Available at
www.ElsevierComputerScience.com
POWERED BY SCIENCE DIRECT®

Signal Processing 84 (2004) 1005–1019

SIGNAL PROCESSING

www.elsevier.com/locate/sigpro

Phoneme recognition using ICA-based feature extraction and transformation

Oh-Wook Kwon[a,*,1], Te-Won Lee[b]

[a] *School of Electrical and Computer Engineering, Chungbuk National University, 48 Gaesin-dong, Heungdeok-gu, Cheongju, Chungbuk 361-763, South Korea*

[b] *Institute for Neural Computation, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0523, USA*

Received 3 August 2003; received in revised form 26 February 2004

Abstract

We investigate the use of independent component analysis (ICA) for speech feature extraction in speech recognition systems. Although initial research suggested that learning basis functions by ICA for encoding the speech signal in an efficient manner improved recognition accuracy, we observe that this may be true for a recognition tasks with little training data. However, when compared in a large training database to standard speech recognition features such as the mel frequency cepstral coefficients (MFCCs), the ICA-adapted basis functions perform poorly. This is mainly due to the resulting phase sensitivity of the learned speech basis functions and their time shift variance property. In contrast to image processing, phase information is not essential for speech recognition. We therefore propose a new scheme that shows how the phase sensitivity can be removed by using an analytical description of the ICA-adapted basis functions via the Hilbert transform. Furthermore, since the basis functions are not shift invariant, we extend the method to include a frequency-based ICA stage that removes redundant time shift information. The performance of the new feature is evaluated for phoneme recognition using the TIMIT speech database and compared with the standard MFCC feature. The phoneme recognition results show promising accuracy, which is comparable to the well-optimized MFCC features.

Keywords: Speech recognition; Independent component analysis; Feature extraction

1. Introduction

Finding an efficient data representation has been a key focus for pattern recognition tasks. Popular methods for capturing the structure of data has been principal component analysis (PCA), which yields a compact representation, and more recently independent component analysis (ICA). In ICA, the data are linearly transformed such that the resulting coefficients are statistically as independent as possible. In a graphical model framework, the ICA can be regarded as data generative model in which independent source signals activate basis functions that describe the observation. The adaptation of these basis functions using ICA has received attention since this adaptation leads to a highly efficient representation of the data. Efficiency is measured in terms of its coding lengths (bits) per unit. Fewer bits corresponds to a lower entropy of the transformed data. Examples in representing

* Corresponding author.
E-mail address: owkwon@chungbuk.ac.kr (O.-W. Kwon).
[1] This work was mostly done while with INC, UCSD.

0165-1684/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.sigpro.2004.03.004

natural scenes include [4,20]. For audio signals, Bell and Sejnowski [3] proposed ICA to learn features for certain audio signals. Speech basis functions were also learned for speech recognition tasks [17].

Feature extraction for speech recognition aims at an efficient representation of spectral and temporal information of non-stationary speech signals. Conventionally speech signals were transformed to the frequency domain by using the Fourier transform and then the spectral coefficients are transformed by the discrete cosine transform (DCT) to the cepstral domain to remove the correlation between adjacent coefficients. The DCT reduces the feature dimension and produces nearly uncorrelated coefficients, which is desirable when back-end speech recognizers are based on continuous hidden Markov models (HMMs) using Gaussian mixture observation densities with diagonal covariance matrices. The resulting mel frequency cepstral coefficients (MFCCs) are one of the most common base features to represent spectral characteristics of speech signals.

Among many on-going research to challenge the MFCC feature, we note two techniques: perceptually linear predicted (PLP) cepstral coefficients and ICA [11,15]. The PLP-based cepstral coefficients were devised to directly reflect the human perceptual characteristics such as loudness and frequency sensitivity [9]. On the contrary, the ICA-based feature extraction is data driven and attempts to find a linear transformation such that the resulting coefficients are as independent as possible.

To reduce the temporal correlation, conventionally delta and acceleration components, equivalent to the second-order regression coefficients, were appended to the base features in the standard HMM-based speech recognizers [27]. The procedure to compute the added coefficients corresponds to a finite-impulse response (FIR) filtering in the temporal direction assuming independence in the spectral direction. Recently, research efforts have been made to replace the FIR filtering by a more efficient feature transformation. Usually, a segment of multiple static feature frames with overlap is regarded as a two dimensional image patch, and then spectro-temporal redundancy is reduced by using orthogonal transforms or linear discriminant analysis (LDA) [14]. The orthogonal transforms include PCA [7], two-dimensional DCT [28], discrete wavelet transform (DWT) [8]. Recently there have emerged a few research works on applying ICA for the same purpose. The MFCCs or mel filter bank coefficients were used as the input signals of ICA transformation [13,23,25].

Prior research on using ICA features for speech recognition resulted in significant improvements [18] but experiments were conducted under constraint settings (small training data). Our goal was to investigate this approach without any constraint setting and provide new analysis and options to cope with the main problems of the standard ICA features, namely in providing features that are phase insensitive and time-shift invariant.

In this paper, we apply ICA to speech signals in order to analyze its intrinsic characteristics and to obtain a new set of features for automatic speech recognition tasks. Although we would like to ideally obtain features in a complete unsupervised manner since the ICA is a data-driven method, however, we are faced with certain problems that need to be addressed for applying ICA features to the speech recognition task. First, the ICA filters (row vectors of the ICA unmixing matrix) are sensitive to phase change of input signals and produce different coefficients with different phase. The fact does not match the human perception mechanism of phase insensitivity [21]. In fact, speech contents can be recognized from zero-phased speech signals while speech signals with magnitude uniform and phase unchanged sound like noise signals. The ICA filters are also sensitive to shift (location) of speech signals in a window especially in the high-frequency band because the corresponding basis functions are localized in the temporal direction. Another problem is that ICA does not consider the human perception characteristics, high sensitivity to low-frequency band and logarithmic perception to loudness [21]. Our goal is to analyze the results to derive a new set of features that makes use of the ICA derived features and copes with the phase and time shift invariance in speech recognition.

In Section 2, we describe the speech model assumed in the paper, explain the phase problem and propose the feature extraction and transformation method, which uses an ICA filter instead of the fast Fourier transform (FFT) and another ICA filter for the DCT and temporal filtering. In Section 3, we analyzed the effects of the window size and showed the potential advantage of ICA by analyzing the conditional

probability distribution of the final coefficients. In Section 4, phoneme recognition results are presented. In Section 5, we discuss several issues related with phase invariance, ICA in the power domain, and speech recognizers. Conclusions are presented in Section 6.

2. Feature extraction and transformation using ICA

2.1. Speech model

Recently, the concept of sparse coding and ICA has been successfully applied to image coding and natural signal representation. Sparse coding of natural images was shown to produce localized and oriented basis filters similar to the receptive fields of simple cells in the primary visual cortex [20]. In their study, an image patch was assumed to be generated by a linear combination of basis patches with their corresponding factor coefficients that were as sparse as possible. ICA was also used to elucidate the basis functions of natural images [4] and sound signals [3,19], assuming that the underlying causes have sparse or in general super-Gaussian distributions.

