INDEPENDENT COMPONENT ANALYSIS FOR NOISY SPEECH RECOGNITION

Hsin-Lung Hsieh[1], *Jen-Tzung Chien*[1], *Koichi Shinoda*[2] *and Sadaoki Furui*[2]

[1] National Cheng Kung University, Tainan, Taiwan
[2] Tokyo Institute of Technology, Tokyo, Japan
{hlhsieh,chien}@chien.csie.ncku.edu.tw {shinoda,furui}@cs.titech.ac.jp

ABSTRACT

Independent component analysis (ICA) is not only popular for blind source separation but also for *unsupervised learning* when the observations can be decomposed into some independent components. These components represent the specific speaker, gender, accent, noise or environment, and act as the basis functions to span the vector space of the human voices in different conditions. Different from eigenvoices built by principal component analysis, the proposed *independent voices* are estimated by ICA algorithm, and are applied for efficient coding of an adapted acoustic model. Since the information redundancy is significantly reduced in independent voices, we effectively calculate a coordinate vector in independent voice space, and estimate the hidden Markov models (HMMs) for speech recognition. In the experiments, we build independent voices from HMMs under different noise conditions, and find that these voices attain larger redundancy reduction than eigenvoices. The noise adaptive HMMs generated by independent voices achieve better recognition performance than those by eigenvoices.

Index Terms**—** Independent component analysis, speech recognition, environment modeling

1. INTRODUCTION

ICA [3] was developed for blind source separation (BSS) where a set of observation data is seen and the underlying source information is unseen. BSS aims to identify the source signals or the mixing weights so as to separate the information sources in signal domain, feature domain or model domain [1]. There are two assumptions in developing ICA algorithm; the source signals should be mutually independent, and non-Gaussian distributed. In ICA model, an $M \times 1$ observation vector $\mathbf{x}$ is produced from M statistically independent sources $\mathbf{s}$ by $\mathbf{x} = A\mathbf{s}$ where A is a $M \times M$ mixing matrix. We are engaged in an inverse problem by finding a demixing matrix W and recovering the source signals by $\mathbf{y} = W\mathbf{x}$. In [2], a mutual information measure was proposed for ICA transformation, and so the inter-cluster data in transformed space were independent and intra-cluster data were dependent. Such an unsupervised learning framework was applied to establish multiple hidden Markov models (HMMs) for compensating the pronunciation variations in speech recognition [1][2].

HMM is a popular paradigm for automatic speech recognition. To deal with the mismatch between training and test data, HMM parameters should be adapted to fit the unknown test environments. The adaptive HMMs achieve the robustness in speech recognition. However, an important issue in speaker or environment adaptation is to find the solution to rapid adaptation from very limited adaptation data. Kuhn *et al*. [8] presented the rapid adaptation in eigenvoice space, which was spanned by the eigenvoice basis vectors. These eigenvoices were estimated by principal component analysis (PCA) from a set of reference models. Since the eigenvoices are orthogonal and representative as the most important components of speaker variations, we can use sparse adaptation data to estimate the coordinate coefficients and find the adapted acoustic model in eigenvoice space. In [10], the eigenvoices were extended to the kernel eigenvoices by performing a nonlinear PCA using the composite kernels.

Although PCA is useful for dimension reduction and the reduction of information redundancy, we challenge that the reduction performance is not as good as ICA. Different from PCA extracting the *orthogonal* components, ICA identifies the *independent* components, which are sufficient to refer as the *basis vectors* to form the minimal spanning set. Accordingly, we present an ICA approach to extract the *independent voices* and construct an independent space where the environmental variations can be effectively represented. With this space, we perform the *sparse coding* to encode the noise adaptive HMMs for speech recognition. In [7], the sparse coding via ICA was developed to extract image features which were rarely and significantly active. In [9], the sparse coding was used to find the statistical structure of male and female speech signals. In this study, we focus on identifying the significantly active components of HMMs using ICA. These components represent the acoustic models in different noise types and signal-to-noise ratios (SNRs). The HMM parameters are adapted by using the *maximum likelihood independent decomposition*. Experimental results on Aurora2 database show that the proposed independent voices achieve higher information redundancy reduction and speech recognition performance in comparison with eigenvoices under different noise conditions and numbers of components and adaptation sentences.

2. SURVEY OF RELATED WORKS

First of all, we address a standard ICA method and explain why ICA is feasible for sparse coding.

