Evaluation of a Feature Selection Scheme on ICA-based Filter-Bank for Speech Recognition

Neda Faraji & S.M. Ahadi

Department of Electrical Engineering,
Amirkabir University of Technology, Tehran, Iran
nfaraji@cic.aut.ac.ir sma@aut.ac.ir

Abstract

In this paper, we propose a new feature selection scheme that can contribute to an ICA-based feature extraction block for speech recognition. The initial set of speech basis functions obtained in independent component analysis training phase, has some redundancies. Thus, finding a minimal-size optimal subset of these basis functions is rather vital. On the contrary to the previous works that used reordering methods on all the frequency bands, we have introduced an algorithm that finds optimal basis functions in each discriminative frequency band. This leads to an appropriate coverage of various frequency components and easy extension to other data is also provided. Our experiments show that the proposed method is very useful, specifically in larger vocabulary size tasks, where the selected basis functions trained using a limited dataset, may get localized in certain frequency bands and not appropriately generalized to residual dataset. The proposed algorithm surmounts this problem by a local reordering method in which contribution of a basis function is specified with three factors: class separability power, energy and central frequency. The experiments on a Persian continuous speech corpus indicated that the proposed method has led to 17% improvement in noisy condition recognition rate in comparison to a conventional MFCC-based system.

1. Introduction

A fundamental problem in applied digital signal processing is to find suitable representations for image, audio or other kind of data for applications such as recognition and denoising. Data representations are often based on linear transformations. Standard linear transformations widely used in signal processing are the Fourier, Haar, cosine transform, etc. It would be most useful to estimate the linear transformation from the data itself, in which case the transform could be ideally adapted to the kind of data that is being processed [1]. Independent Component Analysis (ICA) is a data-driven method that can capture the higher order statistics from the signal. ICA can separate independent components from the signals which are mixtures of the unknown sources and it can be applied to image, speech and medical signal processing [2].

In [3], ICA was used for feature extraction of speech signal. The extracted ICA-based features were then applied to an Isolated-word recognition task. Since the ICA algorithm finds the independent components corresponding to the dimensionality of the input, it may result in redundant components. However, the extracted independent components of speech correspond to the sources of speech production. Some of these sources are irrelevant sources. In reality, these may not be useful for speech recognition and should be removed. Identifying the irrelevant and redundant sources and their removal could be carried out by a feature selection method. In [3], two measures have been used for reordering and selecting of basis vectors:

1) The energy of the basis vector.
2) The variance of the basis vector coefficient.

Although the basis vectors selected with these two simple methods outperformed mel-scale filter-bank in capturing the higher order structures of speech, our experiments have shown that two mentioned methods lead to high degradations of recognition rate in some conditions. In reality, the overall ICA-based feature extraction performance is very sensitive to applied feature selection method. We should, therefore, use an effective feature selection method that considers all aspects of a given problem, here speech recognition.

We propose a new feature selection algorithm that minimizes the available problems in two aforementioned methods by obtaining a nearly optimal subset of the initial ICA-based filters with respect to adaptation and generalization issues. The considerable effect of proposed method in improving recognition rate has been approved by our experiments on a Persian continuous speech corpus. Also, the recognition results obtained on both Aurora 2 and Persian tasks show that the ICA-based features are robust in noisy conditions. It should be noted that in the case of Aurora 2 task, the global reordering method was adequate and only the effect of initial number of filters is evaluated.

The paper is organized as follows. In section 2 we briefly review the use of ICA technique in feature extraction of speech. The proposed feature selection method is presented in section 3. Experimental results are discussed in section 4 and conclusions are drawn in section 5.

