is similar to the architecture of a single-layered neural network. The estimated functions, $\hat{f}_i$'s, are contrasted on the nonlinear function G_k in equation 9 to calculate independence. The contrast function plays the role of a activation function in a neural network. The distance between two arbitrary vectors is equal to the summation of each component's distance in the resultant vector space.

$$\delta(\hat{\mathbf{f}}, \hat{\mathbf{f}}') = \sqrt{\textstyle\sum_{i=1}^{q} (\hat{f}_i - \hat{f}'_i)^2} \qquad (12)$$

If we restrict the normalization matrix to make the transformed components have a unit variance, the contributions of each functions to the whole distance become equal.

In addition, when the transmission function $\mathbf{t}[n]$ in equation 2 is too large, it can cause significant performance degradation due to the mismatch between the training and testing environments. If the data for estimation includes enough different channel conditions, it is possible to suppress ineffective scattered information and to give emphasis to the repetitive speech information in the resultant vector space.

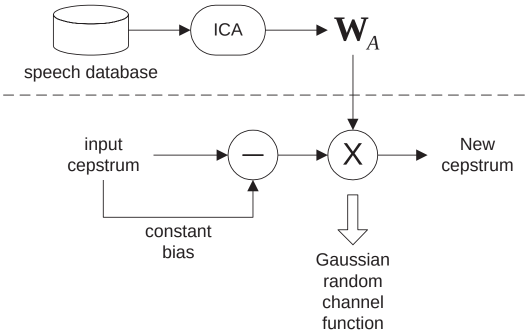

Figure 3: ICA-transformation process

3.3. Speaker Identification System

The preprocessing part of the identification system implemented in our work is shown in figure 3. Whole training vectors should be used and are needed to reflect the global transformation matrix $\mathbf{W}_A$ of the cepstrum vectors in order not to loose the possible independent components of each speaker's cepstrum space. The transformed cepstrum vectors with the ICA transformation matrix $\mathbf{W}_A$ would be used for training and testing sessions.

For speaker modeling, we use a simplified version of CHMM called HMVQM (hidden Markov VQ-codebook model) [5][8]. This incorporates observation probabilities in each state into VQ distance, instead of Gaussian mixture probability, as follows:

$$b_i(\mathbf{x}) = \exp\left(\max_k \|\mathbf{x} - \mu_{ik}\|^2\right) \qquad (13)$$

Based on the Bayesian error criterion, the probability of distortion with the nearest neighbor is at least half of the optimum probability. Two times the optimum error rate are guaranteed in this model. The advantages are fast calculation and reduction of free parameters [8]. To implement a text-independent model, fully-connected ergodic topology is adopted. Whole state transitions in this topology can map possible whole phonetic sequences of training data, because the states form broad phonetic classes in the training phase.

4. EXPERIMENTS

4.1. Database

SPIDRE is a telephone speech database for speaker identification research. It is collected from real long-distance telephone lines, with environmental noise, cross talk, transmission noise, etc. It consists of 45 target speakers, including 23 males and 22 females. From them 42 speakers are selected who have four sessions of approximately five minutes with three channel conditions. That is, we can choose two of them as DS1 in an equal channel condition, and the rest as DS2 in a different channel condition.

DS1 equal channel condition, the channel number of the training and testing data is equal, no channel mismatch exists.

DS2 different channel condition, conversely.

Table 1: Performance comparisons between transformations

(a) DS1: equal channel condition

nonlinearity	1 state	2 states	4 states	8 states
baseline	83.3%	81.5%	83.3%	84.9%
u^2 (PCA)	84.9%	82.3%	84.1%	84.9%
$\log\cosh(u)$	86.5%	83.3%	84.9%	86.5%
$-\exp(-u^2/2)$	85.7%	84.1%	85.7%	84.1%
$\frac{1}{4}u^4$	84.9%	85.7%	84.1%	85.7%

(b) DS2: different channel condition

nonlinearity	1 state	2 states	4 states	8 states
baseline	53.6%	48.4%	53.9%	55.6%
u^2 (PCA)	63.5%	61.9%	62.7%	63.5%
$\log\cosh(u)$	64.3%	62.7%	64.3%	64.3%
$-\exp(-u^2/2)$	61.9%	66.7%	64.3%	66.7%
$\frac{1}{4}u^4$	67.5%	65.9%	67.5%	61.9%

Both data sets are divided into training and testing sessions. In the training session, 30 seconds were used for each speaker. Equal amounts were used in the testing session. But they were divided into 10 second segments to evaluate each speaker's testing data three times.

