Applying Independent Component Analysis for Speech Feature Detection*

Włodzimierz Kasprzak , Adam F. Okazaki

Warsaw University of Technology, Institute of Control and Computation Engineering
ul. Nowowiejska 15-19, PL - 00-665 Warsaw, Poland
e-mail: W.Kasprzak@ia.pw.edu.pl, A.Okazaki@elka.pw.edu.pl

Abstract— **An approach to speech feature detection is developed, which uses the technique of independent component analysis for a blind (unsupervised learning) detection of basic vectors in the Fourier space. This kind of features could replace the Mel Frequency Cepstrum Coefficient (MFCC) features, widely used today for phoneme-based speech recognition. Alternatively, the ICA components could act as basic features in speaker verification systems.**

Keywords: Independent components, speech features, speaker identification, unsupervised learning.

I. Introduction

It is common in automatic speech recognition and identification systems to apply a frame-based segmentation of the signal; i.e. to use short-time frames [1], [2]. Specific features of a single frame are detected either in the time-domain (like LPC features) Fourier space (power coefficients) or in "cepstral" space (a homomorphic filtering via the Fourier space back to the time domain). Among these, the Mel Frequency Cepstrum Coefficients (MFCC) are widely used [2], [3].

Recently it was observed, that statistical cues could offer increased power to speaker recognition systems [4], [5]. In this context the two techniques PCA and ICA can be considered [6], [7].

Assuming, that there exists independent component of measured signals, the goal of ICA is to reconstruct both the estimates of original sources (i.e. the basic vectors of our "independent component" space) and the (inverse of) "mixing" coefficients [6]. Some ambiguity is inherently included; i.e. the permutation order and the scaling factors cannot be reliably detected. In our application these are not drawbacks, as we are interested in the complete set and not in individual vectors, whereas the scaling of vectors is not relevant.

Different authors derive the principal component analysis (PCA) or ICA [4] of the power spectra vectors, which are also smoothed using Mel-scale triangular filters. Resulting features are further narrowed down using a linear discriminate based criterion. The authors of [5] assume that the spectra of sounds generated by a given speaker can be synthesized using a set of speaker specific basis functions - the unknown source in the ICA model.

In this paper we expect that the Fourier power coefficients of a single frame can be mixed from a set of basic vectors. These basic vectors are assumed to be statistically independent. At first we want to verify if the basic vectors itself can be used as an identification criteria. In other words, in first case we are not tracking the mixture coefficients, that should vary depending on the phoneme and person but we only estimate the source set in the speaker's spectra and compare it with the learned set for an individual speaker. In the second case we work with a fixed set of basic vectors - determined in the learning phase - and we care about the mixture coefficients that should vary depending on the phoneme and person.

In section 2 the problems of speaker identification and independent component are shortly introduced. The two proposed approaches are described in section 3. Some simulation results are presented in section 4.

II. The ICA problem and solution

A. The task of ICA

In Independent Component Analysis we assume that there exist m zero-mean source signals, $s_1(t), ..., s_m(t)$, that are scalar-valued and mutually (spatially) statistically independent (in practice: as independent as possible) at each time instant or index value t. The original

* The work reported in this paper was supported by the grant MNiI - 4T11A 003 25.

sources $s_j(t)$ are unknown to the observer, who has to deal with n possibly noisy but different linear mixtures, $x_1(t), ..., x_n(t)$, of the sources (usually for $n >= m$). The mixing coefficients are some unknown constants.

The task of ICA is to find the waveforms $s_i(t)$ of the sources, knowing only the mixtures $x_j(t)$ and the number m of sources [6]. Denote by $x(t) = [x_1(t), ..., x_n(t)]^T$ the n-dimensional t-th data vector made up of the mixtures at discrete index value (usually time) t. The ICA mixing model is equal to:

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) + \mathbf{n}(t) = \sum_{i=1}^{m} s_i(t)a_i + \mathbf{n}(t). \tag{1}$$

Let us assume further that in the general case the noise signal has a Gaussian distribution but none of the sources is Gaussian. In the simplified case at most one of the source signals $s_i(t)$ is allowed to have a Gaussian distribution. These assumptions follow from the fact that it is impossible to distinguish several Gaussian sources from each other.