2.1 Independent component analysis

ICA is an extension of PCA and is specialized in finding the underlying factors or sources, which are as independent as possible. ICA algorithm is designed to identify the independent sources from the mixed signals. The sources are non-Gaussian distributed with higher-order statistics, and so a demixing matrix W can be estimated by optimizing an objective function measuring the independence or non-Gaussianity. There are many ICA objective functions proposed in the literature [1][2][3][6]. The *negentropy* was one of the most popular metrics, which was used to measure the non-Gaussianity and adopted to develop the Fast ICA algorithm [6]. This algorithm builds the negentropy-based contrast function and performs the optimization by

978-1-4244-2354-5/09/$25.00 ©2009 IEEE

$$\hat{W} = \arg\max_{W} \{E[G(W\mathbf{x})] - E[G(\mathbf{v})]\}^2 \quad (1)$$

under the constraint of decorrelation in individual components. In (1), $G(\cdot)$ is a nonquadratic function and $\mathbf{v}$ is a standardized Gaussian vector. There are two useful nonquadratic functions in building contrast function, and given by [6]

$$G_1(s) = \frac{1}{b_1} \log\cosh(b_1 s) \quad (2)$$

$$G_2(s) = -\frac{1}{b_2} \exp(-b_2 s^2 / 2) \quad (3)$$

where $1 \le b_1 \le 2$, $b_2 \approx 1$ are constants. The demixing matrix $\hat{W}$ is accordingly estimated to identify the independent components. This study applies the Fast ICA algorithm [6] to build independent voice space for speaker and environment adaptation.

2.2 Sparse coding

Interestingly, ICA methods are feasible to fulfilling *sparse coding* which is important for feature extraction, data compression, and some other applications. The relation between source coding and statistical density estimation is illustrated by the Shannon's information theorem [4]

$$\begin{aligned} E[l(s)] &\ge \sum_s p(s) \log \frac{1}{q(s)} \\ &= \sum_s p(s) \log \frac{1}{p(s)} + \sum_s p(s) \log \frac{p(s)}{q(s)}. \end{aligned} \quad (4)$$

Here, $p(s)$ and $q(s)$ denote the true density and the approximate density of a source signal *s*, respectively. The lower bound of an expected code length $E[l(s)]$ depends on the entropy of source signal and the Kullback-Leibler (KL) divergence between $p(s)$ and $q(s)$. True density $p(s)$ is unknown. If $q(s)$ equals to $p(s)$, i.e. $\mathrm{KL}(p(s), q(s)) = 0$, the expected code length only relies on the entropy of the data. As we know, Gaussian variable has the maximum entropy among all distributions with a given mean and variance. If a source signal is distributed by a super-Gaussian density following the assumption of ICA, this signal has shorter code length than a Gaussian signal. However, ICA is comparable to transform the mixed signal by some basis functions so that the transformed signal attains the largest non-Gaussianity and can be encoded by the shortest length. This property is crucial to illustrate the capability of high information packing by using ICA transformation.

Accordingly, the non-negative sparse coding [5] was presented to extract the basis vectors from non-negative observed signal $X = \{\mathbf{x}_n\}_{n=1}^N$ with only a few nonzero source signals $S = \{\mathbf{s}_n\}_{n=1}^N$. The objective function is expressed by

$$(\hat{W}, \hat{S}) = \arg\min_{(W,S)} \frac{1}{2} \left\| X - W^{-1} S \right\|^2 + \eta \sum_{n=1}^{N} \sum_{m=1}^{M} \psi(s_{nm}) \quad (5)$$

where ψ is a sparseness measure and $\eta \ge 0$ is a tuning parameter. This method aims to minimize the reconstruction error with the sparse distribution for coefficients and under the constraints of unit column vectors in W^{-1} and positive entries in W^{-1} and *S*. Using this scheme, only a few coefficients affect the estimated basis functions. Sparseness of source signals is proportional to the information conveyed in the entries of basis vectors. This property of sparse coding was applied to extract image features in [7] and was related to Rissanen's minimum description length (MDL) algorithm [11]. In speech recognition, we would like to perform sparse coding and build a set of basis functions, which represents the variations of noisy environments and encodes the acoustic model as efficient as possible.

3. INDEPENDENT VOICE SPACE

This works focuses on building a voice space from a set of reference models containing some redundant information. The information redundancy can be reduced to simplify the model uncertainty as well as to decrease the number of components in a factor model [13]. In what follows, we address why ICA minimizes information redundancy for for environment adaptation.

