4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003), April 2003, Nara, Japan

CEPSTRUM-LIKE ICA REPRESENTATIONS FOR TEXT INDEPENDENT SPEAKER RECOGNITION

Justinian Rosca and Andri Kofmehl

Siemens Corporate Research, 755 College Road East, Princeton, NJ 08540
justinian.rosca@scr.siemens.com,akofmehl@ee.ethz.ch

ABSTRACT

Automatic methods to determine voiceprints in speech samples predominantly use short-time spectra to yield specific features of a given speaker. Among these, the Mel Frequency Cepstrum Coefficient (MFCC) features are widely used today. The speaker recognition method presented here is based on short-time spectra, however the feature extraction process does not correspond to the MFCC process. The motivation was to avoid what we see as shortcomings of present approaches, particularly the blurring effect in the frequency domain, which confuses rather than helps in distinguishing speakers. We introduce a speech synthesis model that can be identified using Independent Component Analysis (ICA). The ICA representations of log spectral data result in cepstral-like, independent coefficients, which capture correlations among frequency bands specific to the given speaker. It also results in speaker specific basis functions. Coefficients determined from test data using a speaker's true basis functions show a low degree of correlation, while those determined using other basis functions do not. This enables the system to reliably recognize speakers. The resulting speaker recognition method is text-independent, invariant over time, and robust to channel variability. Its effectiveness has been tested in representing and recognizing speakers from a set of 462 people from the TIMIT database.

1. INTRODUCTION

Human aural discrimination typically manifests itself by capability of audio separation in frequency (e.g. pure tones separable in frequency) and in time (e.g. successive impulses or clicks). Furthermore, psychologists have identified another discriminative power of the human system, which has been less studied and is called *statistical separation* [1]. Examples of statistical separation and discrimination are the human power for filling in the masked parts of audio, recognizing what language a person speaks, whether one or more people are speaking, or qualitative speech cues such as diction, prosody, rhythm and intonation. Statistical features are associated with high-level perceptual cues in speech, which are extremely difficult to automatically extract by computers [2]. While frequency and time based features are heavily used in today's speaker recognition systems, statistical cues or supra-segmental features [3] could offer increased power to speaker recognition systems.

In this paper we attempt to merge two lines of thought: (1) primary features to be used in speaker recognition are cepstral feature vectors, and (2) independent component analysis extracts useful (higher order) statistics from the data, and may help to capture speaker dependent statistical cues in a text-independent system. The first assumption, dominant in the speech and speaker recognition literature todate, agrees to and exploits discrimination in frequency and time. The second assumption may be related to the statistical separation power of the human system. ICA is to be used for feature extraction. How exactly could we do this in the context of assumption (1) and why would statistics captured by ICA be related to human aural discrimination? This is the subject of the work reported here.

Recent speaker recognition literature discusses the use of ICA-based feature extraction approaches, but to a large extent neglects assumption (1) above. For instance, [4] assumes a linear superposition model $x = As$ in the time domain for speech samples x of a given speaker. Here "sources" or underlying features or basis functions s for a speaker are independent and have generalized Gaussian distributions. The ICA approach attempts to determine the mixing recipe A, and thereby audio features, using matched or learned distributions of s. The relevance of features in not entirely transparent and a relation to studied auditory system features that are specific to a speaker is not clear. By employing time-domain data, the approach may be inherently sensitive to noise and the content of the training data.

An approach for speaker verification similar in spirit to our proposal but different in detail is presented in [5, 6]. The authors derive the principal component analysis (PCA) [5] or ICA [6] of power spectra smoothed using Mel-scale triangular filters. Resulting features are further narrowed down using a linear discriminant based criterion. The approach intuitively follows assumption (1) by using power spectral information, however power spectral features are

muffled by the use of Mel-frequency filters in accordance with present speech and speaker recognition literature. Features such as correlations between frequencies, which are functions of a speaker's glottal shape may be averaged over and lost. In this paper we offer a formal justification about what processing steps make sense and would work well in conjunction with feature extraction techniques such as PCA or ICA.