2. Extracting speech features using ICA

2.1. Extracting basis vectors of speech

The linear model of independent component analysis assumes that the observation is a linear mixture of the independent components and is represented as:

$$\mathbf{x} = \mathbf{A}\mathbf{s} = \sum_{i=1}^{N} \mathbf{a}_i s_i \ , \ \ \mathbf{s} = \mathbf{W}\mathbf{x} \qquad (1)$$

In this model, $\mathbf{a}_i$ is a basis vector that contributes to the generation of the observed data with source coefficient s_i. The estimation of sources and basis functions could be done by maximization of negentropy $J(s_i)$ [1]:

$$max\, J(s_i) = J(\mathbf{w}_i^T \mathbf{x}) = \left[E\{G(\mathbf{w}_i^T \mathbf{x})\} - E\{G(v)\} \right]^2 \qquad (2)$$

where $\mathbf{w}_i$ is the ith row of mixing matrix $\mathbf{W}$, G is a contrast function and v is a standardized Gaussian variable. In this work, we take the exponential function as the contrast function:

$$G(u) = -\exp(-\frac{1}{2}u^2) \qquad (3)$$

To find basis vectors of speech, short-time segments from speech signals are constructed and using an iterative algorithm, negentropy is maximized. We take a fixed point version of the algorithm and an iterative symmetric decorrelation scheme. Finally, the columns of **A** matrix are obtained from the inverse of the estimated **W**. These columns are considered as the initial set of ICA-derived filter-bank.

2.2. Selection of dominant basis vectors

The dimension of the extracted basis vectors is equal to the dimension of the short-time segments[3]. It is desirable to select dominant basis vectors from the initial set so that the feature extraction method is computationally comparable with the conventional one. Also, irrelevant and redundant basis functions are eliminated. We could decide based on the L2-norm of the basis vector or the variance of the basis vector coefficient [3]. The block diagram of ICA-based speech recognition system has been shown in Fig. 1.

3. Proposed feature selection algorithm

The feature selection methods utilized in [3], may omit filters in some frequency bands. We face this effect in certain conditions. For example, most of high energy basis vectors are localized in low frequency bands. Thus, the feature selection method may suppress the high frequency basis vectors. However, when the dimensionality of the training speech frames in ICA algorithm is increased, global reordering methods may result in almost low frequency dominant basis functions. Also, the ICA-derived filters are highly adapted to its normally limited training dataset. It emphasizes on some frequency bands available in ICA training dataset. Thus, the feature selection algorithm may remove less significant filters. These removed filters may have very important role in extracting information from the non-training data. In practice, there should be a trade-off between adaptation and generalization. The flow diagram of the algorithm has been demonstrated in Fig. 2.

3.1. Measure of comparing basis vectors

It is preferable to choose a measure that considers the classification problem more strongly. Thus, we have introduced a weighted version of L2-norm in which L2-norm of a basis vector is weighted with its Proportion of Variance (PoV).

$$\text{measure}(j) = \text{PoV}(j) \times \text{L2-norm}(j) \qquad (4)$$

where j indicates jth filter. PoV [4] is calculated using the ratio of between-class variance to within-class variance. The PoV measure of a given basis vector, shows its capability in separating different classes, and along with L2-norm directs us to choose those high energy basis vectors that also have high discrimination power. We assume that the number of classes is L. Then, the between-class and within-class variances for a given variable x, are defined as follows [5]:

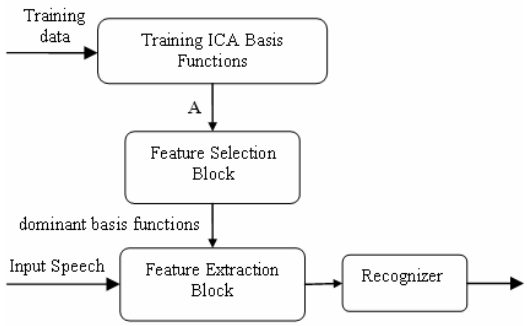

Figure 1: Block Diagram of ICA-based speech recognition system

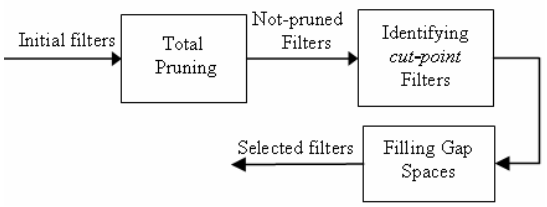

Figure 2 : Proposed Feature Selection Block.