In standard source separation approach, an $m \times n$ separating matrix $\mathbf{W}(t)$ is updated so that the m-vector $\mathbf{y}(t) = \mathbf{W}(t)\mathbf{x}(t)$ becomes an estimate of the original independent source signals. $\mathbf{y}(t)$ is the output vector of the network and the matrix $\mathbf{W}(t)$ is the total weight matrix between the input and output layers.

B. Pre-processing in ICA

The elimination of mean value - results in an algorithm simplification. Let $\mathbf{m}$ be the mean vector of time series (observation vector) $\mathbf{x}(t)$. After estimating the sources in ICA their means can be reconstructed from:$\mathbf{A}^{-1}\mathbf{m}$.

"Whitening" - a linear transformation such that the observation vector elements will be uncorrelated and with unit variances:

$$E\{\tilde{\mathbf{x}}\tilde{\mathbf{x}}\} = \mathbf{I}. \tag{2}$$

Whitening allows the reduction of the ICA search problem from n^2 free matrix coefficients to only $n(n-1)/2$ elements, as the matrix must be kept orthogonal (figure 1). Even the matrix size reduction could be possible, if some eigenvalues λ_j are to small.

After whitening we get a "rectangular"-like distribution of samples in every 2-dimensional subspace - it is sufficient to find one rotation angle (not two) by an ICA procedure. In contrast to whitening, after PCA the mixed vector samples are only rotated and the marginal 1-D distributions are not mutually independent.

C. On-line update rule

A well-known iterative optimization method is the stochastic gradient (or gradient descent) search [8]. In this method the basic task is to define a criterion $J(\mathbf{W}(k))$, which obtains its minimum for some $\mathbf{W}_{opt}$, which is the estimated optimum solution.

For the ICA problem another gradient approach was developed recently - the *natural gradient descent*. This was achieved by Amari et al. [9] (who gave a theoretical justification by using the Riemannian space notation), Cichocki and Unbehauen (which proposed an algorithm justified by computer simulations) and Cardoso and Laheld (that introduced the relative gradient). The natural gradient takes the form:

$$\mathbf{W}(k+1) = \mathbf{W}(k) - \theta(k)\frac{\partial J(\mathbf{W}(k))}{\partial \mathbf{W}(k)}\mathbf{W}^T(k)\mathbf{W}(k). \tag{3}$$

Different theoretical justifications of the ICA, like the Kullback-Leibler divergence minimization, the information maximization and the mutual information minimization lead to the same cost function [6]:

$$J(\mathbf{W}, y) = -log(det(\mathbf{W})) - \sum_{i=1}^{n} log(p_i(y_i)) \tag{4}$$

where the $p_i(y_i)$ -s are pdf's of signals y_i respectively, $det(\mathbf{W})$ is the determinant of the matrix $\mathbf{W}$.

The criterion is minimized if the border distributions (individual variables) are independent. After the approximation of (unknown) distribution shapes by means of fourth order moments, the weight update rule can be derived.

Applying the natural gradient approach we may derive the learning rule for on-line ICA as:

$$\Delta\mathbf{W}(k) = \theta(k)[\mathbf{I} - f(\mathbf{y}(k))\mathbf{y}^T(k)]\mathbf{W}(k). \tag{5}$$

D. "FastICA"

An efficient "batch" approach is the method "FastICA" of Hyvarinen et al. [7]. The batch processing allows a preliminary "whitening" step for the zero-mean mixture signals, which improves the convergence speed of the ICA procedure.

A. Initialize nonzero weights $\mathbf{W}$.

B. Iterate:

1. FOR outputs $p = 1, ..., n$; perform steps (2-4):
2. Vector update:

$$\mathbf{w}_p^+ = E\{\mathbf{x}g(\mathbf{w_p^T x})\} - E\{g'(\mathbf{w_p^T x})\}\mathbf{w}_p \tag{6}$$

where g is a nonlinear function, g' its first derivative with respect to time.