The present speaker verification method is based on the idea that the spectra of sounds generated by a given speaker can be synthesized using a set of speaker specific *basis functions*. This could explain spectrogram correlation among frequencies that are specific to individualized glottal or nasal features. This specificity is exactly what an automatic speaker recognition system should rely on. As speech sounds have co-evolved to be distinct [7], the basis functions are combined in uncorrelated or independent way over time across one person's speech. The basis functions can be determined using independent component analysis techniques. Functions are distinct for every speaker, and coefficients determined from test data using a speaker's true basis functions do show independence or a low degree of correlation. In contrast, coefficients determined using other basis functions depart from this norm. This idea offers a criterion to automatically recognize speakers. Moreover, the ICA representations of log spectral data result in cepstral-like coefficients to capture correlations among frequency bands specific to the given speaker, therefore the literature about properties of the cepstrum of human speech can be carried over and interpreted from the new perspective.

This paper describes in detail some of these ideas. Section 2 introduces the synthesis model at the foundation of the ICA approach, and shows why features extracted have a cepstrum-like interpretation. Section 3 presents PCA and ICA based algorithms for speaker recognition. Section 4 discusses experimental results on a 462 speaker database. We conclude with a summary of main directions to be pursued to finesse the presented approach.

2. CEPSTRUM-LIKE FACTORIAL MODEL FOR SPEAKER RECOGNITION

A model for automatic speaker recognition should be capable of powers available in typical human aural discrimination. It is generally agreed that the most important requirements for an automatic speaker recognition model are capability to capture naturally and frequently occurring cues in speech, invariance over data/time, invariance to environment, insusceptibility to mimicry, and overall effectiveness in uniquely representing speakers [8].

Here we will use ICA [9] to approach the problem of discovering hidden factors that underlie speech and arguably cover some of these most important requirements.

2.1. Simplified speech synthesis model

Every person's voice has unique, distinguishing features. Speech is produced by modifying the air stream provided by the lungs such that a desired sequence of sounds is generated. Differences in speech originate from the characteristics of the vocal tract shape, the vocal cords, and nasal cavity. The vocal tract can then be seen as a set of filters that alter a set of excitation signals [10].

Let us assume a speech signal $x(t)$ with a proper short-time spectral based representation $\mathbf{x}_t(\omega) \doteq \mathbf{x}_t$ (e.g. the log power spectrum $\mathbf{x}_t = \log|\mathbf{x}(\omega, t)|^2$). The idea is to capture correlations among frequencies in the observed representation of speech spectra $\mathbf{x}$ as linear combinations of basis filter functions. Assume that each person is characterized by a specific set of basis functions, which are then combined in a statistically independent manner to form specific sounds:

$$\mathbf{x}_t = \sum_{m=1}^{M} s_{m,t}\, \mathbf{a}_m \quad \text{or} \quad \mathbf{x}_t = \mathbf{A}\mathbf{s}_t \tag{1}$$

where M is the number of basis functions $\mathbf{a}$ and $s_{m,t}$ are independent coefficients (latent random variables). A matrix format is obtained by stacking horizontally vectors $\mathbf{x}_t$ and $\mathbf{s}_t$ ($1 \leq t < L$) and regarding $\mathbf{a}_m$ as columns of matrix A:

$$\mathbf{X} = \mathbf{A} \cdot \mathbf{S} \tag{2}$$

2.2. Model estimation by ICA

If the basis functions can be robustly extracted from speech, i.e. $\mathbf{A}$ can be estimated from observed data to ensure independence of the latent random variables $\mathbf{S}$, then they would be useful for speaker recognition. ICA can fit the bill. The model (2) is identifiable under the assumption that $\mathbf{S}$ have non-Gaussian distributions. The ICA problem is to determine the matrix $\mathbf{W} \simeq \mathbf{A}^{-1}$ which ensures the independence of the coefficients (lines of) $\mathbf{U}$:

$$\mathbf{W} \cdot \mathbf{X} = \mathbf{U} \tag{3}$$

Our sound synthesis model is analogous to the view taken in [11] for natural images, and used for instance in the ICA-based "factorial codes" architecture for face recognition in [12]. This is represented in Figure 1.