3.1 PCA versus ICA

PCA and ICA are useful to reduce information redundancy. In contrast to PCA extracting principal components, ICA identifies the independent components. As pointed in [9], the degree of sparseness in distribution of the recovered signals is proportional to the amount of information conveyed by the transformation or its basis vectors. Typically, sparse distribution has sharp peak and heavy tails. With the sparse distribution, the recovered signals are clearly clustered. In general, the sparseness in independent components is more significant than that in principal components. The reason is that ICA extracts higher-order statistics while PCA performs a linear de-correlation process. Using PCA, the extracted *M* principal components $\{e_1, e_2, \cdots, e_M\}$ have zero mean and obey the uncorrelation property

$$E[e_1 e_2 \cdots e_M] = E[e_1]\, E[e_2] \cdots E[e_M]\,. \quad (6)$$

The first moment of joint distribution $p(e_1, \cdots, e_M)$ equals to the product of the first moments of individual marginal distributions $\{p(e_m)\}_{m=1}^M$. However, the extracted *M* independent components $\{s_1, s_2, \cdots, s_M\}$ in ICA have zero mean and satisfy

$$E[s_1^r s_2^r \cdots s_M^r] = E[s_1^r]\, E[s_2^r] \cdots E[s_M^r] \quad (7)$$

for any integer r. This property holds at different moments. The higher-order correlations in independent components are zero. Independent components are uncorrelated, but the principal components are not sufficiently independent. The extracted independent components give us the means of exploiting information embedded in higher-order statistics of observed data.

Owing to this advantage, we apply ICA method to build up independent voices holding the higher-order uncorrelation. Using these significantly active voices, we estimate a composite model for speech recognition. The composite model using ICA is superior to that using PCA. In this study, we conduct the performance evaluation of different components and their numbers by the metric of minimum description length (MDL) [11] or Bayesian information criterion (BIC) [12]. MDL and BIC are initialized from different aspects, but come up with the same formula which is popular for model selection. In the experiments, we compare BIC values of eigenvoices and independent voices, and investigate the performance of two sets of components in building the adapted acoustic model. A better model shall either attain larger Bayesian information for *model regularization* or produce smaller description length for sparse coding of unknown environment, and so the unexpected variations shall less likely happen. Let *K* denote the number of selected components and λ denote the adopted model: PCA or ICA. We evaluate BIC value for different λ and *K*

$$\mathrm{BIC}\,(\lambda,K)=\log p(X\mid\lambda,K)-\frac{1}{2}\xi\cdot K\cdot\log N \tag{8}$$

where ξ is a control parameter [13].

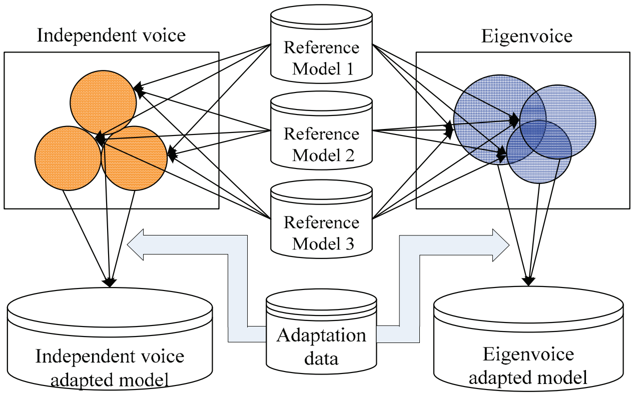

Figure 1 Adaptation data flow using eigenvoices and independent voices

3.2 Adaptation in independent voice space

We construct the independent voice space for representing the acoustic models covering different speakers and noise conditions. The information redundancy in reference models are reduced. With these independent voices, we calibrate a new acoustic model by using sparse adaptation data. Figure 1 shows the adaptation using eigenvoices and independent voices. In the procedure, we gather a set of reference models or HMMs which were trained by stereo speech data in different noisy conditions. Similar to speaker adaptation in eigenvoice space, we first setup N dimensional supervector consisting of Gaussian mean vectors of different words, HMM states and mixture components and form the matrix X using the aligned supervectors from a set of M reference HMMs. In the next step, we reduce the redundancy information in the supervectors or $M\times N$ matrix X by performing the Fast ICA algorithm [6]. An $M\times M$ demixing matrix W is calculated and $A=\{a_{mn}\}=W^{-1}$ is obtained. In [8], the eigenvectors with the largest K eigenvalues are selected as the eigenvoices for building eigenvoice space. Here, we sort the column vectors in A and pick up the independent components according to the *component importance measure* (CIM) [13]

$$\mathrm{CIM}(n)=\frac{1}{M}\sum_{m=1}^{M}|a_{mn}|\,. \tag{9}$$

The independent voices are obtained by selecting demixing vectors $\mathbf{w}_n$ with the largest K CIM values. These vectors act as the basis vectors or minimal spanning vectors to build the independent voice space for noisy speech recognition. As illustrated in Figure 1, the independent voices are salient and feasible to represent the noise variations with larger redundancy reduction compared with the eigenvoices. When performing speaker and environment adaptation, we use a small set of enrollment data and estimate the coordinate vector of the adapted HMM mean vectors in independent voice space. Similar to the maximum likelihood eigen-decomposition (MLED) in eigenvoice method [8], we perform the *maximum likelihood independent decomposition* where the coordinate coefficients are estimated by maximizing the likelihood of adaptation data given the independent voices. Detailed formula of MLED can be found in [8].