Along this line of research, we assume that speech signals are generated by a generative model where speech signals are represented as a linear combination of basis functions weighted by independent source coefficients. A frame of N observed speech samples is represented by a linear combination of N source signals as

$$\mathbf{x} = \mathbf{A}\mathbf{s}, \tag{1}$$

where $\mathbf{x}$ is an $N \times 1$ column vector of the speech samples, $\mathbf{A}$ is an $N \times N$ mixing matrix whose column vectors constitute a set of basis functions and $\mathbf{s}$ is an $N \times 1$ column vector of the source signals. In this work, we assume that the source signals follow a sparse distribution. This sparseness assumption is reasonable when trying to obtain basis functions that produce an efficient coding scheme. On the other hand, one can also adapt the source distribution using a parameterized model of the source density, such as the generalized Gaussian or exponential power density [5]. For representing speech signals however, this parameterized approach leads to source densities that have Laplacian or even sparser density models [12,19]. Both directions, namely a parameterized density model with an independence cost function and the Laplacian prior model yield similar basis functions and properties for speech signal representation. For simplicity, we assume the Laplacian source model to learn the basis functions. With the assumption of the Laplacian source density, we used the Infomax algorithm with the natural gradient [15] to learn all the basis functions in our proposed method. Fig. 1 illustrates the assumed speech model where a speech segment is decomposed into basis functions and the coefficients.

2.2. Phase information of speech signals

When processed in a short segment, speech signals are insensitive to phase variation, as opposed to the case of natural images [4]. To describe this phenomenon, we set up a small experiment. A segment of speech signals is transformed to frequency domain, the phase of the transformed coefficients was set to zero or the magnitude was set to unity with phase preserved. Speech signals of 8 kHz sampling rate were processed block-wise with the window size 20 ms and the shift size 10 ms. Fig. 2 shows the resulting waveforms and the corresponding spectrograms. From the top are the original speech signals, the zero-phased signals, and the unity magnitude signals. Below each waveform is magnified five frames of the waveform after 0.78 s. The uniform magnitude signal is totally different from the original signal, whereas the zero-phase reconstruction lets us recognize the speech contents with minor degradation in speech quality, monotonic tone and loss in speech details. The unity-magnitude signals sounded almost like white noise.

2.3. Proposed method

Phase sensitivity and time variance seemed to be the most profound factors prohibiting the use of the ICA-adapted basis functions for speech recognition tasks. Ideally, we would like the algorithm to learn phase insensitive and time shift invariant filters. However, there exist no algorithm that can handle these invariance and it is still subject to research directions. Instead of a new algorithm, we propose additional steps to cope with the invariance problems. We alleviated the problem of phase sensitivity by using the

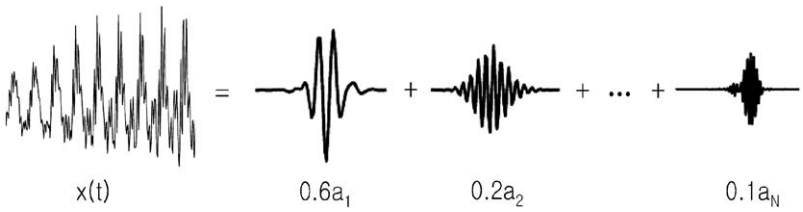

Fig. 1. A speech segment **x** is generated by or decomposed into basis functions and its corresponding coefficients.

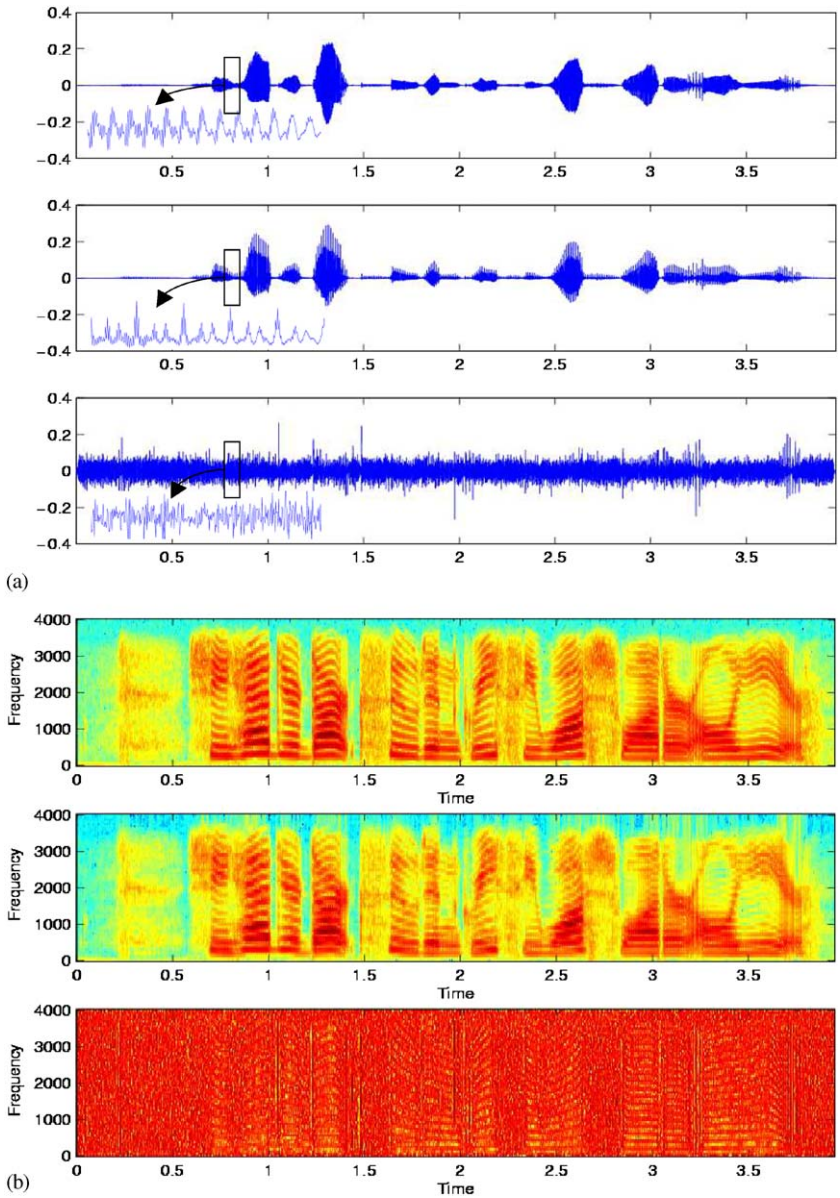

Fig. 2. Original speech signals, zero-phase signals, and uniform magnitude signals are displayed in the time domain (a), in the log spectral domain (b). The magnified waveforms of five frames after 0.78 s show differences in detail. Each sub-figure consists of the original speech signals (top), zero-phased signals (middle) and uniform magnitude signals with phase preserved (bottom).