$\mathbf{X}$'s columns are transformations of frames of speech data (e.g. time-domain frames in [4], smoothed spectra in [6], etc; an appropriate representation for speech frames is to be determined yet.) Speech excitation $\mathbf{S}$ is unknown and represents the speech causes. It activates one person's basic speech components or features (speech basis) $\mathbf{A}$ to result in actual speech frames $\mathbf{X}$. For identification of this model by ICA, speech is filtered using filters $\mathbf{W}$ to generate statistically independent variables (coefficients) for coding speech segments. Once $\mathbf{W}$ and $\mathbf{U}$ are determined by ICA, the given

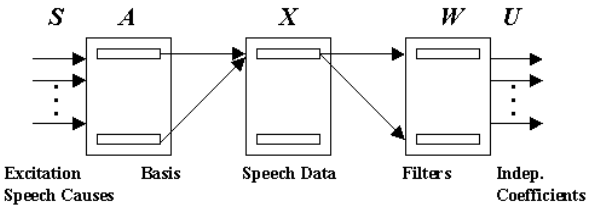

Fig. 1. Speech sythesis model for ICA-based feature extraction by $\mathbf{W} \cdot \mathbf{X} = \mathbf{U}$.

speech data $\mathbf{X}$ can be reconstructed by $\mathbf{X} = \mathbf{W}^{-1} \cdot \mathbf{U}$, which shows that each (frame) column of $\mathbf{X}$ is a linear combination of the columns of $\mathbf{W}^{-1}$. The coefficients of the linear combination are given by the corresponding column of $\mathbf{U}$, and represent statistically independent variables for coding speech. $\mathbf{U}$ is the *ICA factorial representation* of the data.

$$\mathbf{x}_k = \sum_{m=1}^{M} u_{mk} (\mathbf{W^{-1}})_{\mathbf{m}} \tag{4}$$

2.3. Cepstrum-like data representation

If data $\mathbf{x}$ were the log spectrum of the time-domain speech data and if we replaced $\mathbf{W}$ by filters given by the Fourier series, then the factorial representation (4) would correspond to the LPC cepstrum:

$$[\log X(\omega)] = \sum_{-\infty}^{+\infty} u_m \cdot [e^{-jm\omega}] \tag{5}$$

Indeed, assume speech is modeled by a minimum phase all-pole model $\frac{\sigma}{A(e^{j\omega})}$. It's spectrum is given by $S(\omega) = \frac{\sigma^2}{|A(e^{j\omega})|^2}$. The Taylor expansion of $\log S(\omega)$ exists for a stable all-pole model, as $\log A(e^{j\omega})$ is analytic in the unit circle. The expansion is the LPC cepstrum [13]:

$$\begin{aligned} [\log \frac{\sigma^2}{|A(e^{j\omega})|^2}] \quad &= \textstyle\sum_{-\infty}^{+\infty} c_m \cdot [e^{-jm\omega}] \\ c_0 = \log(\sigma^2); \qquad & c_{-m} = c_m \end{aligned} \tag{6}$$

Note that equation (6) is almost identical to (5), which is further paralleled by (4). Cepstrum is a fast decaying sequence with m, and the first several cepstrum coefficients determine the all-pole filter. This, in addition to the symmetry of cepstral coefficients c_m implies that a finite sum in 5 is a good approximation. The variability of these coefficients is primarily due to variations in speaker characteristics and vocal effects. Coefficients are functions of the speaker's vocal tract characteristics. Normally, in speech recognition, such sources of variability are deemphasized by "liftering" (i.e. filtering in the cepstral domain of the data) and "pre-emphasis" processing. On the contrary, the pre-emphasis is not desirable in a speaker recognition problem, where it is exactly such variability that should be preserved.

When filters $\mathbf{W}$ are learned by means of ICA, we get a cepstrum-like factorial representation of speech $\mathbf{U}$. $\mathbf{U}$'s columns represent features to be used in training and in the recognition process itself.