4. EXPERIMENTS

4.1 Experimental setup

In the experiments, we conducted the evaluation of different methods by using Aurora2 speech database. The eigenvoices and independent voices were implemented. The number of extracted components K was changed in the evaluation. In addition to speech recognition, we calculate the kurtosis and BIC to evaluate the information redundancy reduction and the goodness of selected components, respectively. In speech recognition, the multi-conditional training set with four noise materials (subway, babble, car and exhibition hall), four SNR levels (5, 10, 15, 20 dB) and clean data was prepared. There were 34 gender-dependent sets of HMM parameters and one set of multi-conditional HMM parameters trained to act as the reference models. Totally, 35 supervectors ($M=35$) were generated from HMM mean vectors with aligned Gaussian mixture components. PCA and ICA were performed to extract eigenvoices and independent voices, respectively. In the connected digit recognition, each digit was characterized by 16 states and each state was characterized by three mixture components. The silence model was characterized by three states and the short pause was characterized by one state. Each state was characterized by 6 components. All training data were characterized by 39 Mel-frequency cepstral coefficients (MFCCs), which contained 13 MFCCs and their first and second derivatives. The test set A in Aurora2 was used to evaluate the recognition performance. The multi-conditional training method with no adaptation was referred as the baseline system.

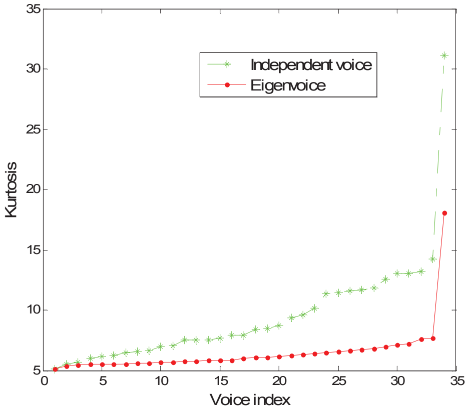

Figure 2 Comparison of information redunancy reduction

4.2 Evaluation for information redundancy reduction

The sparseness of transformed signals reflects the amount of information which is conveyed by the coefficients in basis functions. The fourth-order statistics, called *kurtosis*, is used to measure the sparseness of the sources, or equivalently the significance of information redundancy reduction in PCA and ICA transformation. The kurtosis of a zero-mean signal is given by $\mathrm{kurt}(s)=E[s^4]/E^2[s^2]-3$. Here, we calculate kurtosis for each of 34 supervectors of the transformed HMM mean vectors from different noise environments. Figure 2 displays the kurtosis of 34 sorted eigenvoices and independent voices. The kurtosis values of independent voices are consistently higher than those of eigenvoices. The mean of kurtosis of independent voices is 9.54 and that of eigenvoices is 6.49. These results imply that information redundancy reduction using independent voice approach is more significant than that using eigenvoice approach. The main reason is due to the capability of exploring higher-order statistical structure in using independent voices. Thus, the kurtosis is increased and the distribution becomes peaky.

4.3 Evaluation for noisy speech recognition

In noisy speech recognition, we performed noise adaptation by using 5, 10, 15 adaptation sentences collected in different noise environments. The averaged frame number (N) in a sentence was 176. The word error rates (WERs) in cases of no adaptation and adaptation using eigenvoices and independent voices are shown in Figure 3. The number of basis vectors K was set as 10 and 15 for comparison. The WERs in clean and noisy conditions with different noise types and SNRs were averaged. The averaged WERs were significantly improved by eigenvoice and independent voice methods. The WERs of using independent voices were further decreased. In cases of $K = 10$ and $K = 15$, the averaged WER reduction over different number of adaptation sentences (L=5, 10, 15) is 5.54% and 5.86% by using independent voices relative to eigenvoices, respectively. The WERs were consistently reduced by increasing the number of adaptation sentences. Figure 4 shows the BIC per sentence averaged over clean and different noise conditions. The BIC in (8) was calculated by using the likelihood of adaptation sentences and the regularization term determined by the number of components for the case of $\xi = 0.008$ and $\text{SNR} = 10\,\text{dB}$. This value was increased with large K but was saturated between $K = 10$ and $K = 15$. The independent voices consistently attained higher BIC than eigenvoices under different K. This result indicates that the independent voice model is selected as a better model than eigenvoice model when model regularization is concerned.