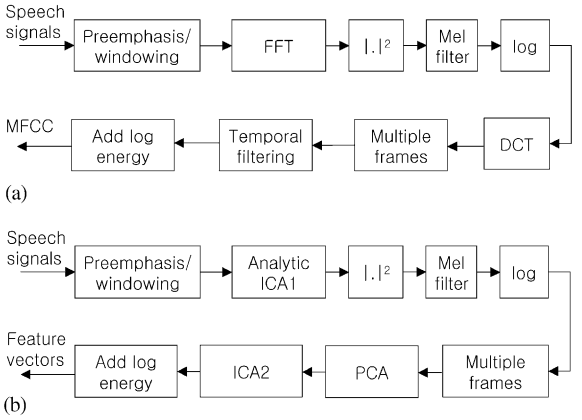

Fig. 3. Block diagrams of feature extraction using MFCC (a) and ICA (b). In ICA-based feature method, speech signals are filtered by analytic ICA filters, the coefficients are taken log of squared magnitude split into mel frequency bands, multiple frames are concatenated, and another ICA in the spectro-temporal domain are performed to produce the final feature vector.

analytic ICA filters obtained from the real ICA filters via the Hilbert transform and taking the magnitude of complex ICA coefficients. We mitigated the shift sensitivity problem by using a mel filter and summing squared magnitude assigned to the same mel band. However, the resulting coefficients have similar characteristics to the standard mel filter bank coefficients except non-uniform center frequencies and non-uniform filter weights. Considering psychoacoustics of speech signals [21], we took log of the obtained coefficients. The coefficients show large correlation because ICA was not learned to optimize the independence of magnitude coefficients. Therefore, we apply an additional ICA transformation for the log spectral coefficients to obtain as independent coefficients as possible. The mel filter and log operation used in the conventional feature extraction were applied to the ICA coefficients in order to reflect the human speech perception characteristics. Fig. 3 compares feature extraction methods using MFCC and ICA. The ICA in the time domain (ICA1) in the proposed method replaces the FFT of the MFCC-based method, and the PCA and ICA in the spectro-temporal domain (ICA2) plays the role of the DCT and temporal filtering. The mel filtering in the proposed method is different from that used in the MFCC-based method as the ICA filters have different center frequencies.

2.3.1. Preemphasis and windowing

Speech signals are preemphasized by using a first-order FIR filter and preemphasis plays a role in weakening the correlation of speech signals. A stream of speech signals is segmented into a series of frames with N samples and each frame is windowed by a Hamming window. These two steps are the standard procedure in feature extraction for speech recognition. In the following sections, we omit the frame index t unless confused, assuming that all processing is done in the frame base.

2.3.2. Analytic ICA in the time domain (ICA1)

We used the Infomax algorithm with the natural gradient extension [1,2] to obtain the basis functions and the corresponding coefficients as described in Section 2.1. To accelerate the convergence, we reduced the dimension of the windowed signals $\mathbf{x}$ and obtained a sphered signal $\mathbf{z}$ by multiplying the sphering matrix $\mathbf{V}_1$ obtained by eigenvector decomposition of the covariance matrix [11]

$$\mathbf{z} = \mathbf{V}_1\mathbf{x}, \tag{2}$$

where $\mathbf{V}_1$ is an $M \times N$ matrix and M is the reduced dimension of the input signals. The updated unmixing matrix $\mathbf{W}_1$ was constrained to an orthonormal matrix [11].

In the recognition mode, we set the mean of the row vectors of the unmixing matrix $\mathbf{B} = \mathbf{W}_1\mathbf{V}_1$ to zero to remove direct current (DC) components bearing no information. To reduce the phase sensitivity, we used the analytic version of the unmixing matrix, which was obtained via the Hilbert transform

$$\mathscr{B} = \mathbf{B} + \mathrm{j}\hat{\mathbf{B}}, \tag{3}$$

where $\hat{\mathbf{B}}$ is the Hilbert transform of $\mathbf{B}$ in the row direction and $j = \sqrt{-1}$. By using the analytic version of the unmixing matrix, we can obtain a smoother estimate of the ith coefficient magnitude $m(i)$, the energy of the windowed signal $\mathbf{x}$,

$$m(i) = \|\mathscr{B}_i\mathbf{x}\|^2, \quad i = 1, \ldots, M, \tag{4}$$

where $\mathscr{B}_i$ is the ith row vector of the analytic unmixing matrix. Using the analytic version is justified in Appendix A.

The difference from using the conventional FFT is that the ICA here uses the filters learned from speech signals having non-uniform center frequencies and

non-uniform filter weights but the FFT does not consider the fact that input signals are speech. Phase sensitivity is a common problem when localized basis functions are used to transform speech signals; energy components were used instead of time samples of filter outputs when the DWT was used for feature extraction [8]. We discuss the phase sensitivity issue further in Section 5.

2.3.3. Mel filter

Mel band energies were obtained by weighting the magnitude coefficients considering the center frequency of the mel bands [6] and the center frequency of the ICA filters. This procedure can be formulated as follows

$$fb(i) = \mathbf{F}(i)\mathbf{m}, \quad i = 1, \ldots, K, \tag{5}$$

where $\mathbf{F}(i)$ is the ith row vector denoting a mel filter whose center frequency is spaced in the mel scale and whose coefficients are weighted according to a triangular shape [6], and K is the number of the mel bands. Because the center frequencies of the ICA filters are different from those of the mel filters, we set the weight for the jth magnitude of the ith filter to

$$F(i,j) = \frac{|f_{\text{ica}}(j) - f_{\text{mel}}(i))|}{f_{\text{mel}}(i+1) - f_{\text{mel}}(i-1)}, \quad f_{\text{mel}}(i+1) \leqslant f_{\text{ica}}(j) \leqslant f_{\text{mel}}(i-1) \tag{6}$$

$$= \varepsilon \quad \text{otherwise}, \tag{7}$$

where $f_{\text{ica}}(j)$ is the center frequency of the jth ICA filters, $f_{\text{mel}}(i)$ is the center frequency of the ith mel band, and ε is a small constant to prevent underflow in taking logarithm in the next step. The $f_{\text{mel}}(0)$ and $f_{\text{mel}}(M+1)$ are assumed to be 0 and half of the sampling frequency. The center frequency of the ICA filters was determined by weighted frequency. The weight was proportional to the energy of each frequency components excluding the DC components. The weight of the DC component was set to zero. Therefore, we obtain a center frequency even if the basis functions have two lobes. In the recognition mode, the mel filtering is done in the time domain because the center frequencies of the basis functions are analyzed in the training stage.