The statistical independence constraint imposed by ICA learning induces a *cepstrum-like* representation of the data. Variations in the transmission/recording channel should not affect learned features. The channel model can be assumed to be convolutional in time domain, i.e. multiplicative in the frequency domain. Therefore, mean subtraction in the *log* spectral domain, analogous to cepstral mean subtraction, can be applied on this representation to factor out channel effects. We conclude that a proper choice of representation is log spectral data, which induces cepstral-like coefficients in $\mathbf{U}$.

ICA coefficients $\mathbf{U}$ inherit a property similar to a distortion measure in the spectral domain, which results by applying Parseval's theorem to a pair of spectra:

$$\sum (u_m - u_m^{'})^2 \sim \frac{1}{2\pi} \int_{-\pi}^{\pi} |\log S(\omega) - \log S^{'}(\omega)|^2 d\omega \tag{7}$$

The L_2 distance alike a cepstral distance could be useful if the ICA representation of the data were employed in more constrained problems such as text dependent verification.

3. ALGORITHMS FOR SPEAKER RECOGNITION

So far we used model (2) and its estimation (3) by ICA. We also concluded that a suitable representation domain of data is the log short-time spectral domain. In order to remove channel convolutional effects, we remove the mean values similarly to the cepstral mean subtraction practice. The basis functions resulting from the ICA estimation process represent speaker-specific features. The training process and the overall speaker recognition algorithm are described in detail below.

3.1. Feature extraction: Learning PCA and ICA basis functions

Data whitening is a useful data preprocessing technique before applying ICA [9]. Preprocessing is done on full resolution frequency domain data for a speaker. Rather than smoothing spectral data to reduce its dimensionality (e.g. by using Mel scale filters), we perform a Karhunen-Loeve low rank approximation [14]. This ensures we preserve sufficient frequency resolution while reducing the dimensionality as well. The step is typically implemented by PCA or

eigenvalue decomposition:

$$\mathbf{Y} = \mathbf{W}_{\text{PCA}} \cdot \mathbf{X} \tag{8}$$

Dimensions in new space correspond to a linear combination of original features with the largest variance in the data, therefore the approximation of the data in the reduced dimensionality space is optimal in the minimum mean square error (MMSE) sense for data reconstruction. The $\mathbf{W}_{\text{PCA}}$ preprocessing of the data compresses data to a space of dimension N, $N \leq M$, where data in each new dimension is decorrelated from the others and has unit variance.

$\mathbf{W}_{\text{PCA}}$ could well be used as the basis functions. This transformation of the data results in uncorrelated components only, rather than independent components. In order to obtain independent components, the ICA step is required:

$$\mathbf{U} = \mathbf{W}^{'} \cdot \mathbf{Y} = \mathbf{W}_{\text{ICA}} \cdot \mathbf{X} \tag{9}$$

The overall data transformation is a matrix $\mathbf{W}$ to project data into its new representation $\mathbf{U}$. Thus we cover both the PCA and ICA cases:

$$\mathbf{U} = \mathbf{W} \cdot \mathbf{X} \tag{10}$$

The feature transformation steps for learning speaker specific basis functions from training data are presented in Figure 2.

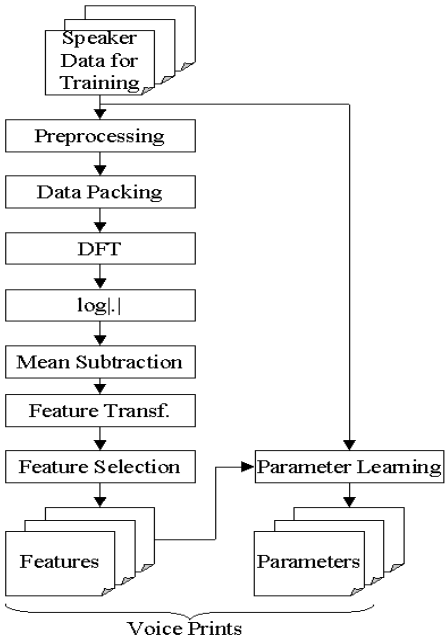

Fig. 2. Architecture of speaker recognition system. The feature transformation is performed by PCA/ICA.