$$S_W = \sum_{i=1}^{L} P_i E\left[(x-\mu_i)(x-\mu_i)^T \,|\, w_i \right] = \sum_{i=1}^{L} P_i \Sigma_i \qquad (5)$$

$$S_B = \sum_{i=1}^{L} P_i \left[(\mu-\mu_i)(\mu-\mu_i)^T \right], \quad \mu = \sum_{i=1}^{L} P_i \mu_i \qquad (6)$$

where μ_i, Σ_i and P_i are the mean, variance and the a priori probability of the ith class respectively. Then we write:

$$\text{PoV} = \frac{S_B}{S_W} . \qquad (7)$$

The aforementioned measure in Eq. 4 is very useful when we want to select dominant basis vectors from a set of basis vectors with the same central frequencies. It is probable that a low-power high-frequency basis vector has more capability to discriminate classes than a high-power low-frequency one. In other words, comparing two basis vectors without considering their central frequencies is not correct and finally may result in giving more weight to a certain band in comparison to the others.

This has been the main motivation behind our work on providing an algorithm to select dominant basis vectors automatically. In reality, the problem may be solved with local decisions instead of global ones.

The proposed algorithm includes three parts:

1) Total pruning
2) Identifying the *cut-point* filters
3) Filling gap spaces.

3.2. Total pruning

In this step, each filter competes with the filters whose central frequencies are near its. The competition is local and the difference between central frequencies should be below a threshold value. These filters constitute a group and one or more filters from this group with the largest PoV-weighted L2-norm are selected. It should be noted that a very small threshold value may lead to no filter pruning. Meanwhile, a very large threshold value destroys the locality assumption. We have selected a threshold value equal to 80 Hz. This value is about half of the minimum bandwidth in a conventional mel-scale filter-bank.

3.3. Identifying *cut-point* filters

The aim of this part of algorithm is identifying the critical filters. The critical filters have more distinct characteristics in comparison to the others and would be kept to improve generalization and classification.

The central frequency of a filter is an index that can separate it from the others. A hierarchical clustering algorithm is applied to the central frequency values to reach a certain number of central frequency clusters. The number of clusters is increased by one, if no cluster is found to have a single member. Single member clusters include the filters that have more distinct central frequencies than the other filters. These filters have an important role in generalization.

Concurrently, another clustering is performed on PoV-weighted L2-norm values to reach 3 clusters corresponding to low, medium and high values. In this case, the single member clusters have more distinct PoV-weighted L2-norm values. We are interested in keeping the filters with high PoV-weighted L2-norms. These filters can improve the classification. The filters found from two aforementioned clustering approaches are *cut-point* filters and can divide the whole frequency band into sub-bands.

After finding the initial *cut-point* filters, the two explained clusterings are performed on each sub-band filters. Sub-band clustering is performed from low to high frequency sub-bands. For the algorithm to be well conditioned, we have applied constraints on it. The constraints were put on:

1) The initial number of central frequency clusters in the beginning of the algorithm.
2) The maximum number of clusters in each sub-band.

The large number of clusters in the beginning of the algorithm lead to fast but not accurate converging. We begin the algorithm with the minimum number of central frequency clusters, so that it can find *cut-points* (single-member clusters) gradually.

In sub-band clustering, the low frequency sub-bands are prior to the high frequency sub-bands. If we do not limit the number of clusters in each sub-band, the algorithm finds the *cut-point*s only in low frequency bands. However, the maximum number of clusters in each sub-band specifies the final number of *cut-point* filters found.

3.4. Filling gap space

According to the values of constraints, there would remain some gaps between the *cut-point* filters found. We heuristically fill these gap spaces.

Table 1: The recognition results in single Gaussian HMM and Persian corpus. The number of filters in each sub-band has been selected heuristically.

		Recognition Rate
MFCC		43.57
ICA	**Global reordering**	32.22
	4 sub-band	33.92
	8 sub-band	39.11
	13 sub-band	40.08
	18 sub-band	43.68
	Proposed method	**44.52**

Table 2: The recognition rate in different noisy conditions and 15-Gaussian mixture HMM.