This step mitigates the problem of time shift variance of the basis functions by summing several coefficients assigned in the same frequency band. Investigating the spectro-temporal characteristics of the basis functions of ICA in the time domain, which will be shown in Section 3.2, the peaks of adjacent basis functions are usually located in different non-overlapping temporal positions for middle-to-high frequency bands. As for the low bands, basis functions have large time span and have less serious problems. That is the same as with the ICA filters. Therefore by summing coefficients we can get a coefficient less sensitive to time shift. That is useful especially in the high frequency band where the corresponding basis functions have narrow temporal span.

The logarithm of the resulting coefficients were taken from the fact that human auditory system is sensitive to speech loudness in the logarithmic scale

$$\mathbf{g} = \log \mathbf{fb}. \tag{8}$$

The output vector $\mathbf{g}$ is used as a base for the following feature transformation stage to remove temporal dependencies between frames and obtain components that are as independent as possible by using another ICA step. As will be shown later, $\mathbf{g}$ has strong conditional dependency between components and thus needs further processing.

2.3.4. Concatenating multiple frames

We concatenate $2\Delta + 1$ consecutive frames to form a new vector at time t, $\mathbf{h}(t)$,

$$\mathbf{h}(t) = \begin{bmatrix} \mathbf{g}(t-\Delta) \\ \cdots \\ \mathbf{g}(t) \\ \cdots \\ \mathbf{g}(t+\Delta) \end{bmatrix}. \tag{9}$$

Because the DC component of $\mathbf{h}$ does not have a sparse distribution, we subtracted the local mean of $\mathbf{h}$ [11]. If Δ is equal to 0, no temporal filtering is done and only spectral transformation takes place as the DCT in the MFCC computation.

Conventionally, a fixed temporal regression filter was used to extract velocity and acceleration components assuming that each frequency component is

independent of other frequency components in the different frame index [27]. We increase independence of the coefficients in the spectral and temporal direction by using concatenated multiple frames for feature transformation.

2.3.5. PCA

Before applying the second ICA, we performed PCA to reduce dimension first to the number of the target coefficients L, following the procedure described in Section 2.3.2. We attempted to reduce the dimension to a larger dimension than L and select L basis functions, which turned out to yield worse accuracy. Therefore we directly reduced the dimension to L before ICA in the spectral domain.

$$\mathbf{p} = \mathbf{V}_2\mathbf{h}, \tag{10}$$

where $\mathbf{V}_2$ is the sphering matrix in the spectral domain.

2.3.6. ICA in the spectro-temporal domain (ICA2)

The basis functions of ICA in the spectro-temporal domain is learned in the same way as those of ICA in the time domain except that the input vectors are subtracted by local mean. The coefficient vector $\mathbf{q}$ is obtained as

$$\mathbf{q} = \mathbf{W}_2\mathbf{p}, \tag{11}$$

where $\mathbf{W}_2$ is the learned unmixing matrix. The resulting coefficients can be used for recognition purpose. In this case, phase sensitivity indicates spectral change or presence of phoneme boundaries. PCA is optimal to decorrelate signals generated by a single Gaussian density. But speech recognizers use Gaussian mixture models and hence an independence constraint is more desirable as in this case.

2.3.7. Adding log energy

The log energy was appended as a component of the final feature vector:

$$\mathbf{r} = \begin{bmatrix} \mathbf{q} \\ E(t) \end{bmatrix}. \tag{12}$$

The $E(t)$ is computed as the mean of the $\mathbf{h}$:

$$E(t) = \sum_{i=1}^{(2\Delta+1)K} h(t,i), \tag{13}$$

where $h(t,i)$ is the ith output coefficient at frame t.

3. Analysis of ICA basis functions

3.1. Speech database

Using the TIMIT speech database, we analyzed the basis functions of ICA and evaluated the performance of the ICA-based feature in a speaker-independent phoneme recognition task. The database was recorded in sound-proof room and is widely used as a reference for comparison of speech recognition performance. The sampling rate of the database was down-converted to 8 kHz to reduce the training time for ICA. We conjecture that the tendency and phenomena observed in this work are also applicable to the case of 16 kHz sampling rate without degradation. To train the ICA filter, we used 1 h of speech data from the training set, which corresponds to DR1 (New England dialect) and DR2 (Northern dialect). For evaluation, we used the core test set which includes 192 sentences by 16 male and eight female speakers. The amount of the training data is sufficient to provide a reliable estimate of the ICA unmixing matrices.

3.2. Basis functions of ICA in the time domain

Fig. 4 shows the basis functions sorted by the L2 norm and the corresponding frequency responses when the frame size is 10 ms and the number of sources is 80. Most of the basis functions show a wavelet function-like waveform. To obtain these basis functions, we updated the ICA filter matrices every 1000 frames, with the convergence factor $\eta(t)$ linearly decreasing from 0.0001 to 0.000001. The mixing matrix $\mathbf{A}$ was computed by inverting the unmixing matrix $\mathbf{B}$. In general, one cannot order the basis functions due to the scale indeterminacy [11]. But in this case, it was possible to sort the basis functions because we sphered the input signal and constrained the ICA filters for the sphered signal to be orthonormal in learning the filters.

Fig. 5 shows the Wigner–Ville distribution (WVD) of the basis functions [18] when the window size is 10 and 20 ms, respectively. Each contour line represents the locus of half of the maximum amplitude and each cross denotes the location where the maximum amplitude occurs. The use of the WVD caused narrower time span in the low-frequency region than the exact time span obtained from visual inspection of the

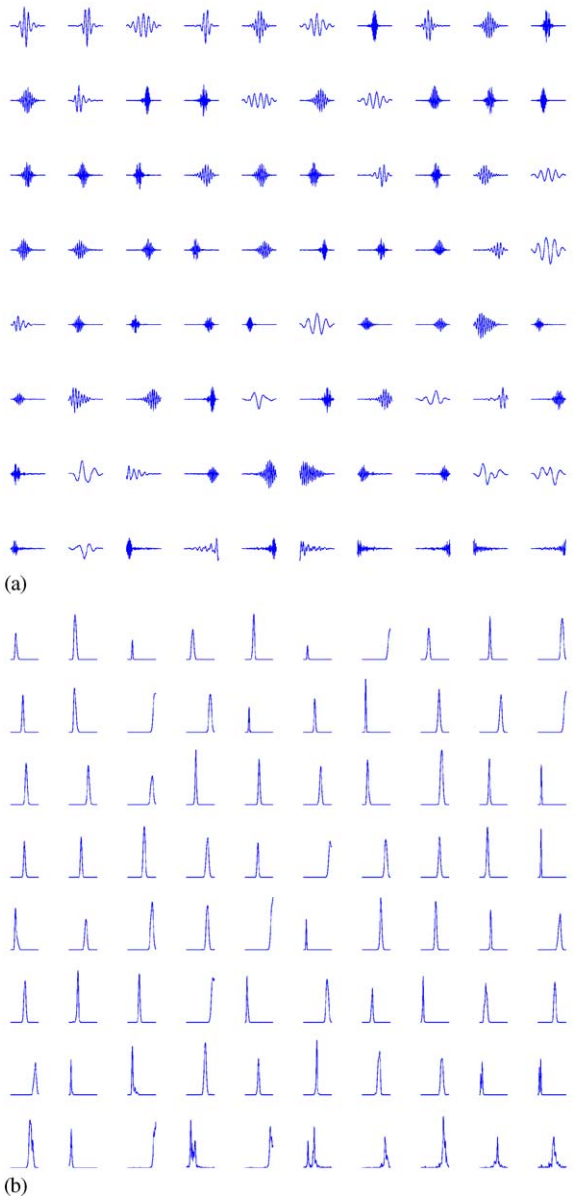

Fig. 4. Basis functions (a) and their frequency response (b) of ICA in the time domain. To learn the basis functions, we used the 10 ms window size and assumed 80 source signals. The basis functions were sorted in the non-increasing order by the L2 norm and most of them show wavelet function-like characteristics.