3.2. Classification

The transformation $\mathbf{W}$ is created to induce uncorrelated or statistically independent coefficients for speech data from a given speaker. Assume that base vectors $\mathbf{W}^{\text{TRAIN}}$ are learned from a large set of training data of the speaker. Test data from the same (true) speaker will induce a similarly low degree of correlation or independence when test data is projected on the learned basis vectors. However, if test data from a different speaker is used, her data is unlikely to produce a similarly low degree of correlation. The basis vectors are not designed to minimize correlation or independence on data coming from a speaker characterized by a different correlation structure in the frequency domain. Thus a recognition score Γ is defined on the transformed data $\mathbf{U}^{\text{TEST}} = \mathbf{W}^{\text{TRAIN}} \cdot \mathbf{X}^{\text{TEST}}$:

$$\Gamma_{\mathbf{W}^{\text{TRAIN}}}(\mathbf{U}^{\text{TEST}}) = \sum_{i<j}^{N} |r_{ij}|^{\alpha} \tag{11}$$

where r_{ij} is either the crosscorrelation of $\mathbf{U}^{\text{TEST}}$ or the normalized mutual information between random variables u representing independent components, and α is a positive constant (e.g. 1). For speaker verification, a yes/no-decision has to be made by comparing the recognition score against a threshold τ:

$$\text{Recog. answer} = \begin{cases} \text{'accept'} & \text{if } \Gamma(\mathbf{U}^{\text{TEST}}) \leq \tau \\ \text{'reject'} & \text{otherwise.} \end{cases} \tag{12}$$

3.3. Recognition threshold

The threshold is learned from training data. Subsequences of the training speech sequence are used as "test" input in order to assess the degree of correlation that can be expected from a real test sequence. Consider that n subsequences are processed the same way as regular test data, to return n different correlation scores. Empirical results indicate that the distribution of score values for the true identity can be modeled by a Gamma distribution. We select the threshold such that a certain percentage of the test data subsequences (e.g. 99.5%) yield scores below it. The small percentage remaining above it and thus rejected will account for the *false reject ratio* in recognition tests. The recognition score was observed to be declining as a function of length of the test data, but if enough data is used the effect is negligible.

More generally, the procedure for learning the threshold for a given speaker identity can take into account the distribution of recognition scores for training data from other speakers. This corresponds to cohort modeling: the threshold is chosen such that the probability mass of false rejects equals the probability mass of false acceptance.

Finally, the voiceprint corresponding to a given speaker consists of the learned basis functions and threshold parameter (see Figure 2).

4. EXPERIMENTAL RESULTS

We tested both the ICA and PCA variations of the speaker recognition algorithm. The ICA implementation used was JADE [15]. We only used the recognition measure computed with correlations, due to its simplicity. This does not yet fully exploit the power of the independent component representation with ICA. We used TIMIT data downsampled at 8 kHz: a total of 462 speakers with ten sentences per speaker. We trained with eight out of the ten sentences. After silence removal we were left with 17.9 seconds of training data on average. Testing was carried out with the remaining two sentences not used for training. The average length of the test data was 5.0 seconds.

The short-time representation of signals was computed on frames of M=256 or 512 samples and an overlap b of 16 or 32 samples. Data analysis used a Hamming window and preserved the logarithm of the absolute value of half of the coefficients.

The system was trained for every single speaker identity. Accordingly, it retained 462 voiceprints: individual feature matrices and their corresponding threshold values. Thresholds were determined from training data by modeling subsequence results with a Gamma distribution. Every test sequence was used with every voiceprint in recognition tests. Altogether, $462 \times 462 = 213444$ combinations had to be evaluated, all of which returned a single recognition score and eventually an 'accept' or 'reject' answer.

Figure 3, shows the decline of the eigenvalues in a typical case. The information contained in $\mathbf{X}$ can be represented by a small number of base vectors. $N = 16$ was used here.