3. Normalize to a unitary-length vector:

$$\mathbf{w}_p = \frac{\mathbf{w^+}_p}{||\mathbf{w^+}_p||}. \tag{7}$$

4. De-correlation of current vector against the previous vector set (by a Gram-Schmidt orthogonal)

$$\mathbf{w}_p = \mathbf{w}_p - \sum_{j=1}^{p-1} \mathbf{w}_p^T\mathbf{w}_j\mathbf{w}_j; \; \mathbf{w}_p = \frac{\mathbf{w}_p}{||\mathbf{w}_p||}. \tag{8}$$

Alternatively a symmetric de-correlation of a whole weight matrix can be performed:

$$(\mathbf{W} = (\mathbf{W}\mathbf{W^T})^{-1/2}\mathbf{W}. \tag{9}$$

5. If $\mathbf{W}$ has not yet converged then repeat from step 1.

III. The two approaches

A. The standard MFCC features

The short-term power spectrum is computed by applying the discrete Fourier Transform (DFT) (in fact the FFT) to each windowed signal and taking directly the magnitudes of Fourier coefficients raised to the power of two. The power spectrum is usually represented on a log scale. Due to the properties of the log function, the shape of the log power spectrum is preserved, independent of the input signal strength.

A MEL scale (empirical result) adopts the frequency bandwidths to the bandwidths recognized by the human auditory system. The set of Fourier features is reduced by considering bandwidths, centered around some MEL scale frequencies. Usually one uses a set of l triangle filters $D(l, k)$ to compute l so called Mel-spectral coefficients MFC(l, t) for each signal window.

Since the vocal tract is smooth, energy levels in adjacent bands tend to be correlated. The inverse DFT (in fact only the cosine transform as the transformed MFC's are real-valued) converts the set of logarithm-scaled energies to a set of cepstrum coefficients (usually m = 12), which are largely un-correlated.

B. Approach 1: ICA sources for a single speaker

In this approach we consider learning samples for a single speaker only. The Fourier coefficients for given frame $FC(., t)$ constitute a single vector $\mathbf{x}(t)$ - a single (mixture) input to the ICA learning procedure (the size of vector is N). This is expected to be a particular mixture of $m < N$ independent sources. From the spoken word we get a learning set of frames for given speaker: $x_i(t)(i = 1, ..., n; t = 1, ..., N)$.

The basic mixing model in ICA (without noise) is assumed. $\mathbf{x}(t)$ is a matrix of n time-varying vector signals, each of size N. $\mathbf{a}_i$ is a set of n mixing vectors (each of size m) combined to a mixing matrix $\mathbf{A}$ (every $\mathbf{a}_i$ is a single row of matrix $\mathbf{A}$). $\mathbf{s}_i$ is a set of m sources - each of size N.

After ICA both unknown sources and unknown mixing coefficients are determined - on base of given sequence of observations (frames) $x_i(t)$ the vector $\mathbf{s}$ and weight matrix $\mathbf{W}$ are estimated.

We obtain for each speech sample its corresponding best-suitable set of sources. The decision (classification) rule can take into account, if this source set is similar to the learned source set of given speaker or not.

C. Approach 2: ICA sources for many speakers

In this approach we try to learn a set of sources in ICA (basis vectors), which is common to some (all available) speakers, but limited to one specific utterance (word). Every word is now recognized by a unique set of mixing coefficients. Now the word recognition process consists of following steps:

1) compute a selected spectrogram for tested speech;

2) with the ICA components, obtained for a word to be recognized, estimate the coefficients $\mathbf{W}$ corresponding to tested spectrogram;

3) classify the image of coefficients $\mathbf{W}$.

IV. Experimental results

The first approach described in section 3 was implemented and tested on speech signal examples, acquired with sampling frequency of 22 kHz. Polish spoken digits from 18 persons (both male and female) were available for testing. Figures 1-3 document the process of computing PCA components (either the reference- or tested components) for two speakers.

For the purpose of speaker identification or word recognition the ICA components obtained for currently tested speech need to be matched with the reference components of given speaker (or given word). During the comparison of these two sets we need to establish the proper permutation index, the scaling and even the sign of amplitude for the individual components. The matching procedure is as follows ([10]):

(1) The amplitudes of all components are re-scaled to an uniform interval of <-1, 1>.