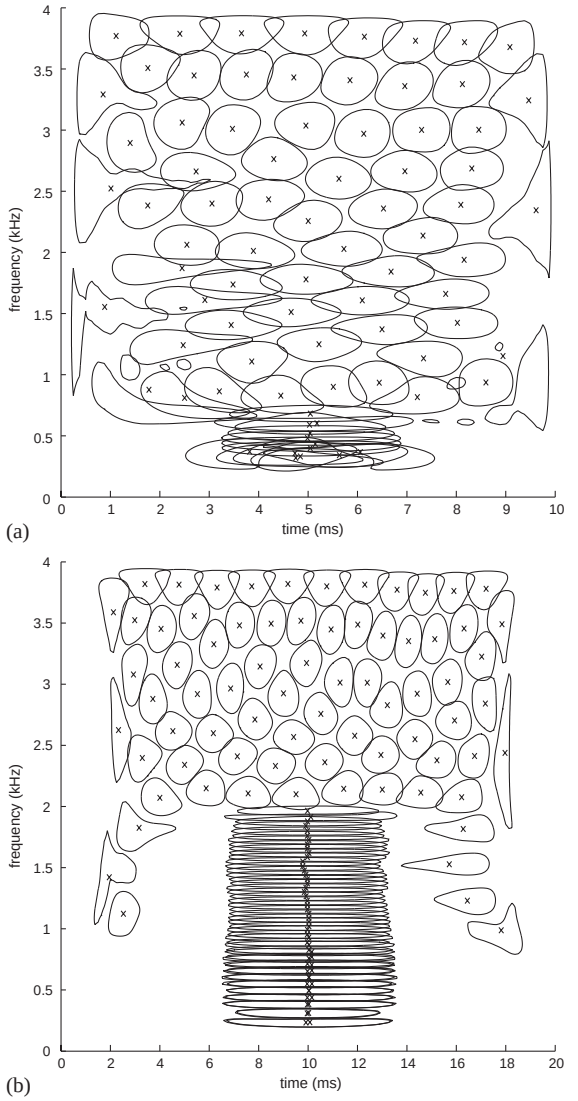

Fig. 5. WVD of the basis functions with different window sizes of 10 ms (a) and 20 ms (b). In case of the 20 ms window size, we assumed 128 source signals and performed sphering by eigenvalue decomposition before learning ICA. Each contour line denotes the locus of half of the maximum amplitude and each cross denotes the location of the maximum amplitude.

basis functions. The figures indicate that the basis functions in the middle-to-high frequency band are localized both in the frequency and temporal directions.

The first 20 basis functions covered the 0–4 kHz frequency range where most of speech signal energy is distributed. The narrow width of the basis functions in the temporal direction implies the shift variant property of the basis functions. In case of 20 ms window

size, the temporal range of the basis functions in the high-frequency region was similar to that of 10 ms window size. In both cases, the basis functions in the low-frequency region spanned the whole frame length. In case of 20 ms window size, the basis functions resembled sinusoidal waveforms and more basis functions covered the whole time span, which is similar to the animal vocalization case in [19]. This fact implies that more basis functions cover whole time span as the window size increases. Note that the WVD figures in this work are a little different in the low-frequency region from [18] because input signals were preemphasized.

3.3. Basis functions of ICA in the spectro-temporal domain

To illustrate the basis functions of the ICA in the spectro-temporal domain, we trained the ICA filters by using the concatenated log energy coefficients, which are more Gaussian-like than power signals. In terms of signal processing, it may be reasonable to use power signals as in [22]. However in view of phoneme recognition, we argue that log energy is a plausible quantity considering logarithmic loudness sensitivity curve of the human auditory system [9]. Recent study results also shows the validity of ICA in the spectral domain [23,24].

The parameters used for learning were the same as those for the ICA1. The window size of the ICA1 was 25 ms and the number of the source signals was 128. We decided to produce the same number of the final coefficients to compare with the standard MFCC feature, which made us use 38 basis functions. We tried to make the number of coefficients at the final output equal to the conventional MFCC feature with the delta and acceleration components (39). This 39-dimensional feature is called the MFCC_E_D_A feature hereafter. Because the more coefficients produce the higher accuracy, the equal number of coefficients is required. Therefore, adding log energy to the 38 coefficients, we could get the same number of coefficients. The 38 coefficients correspond to the highest 38 eigenvalues. We also kept the number of coefficients used in temporal filtering the same as in the standard temporal filtering case, which implies nine frames of the 23-dimensional coefficient vectors. We used the same ICA algorithm to learn the unmixing matrix. After learning the unmixing matrix, the average kurtosis of the resulting coefficients was 7.5, which is smaller than 55 of the coefficients in the time domain (ICA1) due to the logarithmic scaling of energy coefficients.

Fig. 6 shows the learned 38 basis functions. A basis function was displayed as an image patch to reveal its spectro-temporal structure. The image patch size was determined from the fact that the equivalent frame width ($2^*\Delta + 1$) is 9 in the MFCC_E_D_A case and the number of mel bands is 23. The number of mel bands is optimized from the human perception study. The horizontal axis denotes temporal index and the vertical axis denotes the index to the mel-filtered log coefficients. The gray level of the image patches denotes the value of the basis functions normalized to 0 (black) to 255 (white) between the minimum and maximum values of the basis functions. The first five basis functions represent the temporal changes at phoneme boundaries and the horizontal stripes represent the spectral distribution of speech signals. We also observe that some basis functions (e.g., the 21st, 22nd, 24th, 25th ones) are localized in the spectral direction.

We computed the conditional probability distribution to check the independence of output signals in each step of the proposed method. Fig. 7 shows the conditional probability distributions of signals in each stage of the brief block diagram above. Among many possible pairs of conditional distributions, we illustrated only one sample conditional distribution of the fifth coefficient given the first coefficient. The figures show that the output coefficients after the ICA2 show more independence than the MFCC (the third figure) or PCA coefficients (the fourth figure). In particular, the ICA2 output was shown to have smallest dependency of all cases within one standard deviation range around mean. In contrast with other methods to remove dependency in the magnitude coefficients of the first-stage ICA, by using ICA in the spectro-temporal domain, we do not need computation-intensive maximization steps as in divisive normalization [24].