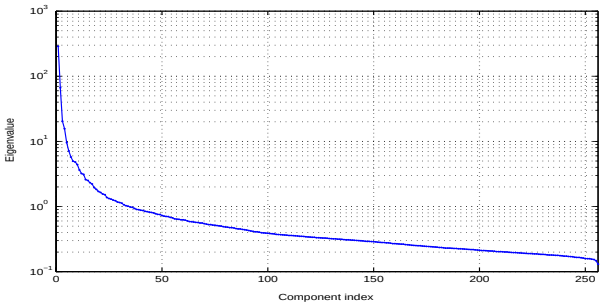

Fig. 3. Eigenvalues of the principal components sorted in descending order.

The resulting sets of PCA and ICA base vectors are shown in Figure 4. The figure highlights the accomplishment of the desiderata to capture specific speaker characteristics. Indeed, it is apparent that both PCA and ICA bases capture correlations in the frequency domain (the axis representing M spectral coefficients). In particular, there exist correlations in the low order coefficients, known to characterize speaker features. These coefficients are strong functions of the speaker's vocal tract and vocal chord duty cycles [13]. In contrast to the practice reported in the literature, we do not smooth short time spectral data.

	PCA		ICA	
M	128	256	128	256
$b = 16$	4.6 %	—	4.3 %	—
$b = 32$	4.6 %	6.1 %	4.5 %	5.5 %

Table 1. Equal error rates for various parameter settings.

The test results are shown in Figure 5. The curves show the variation of the false accept (FA) and the false reject (FR) ratio when the threshold values are altered. The value of the threshold parameter offers the handle to trade off the two ratios. The different curves correspond to different parameter settings/algorithms. The equal error rates results are reviewed in Table 4.

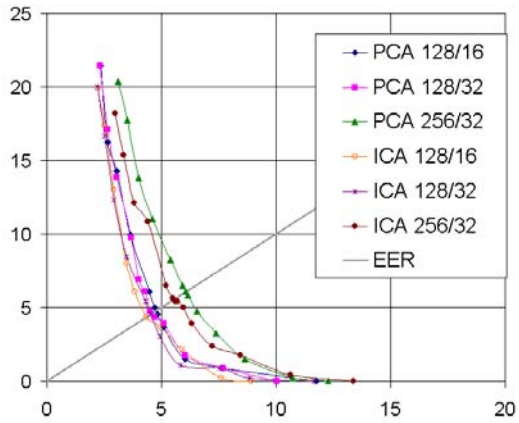

Fig. 5. FR vs. FA percentage for different parameter settings and PCA and ICA algorithms.

5. CONCLUSIONS

The speaker recognition literature surprisingly reduces the number of discrete Fourier transform samples by averaging frequency bins together. This work was motivated by the goal to overcome this prevalent but disadvantageous decision. The model and representation we employ enables us to bypass such a decision and outlines the transformation of the data by ICA. Independent components play the role