(2) FOR all tested components $y_i, (i = 1, ..., n)$:

FOR all reference components $s_j, (j = 1., ..., n)$:

compute the mean square error of approximating s_j by y_i or by $-y_i$: $MSE[y_i, s_j]$ and $MSE[-y_i, s_j]$ and select the better one, i.e. with lower value;

(3) All selected MSE-s are transformed into elements of a new created matrix $\mathbf{P} = [a_{i,j}]_{n \times n}$, where $a_i = 1/\sqrt{MSE[y_i, s_j]}$.

(4) The error index $EI(\mathbf{P})$ is computed as:

$$\frac{1}{n}[\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{a_{ij}}{max_i(a_{ij})} - n] + \frac{1}{n}[\sum_{j=1}^{n}\sum_{i=1}^{n}\frac{a_{ij}}{max_j(a_{ik})} - n].$$

The first part of above sum expresses the average error for matching a tested ICA component with one reference component, whereas the second part is equivalent to a penalty score, if a single reference component is matched with more than one tested component.

Some experiments of both approaches are summarized in tables 1-2, where the EI index values were computed while matching the tested sample components with the proper reference components. From Table 1 it is evident, that the errors are quite independent from the speaker. Hence, let us fix one set of reference ICA components $\mathbf{S}$ and estimate the mixing coefficients: $\mathbf{W} = \mathbf{X}\mathtt{inv}(\mathbf{S})$, where $\mathbf{X}$ is the selectively chosen spectrogram for given speech sample. In Fig. 4 two sets of coefficients $\mathbf{W}$ are presented for the same word "zero" pronounced by two speakers. It is evident, that both sets are more similar than their spectrograms.

From Table 2 we conclude that ICA components, obtained for the same speaker but for different words are also similar. Hence our approach seems to produce a general base for speech recognition. In Fig. 5 we see that the coefficient matrices $\mathbf{W}$ for words "jeden" and "dwa"

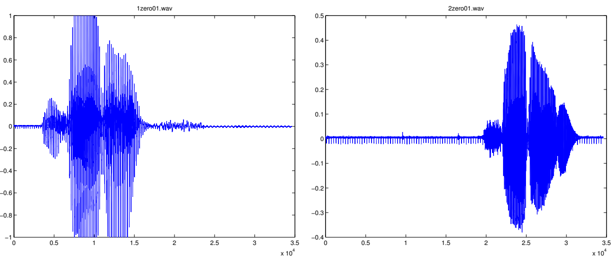

Fig. 1. Waveforms of the word "zero" pronounced by two speakers (male and female).

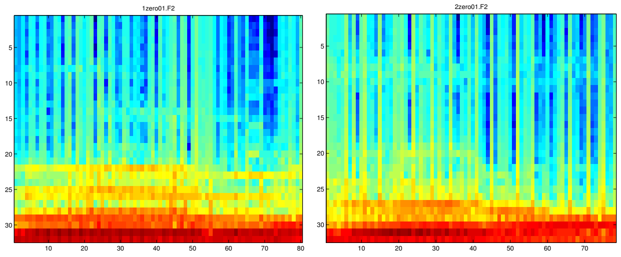

Fig. 2. The spectrograms (selected frames with sufficient energy only) for above words "zero".

acquired from the mixture with the ICA components coming from word "zero" are quite different, than the coefficients for word "zero". An easy word detection should follow.

V. Conclusion

We have proposed two approaches for the use of ICA in speaker identification and word recognition. The first experiments seems to be promising. Although speaker identification and word recognition have quite different goals - to discriminate between different pronunciations or to generalize pronunciations - in both cases the ICA provides a general basis for feature detection.

Table 1. Comparison of the error index $EI(\mathbf{P})$ in the first experiment - the same word ("zero") but 4 different speakers (31 components with 32 elements each).

Reference components	M1	F1	M2	F2
Tested components				
Male 1	6.04	4.46	5.15	3.90
Female 1	6.15	4.62	5.85	5.56
Male 2	6.21	4.47	5.13	4.70
Female 2	9.03	8.47	7.45	7.92

Table 2. Comparison of the error index in the second experiment - different words ("zero", "jeden", "dwa") but the same speaker (31 components with 32 elements each).