To obtain a quantitative measure of independence, we computed mutual information among the coefficients. The lower mutual information is, the less dependency exists. Table 1 shows mutual information averaged over all coefficient pairs. The results showed that the output coefficients from the ICA2 showed the

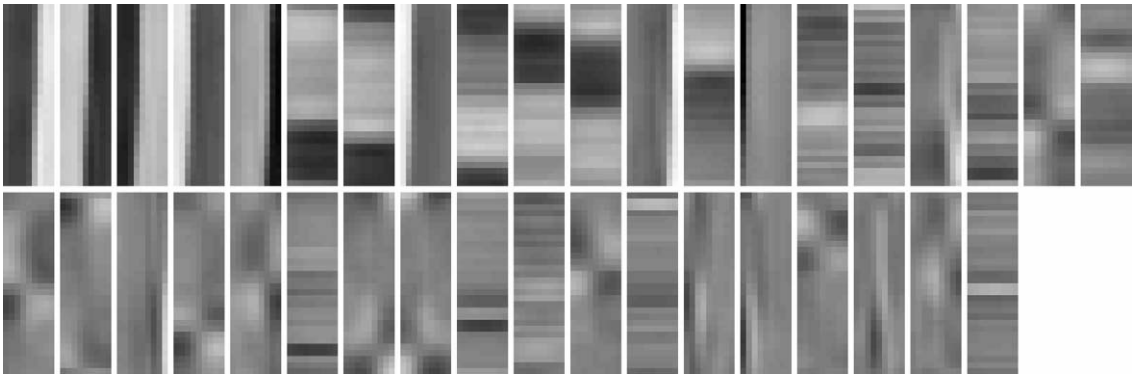

Fig. 6. Basis functions of ICA in the spectro-temporal domain were displayed in a image patch to reveal their spectro-temporal structure. The horizontal axis denotes the temporal index and the vertical axis denotes the index of the mel-filtered log coefficients. The gray level of the image patches denotes the value of the basis functions normalized to 0 (black) to 255 (white) between the minimum and maximum values of the basis functions. The height of the image patches were decided to match with the number of the mel bands.

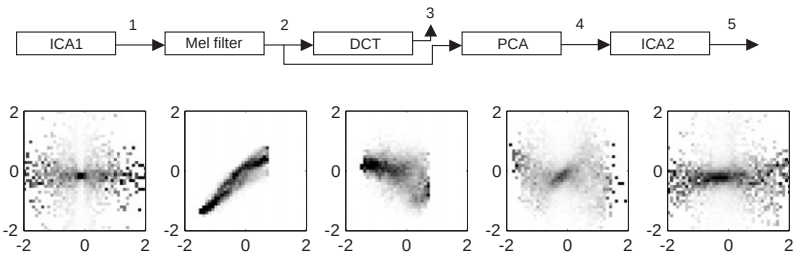

Fig. 7. Conditional probability distributions of output coefficients of ICA1, mel filter bank coefficients, coefficients transformed by DCT, coefficients transformed by PCA, and coefficients by ICA. The block diagram (top) illustrates what signals are used for each distribution (bottom). Each feature dimension was normalized to have zero mean and unity variance.

Table 1
Comparison of mutual information averaged over all coefficient pairs

ICA1	Mel	DCT	PCA	ICA2
0.234	1.233	0.316	0.229	0.220

least average mutual information, which implies the minimum dependency among the five cases. Although most of mutual information reduction are achieved by PCA, the 4% reduction through ICA2 shows that there exists higher-order dependency which cannot be removed by an orthogonal transform optimized by using the second-order statistics only. Increase of mutual information for mel-filtered coefficients explains the fact that the ICA1 coefficients have small dependency but the magnitude of the coefficients have large dependency.

4. Phoneme recognition results

For comparison purposes, the hidden Markov toolkit (HTK) was used to extract the standard MFCC feature with 23 frequency bands [27]. The window size was 25 ms and each frame was shifted by 10 ms. The speech signals were preemphasized by an FIR filter with a factor of 0.97. We did not use any language model to evaluate only the performance of feature extraction. We used 48 context-independent phoneme models including silence. Every phoneme model except short pause was modeled by a three-state left-to-right HMM and short pause was modeled by a

single-state HMM. We used continuous HMM with observation probability distribution of four Gaussian mixtures for each state. Phoneme accuracy was converted to a value corresponding to the standard 39 phoneme set [16].

We first evaluated accuracy of the baseline system with the standard features, which was used as the reference for the subsequent experiments. With the standard 23-dimensional filter bank feature, we obtained the phoneme accuracy of 32.18%. The DCT applied to the log energy coefficients of the filter bank feature yielded 36.45% of phoneme accuracy. The DCT improved accuracy by decorrelating the filter bank coefficients. We obtained 50.87% accuracy by appending the delta and acceleration components (MFCC_E_D_A) with $\Delta = 2$. The delta and acceleration components were obtained only through temporal filtering from the assumption that two streams of cepstral coefficients with different index are independent. We note that the feature MFCC_E_D_A has been well refined with regard to the mel frequency mapping, the shape of band pass filters, and the number of mel filters. For the same context-independent phoneme recognition task, Young reported the phoneme accuracy of 52.7% using the same HTK package and the same feature but the higher sampling rate of 16 kHz [26].

Next, we evaluated the performance of the time-domain ICA feature only. We selected 23 coefficients according to the L2 norm of the basis functions and took log (ICA-NORM-FBANK), and applied the DCT to the log energy coefficients and selected the first 13 coefficients (ICA-NORM-CEP), which uses the equal number of parameters to the standard MFCC features. Experimental results showed that the 23-dimensional ICA-NORM-FBANK feature and the 13-dimensional ICA-NORM-CEP feature produced 12.77% and 13.83% accuracy, respectively. The results implies that we cannot achieve good recognition accuracy by using coefficients from the selected basis functions even though the DCT improves a little by giving more decorrelated coefficients.

We evaluated the performance of using ICA filters both in the time domain and spectro-temporal domain, with the window size and the number of sources are equal to 80 and Δ is 0. We computed the center frequency of the ICA1 filters and applied mel filtering to get 23 coefficients. Then we applied the ICA2 filters to reduce the dimension and obtained 13 coefficients including the log energy. Phoneme accuracy with the 23-dimensional feature after the mel-filtering was 25.90%, with the 13-dimensional feature after PCA, 27.91%, and with the 13-dimensional feature after ICA2, 28.48%. These results are inferior to the ICA-based feature of the previous study [17] because our phoneme recognition task include more speech variability than the small isolated word database which was used in the previous study.