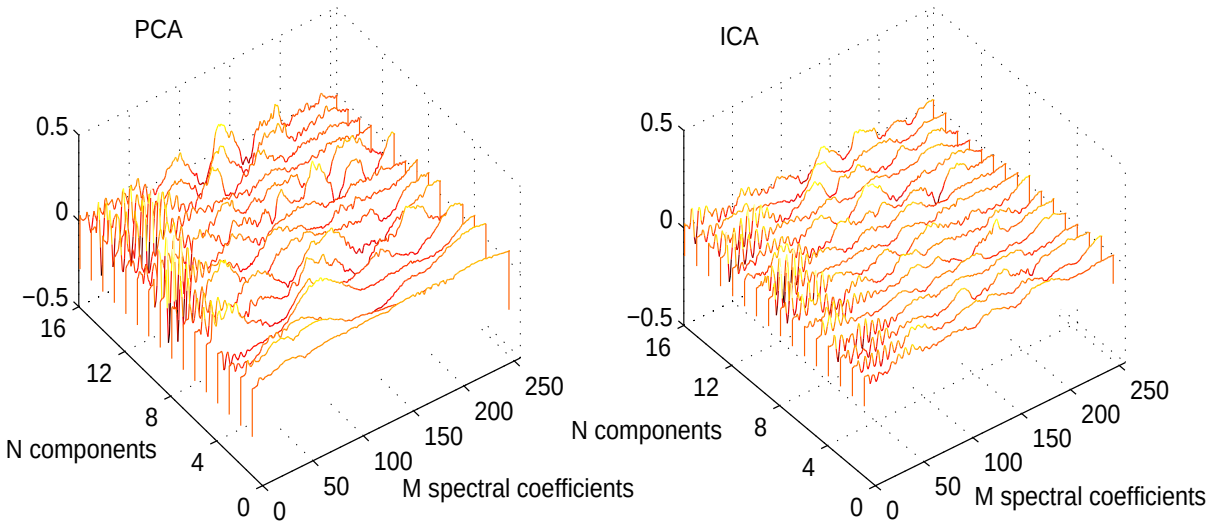

Fig. 4. PCA (left) and ICA (right) basis vectors capture correlations in the frequency domain, particularly in the low order coefficients.

of cepstral-like coefficients, but furthermore they powerfully capture those correlations among frequencies in one's speech that underly the data frequently. This capability gives our method the attribute of tacitly exploiting statistical discrimination features in the data. The clear significance of the various processing steps makes it easy to assemble the pieces together into a computationally efficient algorithm.

Tests on TIMIT data show indeed that performance of a speaker recognition system is drastically improved with the new method. While state-of-the-art systems of text-independent speaker recognition report equal-error-rates of the order of 7-15% for two minutes of training data and 30 seconds of test phone data [2] (results matched in our tests when using standard MFCC features), our ICA-based system showed equal error rates of 4.3% when using just about 18 seconds of training and 5 seconds of test training data. Experiments are ongoing to verify the claim of increased robustness in the presence of increased noise. In addition, future work will explore a recognition measure based on mutual information and the integration of alternative classification or speaker modeling approaches such as vector quantization and Gaussian mixture models.

6. REFERENCES

[1] E. C. Cherry, "Some experiments on the recognition of speech, with one and two ears," *The Journal of the acoustical society of America*, vol. 25, no. 5, 1953.

[2] D. Reynolds, "An overview of automatic speaker recognition technology," in *Proceedings ICASSP*, 2002, vol. IV, pp. 4072–4075.

[3] S.Furui, "Future directions in speech information processing," in *Acoustical Society of America ASA-98*, 1998.

[4] G.-J. Jang, T.-W. Lee, and Y.-H. Oh, "Learning statistically efficient features for speaker recognition," in *Proceedings ICASSP*, 2001.

[5] P. Ding and L. Zhang, "Speaker recognition using principal component analysis," in *Proceedings ICONIP*, 2001.

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

[7] J.R. Hurford, M. Studert-Kennedy, and C. Knight, Eds., *Synthesyzing the origins of language and meaning using co-evolution, self-organization and level formation*, pp. 384–404, Cambridge University Press, 1998.

[8] J. J. Wolf, "Efficient acoustic parameters for speaker recognition," *J. Acoust. Soc. Am.*, vol. 51, no. 6, pp. 2044–2056, 1972.

[9] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley, 2001.

[10] D. O'Shaughnessy, *Speech Communications: Human and Machine*, IEEE Press, 2000.

[11] A.J. Bell and T.J. Sejnowski, "The independent components of natural scenes are edge filters," *Vision Research*, vol. 37(23), pp. 3327–3338, 1997.

[12] M. S. Bartlett, H. M. Lades, and T. Sejnowski, "Independent component representations for face recognition," in *Proceedings of the SPIE Symposium on Electronic Imaging*, 1998.

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

[14] S. Haykin, *Adaptive Filter Theory (3rd Ed.)*, Prentice-Hall, 1996.

[15] J-F. Cardoso and A. Souloumiac, "Blind beamforming for non Gaussian signals," *IEE Proceedings-F*, vol. 140, no. 6, pp. 362–370, Dec. 1993.