Reference components	"zero"	"jeden"	"dwa"
Tested components			
"zero"	3.46	2.50	1.98
"jeden"	2.33	2.82	1.20
"dwa"	2.66	2.94	1.85

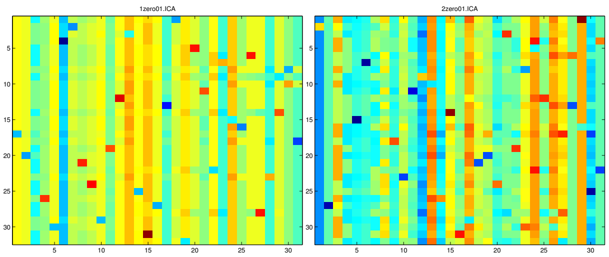

Fig. 3. The detected 31 basic vectors (one column represents one vector with 32 elements) after ICA was applied to above two spectral images.

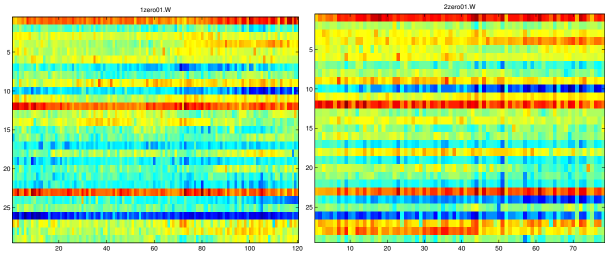

Fig. 4. The coefficients $\mathbf{W}$ (one column represents one vector of coefficients for one signal frame) for two speech samples from different speakers, with the same ICA components computed for the word "zero".

References

[1] J.-C. Junqua, J.-P. Haton: *Robustness in automatic speech recognition*, Kluwer Academic Publications, Boston etc., 1996.

[2] L. Rabiner and B. Juang: *Fundamentals of Speech Recognition*, Prentice Hall, 1993.

[3] S. Grocholewski: *Statystyczne podstawy systemu ARM dla jezyka polskiego*, Rozprawy Nr. 362, Wyd. Politechniki Poznanskiej, 2001.

[4] P. Ding, X. Kang, and L. Zhang, "Personal recognition using ICA", *Proceedings ICONIP*, 2001.

[5] J. Rosca, A. Kofmehl: "Cepstrum-like ICA representations for text independent speaker recognition", *Procceedings of ICA'2003*, (Nara, Japan, April 2003), Publ. by NTT Kyoto, Japan, .

[6] A. Cichocki, S. Amari: *Adaptive Blind Signal and Image Processing*, John Wiley, Chichester, UK, 2002.

[7] A. Hyvarinen, J. Karhunen, E. Oja: *Independent Component Analysis*, John Wiley & Sons, New York etc., 2001.

[8] R.O. Duda, P.E. Hart: *Pattern classification and scene analysis*, John Wiley & Sons, New York, 1973.

[9] S. Amari, S.C. Douglas, A. Cichocki, H.Y. Yang: "Novel online adaptive learning algorithms for blind deconvolution using the natural gradient approach". *IEEE Signal Proc Workshop on Signal Processing Advances in Wireless Communications*, April 1997, Paris, 107 - 112.

[10] W. Kasprzak: *Adaptive computation methods in digital image sequence analysis*, Prace Naukowe - Elektronika, Nr. 127 (2000), Oficyna Wydawnicza PW, Warszawa, 170 pages.

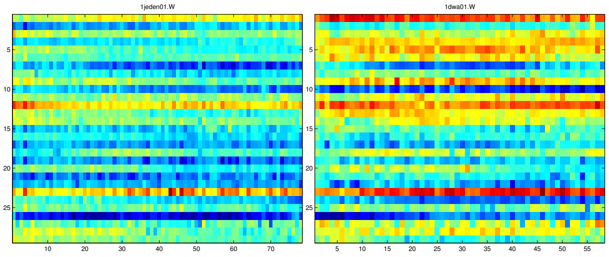

Fig. 5. The coefficients $\mathbf{W}$ for different words "jeden" and "dwa" from the same speaker. The ICA components come from a different word "zero".