We further tested the proposed ICA-based feature with different window sizes and different numbers of source signals. We set Δ to 4 so that the number of coefficients used for temporal filtering is equivalent to the standard feature case (MFCC_E_D_A). We found that the window size 160 (20 ms at the 8 kHz sampling rate) achieved the best accuracy as shown in Table 2. The ICA-based feature with multiple frames yielded recognition results comparable with the MFCC_E_D_A feature.

Further study is needed to obtain better representation of time domain signals in the ICA1. Regarding the ICA2, it successfully decorrelates the mel filter bank coefficients. The performance difference between PCA and ICA in decorrelation the multiple frame coefficients is more remarkable when a small number of mixture Gaussian densities are used in phoneme recognizers. Our results are different from [23] and [25] in that our method uses the mel-filtered band energies from the time domain ICA as the base feature while they use the MFCC instead. When the number of mixtures increases, the performance difference is narrowed as pointed out in [25].

5. Discussion

5.1. Sparse coding and speech recognition

Although one would expect better speech recognition accuracy due to improved encoding of the speech signal, we were not able to meet the expectation mostly due to the nature of current speech recognition systems that are optimized for features that produce a Gaussian density fit. ICA for speech signal representation assumes a super-Gaussian density model for source signals, which is different from a Gaussian density used as the observation density in common HMM-based

Table 2
Phoneme accuracy (%) with different window sizes and numbers of source signals when $\Delta = 4$

Window size (N)	Number of sources (M)	ICA-FBANK (23)	ICA-PCA (39)	ICA-ICA (39)
80	80	25.90	45.97	48.18
128	128	27.57	47.14	48.23
160	128	30.52	50.44	50.89
200	128	29.88	49.70	50.78

speech recognizers. There is still some mismatch between some known properties of speech perception and recognition and the adopted model in this paper. Further investigation at this level is desirable.

To exploit the independence and the super-Gaussian nature of ICA coefficients, a different density model is desirable for speech recognition. But in this case, we lose the advantage of easy parameter estimation in Gaussian mixture modeling (GMM). Therefore, we need to either adapt the ICA outcome to statistics that are accurately modeled by the GMM and the HMM, or we need to change the mixture modeling and HMM, respectively, to account for the changes due to the new representation of features. This process will increase recognition accuracy and therefore optimize the system toward the use of the ICA model.

5.2. Phase and shift invariant ICA

The biggest disadvantage of using ICA for speech feature extraction in the time domain is that the learned basis functions itself are phase sensitive and shift variant. This is not well matched with the fact that speech recognizers proved to be successful with just magnitude information in a certain degree. To implement phase and shift invariance, one may use shift the basis functions and sum the resulting coefficients. That procedure has the same effect as using Fourier transform-like basis functions with non-uniform weight and frequency. Another approach to tackle phase sensitivity of the basis functions in the conventional ICA may be to use nonlinear algorithms such as subspace and topographic ICA [10].

5.3. Use of other ICA algorithms

Although we chose the Infomax algorithm to learn the basis functions, we believe that the results here are generic and are not dependent on use of other ICA algorithms such as the joint approximate diagonalization of eigenmatrices (JADE) algorithm and the Fast ICA algorithm [11]. The topographic ICA algorithm has a different optimization measure and may be useful in speech feature extraction by virtue of its phase invariance property. However, there are many other methods that can be explored for this purpose.

5.4. ICA in the power domain

A few researchers studied feature transformation in the spectral power domain, specifically using the squared magnitude of frequency band energy instead of log spectral energy [22]. The approach was based on the fact that power components of two different orthogonal signals can be linearly combined in the power domain. We also found that ICA in the power domain produce larger kurtosis in output coefficients and more localized basis functions. However, in our preliminary experiment, we could not achieve similar accuracy to the conventional MFCC. Therefore, we decided not to pursue that track any further. For speech recognition purpose, a signal processing-based approach should be combined with information from the human perception characteristics.

6. Conclusions

We investigated the effectiveness of ICA-based feature extraction in the time domain and feature transformation in the spectro-temporal domain using the ICA learning algorithm. To alleviate the phase shift sensitivity and time shift problems in using the conventional ICA coefficients, the analytic version of ICA filters was used and the outputs from adjacent mel bands were summed. We applied a mel filter to obtain

spectral information and took logarithm of the coefficients, which matches well to the human auditory system. It was found that applying nonlinear operations such as the log operation and converting ICA coefficients by using an additional transform was an effective method to reduce the dependencies among the source coefficients. We analyzed the time-frequency characteristics of the learned basis functions with different window sizes, and the ICA filters and the corresponding frequency response. In analysis using a large database, the proposed methods performed similarly to the MFCC-based feature speech recognition system. Note that the MFCC features have been optimized for several decades of speech recognition research. Our results are encouraging and further research is needed in finding appropriate nonlinear transformations to accommodate the human perceptual mechanisms in the spectral domain.

Appendix A.

A.1. Phase sensitivity of ICA filters

Assume that the input signal to a band-pass filter is represented by band-pass signals

$$x(t) = a(t)\cos(\omega t + \phi) \tag{A.1}$$

and a band-pass filter is represented by

$$B(t) = b(t)\cos(\omega t), \tag{A.2}$$

where $a(t)$ and $b(t)$ are tapered to zero in the positive and negative time axes and slowly varying compared to the carrier frequency ω, and ϕ is the phase offset (shift) between $x(t)$ and $B(t)$. Then the output signals of the filter becomes

$$y(t) = \sum_{\tau=1}^{M} x(t-\tau)B(\tau) \tag{A.3}$$

$$= \sum_{\tau=1}^{M} a(t-\tau)b(\tau)\cos(\omega(t-\tau)+\phi) \times \cos(\omega\tau) \tag{A.4}$$

$$= \sum_{\tau=1}^{M} a(t-\tau)b(\tau)\left[\cos(\omega t+\phi)\frac{1+\cos(2\omega\tau)}{2} - \sin(\omega t+\phi)\frac{\sin(2\omega\tau)}{2}\right]. \tag{A.5}$$

Then we obtain an approximation of the real ICA filter as

$$y(t) \approx \frac{1}{2}\sum_{\tau=1}^{M} a(t-\tau)b(\tau)\cos(\omega t+\phi) \tag{A.6}$$

and the magnitude as

$$|y(t)| \approx \frac{1}{2}\left|\sum_{\tau=1}^{M} a(t-\tau)b(\tau)\cos(\omega t+\phi)\right|. \tag{A.7}$$

Here we used the trigonometric identities and the following assumption:

$$\sum_{\tau=1}^{M} a(t-\tau)b(\tau)\cos(2\omega\tau) \approx 0, \tag{A.8}$$

$$\sum_{\tau=1}^{M} a(t-\tau)b(\tau)\sin(2\omega\tau) \approx 0, \tag{A.9}$$

where the carrier frequency ω is much higher than the frequency band of the tapered signal $a(t)$, the frame size M is large enough so that the average of the carrier component becomes nearly zero. Hence, we get the resulting magnitude of the output coefficient is dependent on the phase offset ϕ.