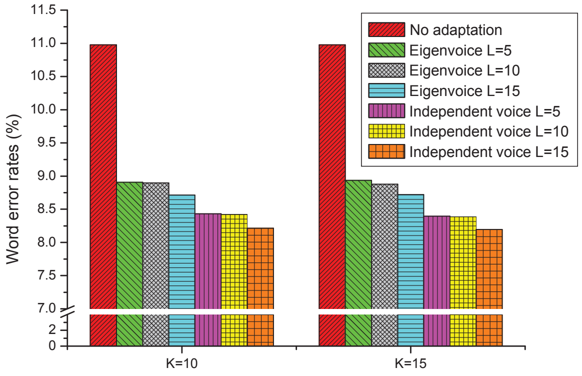

Figure 3 Comparison of word error rates

5. CONCLUSION

This paper presented the construction of independent voice space using ICA method and estimated the composite model for adaptive speech recognition. These independent voices were illustrated to perform sparse coding to reduce the information redundancy or model uncertainty in generation of speech HMMs in different noise conditions. High-order statistical structure was explored in using independent voices. The maximum likelihood decomposition was implemented to estimate a composite model for noise adaptation. Experiments showed that the independent voices reduced larger information redundancy and achieved better model description than the eigenvoices for different number of components. Independent voices obtained lower word error rates than eigenvoices for speech recognition in different noise environments using different number of adaptation sentences.

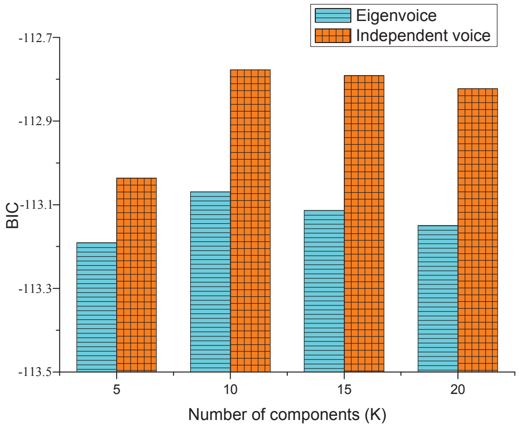

Figure 4 Comparison of BIC values

6. REFERENCES

[1] J.-T. Chien and B.-C. Chen, “A new independent component analysis for speech recognition and separation”, *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1245-1254, 2006.

[2] J.-T. Chien, H.-L. Hsieh and S. Furui, “A new mutual information measure for independent component analysis”, in *Proc. of ICASSP*, pp. 1817-1820, 2008.

[3] P. Comon, “Independent component analysis, a new concept?”, *Signal Processing*, vol. 36, pp. 287-314, 1994.

[4] T.-M. Cover and J.-A. Thomas, *Elements of Information Theory*, John Wiley, 1991.

[5] P.-O. Hoyer, “Non-negative sparse coding”, *Proc. IEEE workshop on Neural Networks for Signal Processing*, pp. 557-565, 2002.

[6] A. Hyvarinen, “Fast and robust fixed-point algorithms for independent component analysis”, *IEEE Trans. Neural Networks*, vol. 10, no. 3, pp. 626-634, 1999.

[7] A. Hyvarinen, E. Oja, P. O. Hoyer and J. Hurri, “Image feature extraction by sparse coding and independent component analysis”, in *Proc. of ICPR*, pp. 1268–1273, 1998.

[8] R. Kuhn, J.-C. Junqua, P. Nguyen and N. Niedzielski, “Rapid speaker adaptation in eigenvoice space”, *IEEE Trans. Speech and Audio Processing*, vol. 8, no. 6, pp. 695-707, 2000.

[9] T.-W. Lee and G.-J. Jang, “The statistical structures of male and female speech signals”, in *Proc. of ICASSP*, vol. 1, pp. 105-108, 2001.

[10] B. Mak, J.-T. Kwok and S. Ho, “Kernel eigenvoice speaker adaptation”, *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 5, pp.984-992, 2005.

[11] J. Rissanen, “Modeling by shortest data description”, *Automatica*, vol. 14, pp. 465-471, 1978.

[12] G. Schwarz, “Estimating the dimension of a model”, *Annals of Statistics*, vol. 6, no. 2, pp.461-464, 1978.

[13] M. Xu and M.-W. Golay, “Data-guided model combination by decomposition and aggregation”, *Machine Learning*, vol. 63, pp.43-67, 2006.