4.2. Experimental Results

The original feature parameters are the 12-order mel-frequency cepstrum plus the $\log$ energy. To minimize the channel difference between the training and testing environments, the cepstral mean subtraction (CMS) [6] is applied before transformation. This is equivalent to the constant bias removal in figure 3. A baseline system is implemented for these 13-dimensional feature vectors. The PCA and ICA system also use 13-dimensional feature vectors, transformed from the original cepstrum. All three types of nonlinear contrast functions in equation 9 are used for ICA. The numbers of states are coupled with the numbers of VQ centroids. Tried configurations are (1, 32), (2, 16), (4, 8), and (8, 4).

The performances are compared with the various numbers of states and the VQ codebook size in DS1 and DS2. The results are shown at table 1. With DS1, both the PCA and ICA methods were slightly better than the baseline system. With DS2, the identification rates of both the PCA and ICA methods increased from about 1∼14% for most configurations. Kurtosis ($\frac{1}{4}u^4$) showed the best performance average among the contrast functions. It can be resoned that the ICA using kurtosis could separate outliers more definitely because the entropy estimated by kurtosis is most sensitive. Differences among speakers increased for this reason. In PCA methods with 4 states configuration, there was almost no increment in DS1 and an 8.8% increment in DS2. In ICA methods with kurtosis contrast function, although there was only 0.8% increment in DS1, there was a 13.6% increment in DS2. The performance improvement with DS1 is much larger than those with DS2. The transformation methods were more effective when the input speech was more degraded, and the ICA method was better than the PCA method.

5. CONCLUSION

We introduced a novel feature transformation method using ICA, which is a generalization of PCA. It is necessary to separate language and speaker information in speech signals to improve the performance of speaker identification systems. The proposed method uses ICA for this purpose because the characteristic of the cepstrum vector agrees with the assumption of the ICA transformation. We implemented this speaker identification system for telephone speech using ICA transformation, and improved the performance over the baseline system. The experimental results showed that the ICA method is more effective, up to 3% increment in identification rate, as environments become more adverse.

Future work includes measuring the importance of each component to remove less significant components from the feature space, and searching for new contrast functions suited to speaker recognition.

ACKNOWLEDGEMENTS

The authors would like to thank Yi Geun Chae and Yeon-Jun Kim for valuable discussions about ICA, and Ji-Hwan Kim, currently at Cambridge University, for foundations of this work.

6. REFERENCES

[1] Chih-Chien T. Chen, Chin-Ta Chen, and Chih-Ming Tsai. Hard-limited Karhunen-Loéve transform for text independent speaker recognition. *Electronics Letters*, 33(24):2014–2016, 11 1997.

[2] Pierre Comon. Independent component analysis, A new concept? *Signal Processing*, 36:287–314, 1994.

[3] Aapo Hyvärinen. A family of fixed-point algorithms for independent component analysis. In *Proceedings of ICASSP*, pages 3917–3920, 1997.

[4] Aapo Hyvärinen. Independent component analysis by minimization of mutual information. Technical Report A46, Helsinki University of Technology, 1997.

[5] Ji-Hwan Kim, Gil-Jin Jang, Seong-Jin Yun, and Yung-Hwan Oh. Candidate selection based on significance testing and its use in normalisation and scoring. In *Proceedings of ICSLP*, pages 141–144, 1998.

[6] Richard J. Mammone, Xiaoyu Zhang, and Ravi P. Ramachandran. Robust speaker recognition: a feature-based approach. *IEEE signal processing magazaine*, pages 58–71, 9 1996.

[7] Danny R. Moates and Z.S. Bond. Same talker, different language. In *Proceedings of ICSLP*, pages 97–100, 1998.

[8] Yun Seong-Jin. Performance improvement of speaker recognition system for small training data. Master thesis, KAIST, in Korean, 1994.