When we use the analytic version of the ICA filter, the output signal becomes

$$y(t) = \sum_{\tau=1}^{M} x(t-\tau)\mathscr{B}(\tau) \tag{A.10}$$

$$= \sum_{\tau=1}^{M} a(t-\tau)b(\tau)\cos(\omega(t-\tau)+\phi)[\cos(\omega\tau) + \mathrm{j}\sin(\omega\tau)], \tag{A.11}$$

where $\mathscr{B}(t)$ is the analytic version of the ICA filter. Here we used the modulation property of the Hilbert transform regarding a band-pass signal [29]

$$H(m(t)\cos(\omega t)) = m(t)H(\cos(\omega t)) = m(t)\sin(\omega t), \tag{A.12}$$

where $H(\cdot)$ denotes the Hilbert transform, $m(t)$ is a non-overlapping low-pass signal and ω is the carrier frequency. Following similar steps to the previous case, we obtain the complex output coefficient

$$y(t) \approx \frac{1}{2}\sum_{\tau=1}^{M} a(t-\tau)b(\tau)[\cos(\omega t+\phi) - \mathrm{j}\sin(\omega t+\phi)] \quad \text{(A.13)}$$

and the corresponding magnitude

$$\|y(t)\| \approx \frac{1}{2}\left|\sum_{\tau=1}^{M} a(t-\tau)b(\tau)\right|, \quad \text{(A.14)}$$

which is mostly insensitive to the phase offset ϕ.

References

[1] S. Amari, Neural learning in structured parameter spaces —natural Riemannian gradient, in: Advances in Neural Information Processing System, Vol. 9, MIT Press, Cambridge, MA, 1997, pp. 127–133.

[2] A. Bell, T. Sejnowski, An information-maximization approach to blind separation and blind deconvolution, Neural Comput. 7 (1995) 1129–1159.

[3] A.J. Bell, T.J. Sejnowski, Learning the higher-order structure of a natural sound, Network Comput. Neural Syst. 7 (1996) 261–266.

[4] A.J. Bell, T.J. Sejnowski, The 'independent components' of natural scenes are edge filters, Vision Res. 37 (23) (1997) 3327–3338.

[5] G.E.P. Box, G.C. Tiao, Bayesian Inference in Statistical Analysis, Wiley, New York, 1992.

[6] ETSI Standard, Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms, ETSI ES 202 050 v1.1.1, October 2002.

[7] T. Fukuda, et al., Peripheral features for HMM-based speech recognition, in: Proceedings of the International Conference on Acoustics, Speech, Signal Processing, Salt Lake City, UT, 2001.

[8] R. Gemello, et al., Integration of fixed and multiple resolution analysis in a speech recognition system, in: Proceedings of the International Conference on Acoustics, Speech, Signal Processing, Salt Lake City, UT, 2001.

[9] H. Hermansky, et al., RASTA-PLP speech analysis technique, in: Proceedings of the International Conference Acoustics, Speech, Signal Processing, San Francisco, CA, March 1992, pp. 1121–1124.

[10] A. Hyvärinen, P.O. Hoyer, M. Inki, Topographic independent component analysis, Neural Comput. 13 (2001) 1527–1558.

[11] A. Hyvärinen, J. Karhunen, E. Oja, Independent Component Analysis, Wiley, New York, 2001.

[12] G.-J. Jang, T.-W. Lee, A probabilistic approach to single channel blind signal separation, in: Advances in Neural Information Processing Systems, 15, MIT Press, Cambridge, MA, 2003.

[13] G.-J. Jang, S.-J. Yun, Y.-H. Oh, Feature vector transformation using ICA and its application to speaker verification, in: Proceedings of the EUROSPEECH 99, Budapest, Hungary, September 1999, pp. 767–770.

[14] S. Kajarekar, et al., A study of two dimensional linear discriminants for ASR, in: Proceedings of the International Conference on Acoustics, Speech, Signal Processing Salt Lake City, UT, 2001.

[15] T.-W. Lee, Independent Component Analysis: Theory and Applications, Kluwer Academic Publishers, Dordrecht, 1998.

[16] K.-F. Lee, H.-W. Hon, Speaker-independent phone recognition using hidden Markov models, IEEE Trans. Acoust. Speech, Signal Process. 37 (11) (November 1989) 1641–1648.

[17] J.H. Lee, H.Y. Jung, T.W. Lee, S.Y. Lee, Speech feature extraction using independent component analysis, in: Proceedings of the International Conference Acoustics, Speech, Signal Processing, Istanbul, Turkey, June 2000, pp. 1631–1634.

[18] J.-H. Lee, T.-W. Lee, H.-Y. Jung, S.-Y. Lee, On the efficient speech feature extraction based on independent component analysis, Neural Process. Lett. 15 (3) (June 2002) 235–245.

[19] M.S. Lewicki, Efficient coding of natural sounds, Nat. Neurosci. 5 (4) (2002) 356–363.

[20] B.A. Olshausen, D.J. Field, Emergence of simple-cell receptive field properties by learning a sparse code for natural images, Nature 381 (1996) 607–609.

[21] D. O'Shaughnessy, Speech Communication: Human and Machine, IEEE Press, New York, 1999.

[22] L. Parra, C. Spence, P. Sajda, Higher-order statistical properties arising from the non-stationarity of natural signals, in: Advances in Neural Information Processing Systems, Vol. 13, MIT Press, Cambridge, MA, 2001.

[23] L. Potamitis, N. Fakotakis, G. Kokkinakis, Independent component analysis applied to feature extraction for robust automatic speech recognition, Electron. Lett. 36 (23) (November 2000) 1977–1978.

[24] O. Schwartz, E.E. Simoncelli, Natural sound statistics and divisive normalization in the auditory system, in: Advances in Neural Information Processing Systems, Vol. 13, MIT Press, Cambridge, MA, 2001.

[25] P. Somervuo, Experiments with linear and nonlinear feature transformations in HMM based phone recognition, in: Proceedings of the International Conference on Acoustics, Speech, Signal Processing, Hong Kong, China, April 2003, pp. I-52–I.55.

[26] S.J. Young, The general use of tying in phoneme-based HMM speech recognisers, in: Proceedings of the International Conference on Acoustics, Speech, Signal Processing, San Francisco, CA, March 1992, pp. 1569–1572.

[27] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, P. Woodland, The HTK Book, Cambridge University Engineering Department, Cambridge, UK, 2002.

[28] Q. Zhu, A. Alwan, An efficient and scalable 2D DCT-based feature coding scheme for remote speech recognition, in: Proceedings of the International Conference on Acoustics, Speech, Signal Processing, Salt Lake City, UT, 2001.

[29] R.E. Ziemer, W.H. Tranter, Principles of Communications: Systems, Modulation, and Noise, 5th Edition, Wiley, New York, 2002, pp. 76–83.