	Recognition Rate	
	ICA	**MFCC**
Clean	72.83	76.43
Babble	27.54	20.01
Car	70.96	50.32
F16	28.72	19.6
Average	**42.4**	**29.97**

Table 3: The recognition rate of ICA-based features in AURORA 2 task (12 Ceps.coeff+C_0 in ICA features, 12 Ceps.coeff+log Energy in baseline system)

	Recognition Rate		
	35	**50**	**baseline**
Set A	65.44	63.94	61.13
Set B	64.85	64.1	55.57
Set C	60.28	65.57	66.68
Average	**63.52**	**64.53**	**61.12**

4. Experiments and results

4.1. Experiments on Farsi continuous speech corpus

The Farsi continuous speech corpus, FARSDAT was used in this work [6]. This corpus is the only continuous speech corpus of Farsi, which is available to public. It consists of 6000 sentences from 300 speakers, each uttering 20 sentences selected from a set of 392 available sentences. We have used 1819 sentences from 91 speakers for building a 3-state phoneme-based HMM system. Also, the test set includes 888 sentences from 46 speakers. Training and test have been carried out using HTK [7] and with different number of Gaussian mixtures in each state. The noise was then added to the speech in different SNRs. The noise data was extracted from the NATO RSG-10 corpus [8]. We have considered babble, car and F16 noises and added them to the clean signal at 20, 15, 10, 5 and 0 dB SNRs. Our experiments were carried out using MFCC (for comparison purposes) and ICA features. The features in two cases were computed using 25 msec. frames with 10 msec. of frame shifts. Pre-emphasis coefficient was set to 0.97 and a Hamming window was applied. The feature vectors for the two methods were composed of 12 cepstral and a log-energy parameter.

For extracting MFCC features, a 24-channel mel-scale filter-bank was used. Also, for training the initial set of ICA-based filter-bank, the 100 sample short-time segments with frame shifts of 30 samples were used. These segments were

constructed using 110 sentences from 11 speakers. After training, the proposed feature selection block selected 24 dominant filters from a total of 100 filters. The selected filters were then replaced mel-scale filters in feature extraction block.

In step 1 of feature selection algorithm, threshold value was set to 80Hz and 50 filters were pruned. From the 50 remaining filters, 20 *cut-point* filters were selected in step 2. This final number of *cut-point* filters was obtained with these parameters:

- The initial number of central frequency clusters was set to 4 at the beginning of the algorithm.
- The maximum number of clusters in each sub-band was set to the number of sub-band members minus one.
- The initial number of clusters in each sub-band clustering was set to the minimum value between 4 and half of the sub-band members.

Finally, 4 filters filled gap spaces to get 24 filters. In all steps, the comparison of filters was carried out by PoV-weighted L2-norm measure. Also, the PoV values of 100 initial filters were calculated using 119 phoneme-labeled sentences.

Table 1 lists the recognition results of ICA-based features using sub-band reordering methods, in which the whole frequency band was divided into some sub-bands and in each sub-band, the L2-norm measure was used for selecting dominant filters. The number of filters in each sub-band is selected according to the number of mel-scaled filters in that sub-band. The results have been brought from clean condition and 1-mixture phoneme-based HMM recognizer. Also, the recognition result of MFCC features is given for comparison. The indicated results show the considerable effect of used feature selection method in overall result. As seen in Table. 1, proposed feature selection algorithm has increased the recognition rate about 1.6% in comparison to MFCC.

Table 2 lists the recognition results obtained in clean and noisy conditions, on a 15-component HMM-based system using the proposed feature selection method. As indicated in this table, the performance of ICA-based features is worse than MFCC in clean condition when we used 15-Gaussian HMM system. This effect is the natural result of lower-variance ICA features. The results also show the robustness of ICA features in noisy conditions. The average values mentioned in this table are calculated over the results obtained from 0 dB to 20 dB SNRs , omitting the clean ones.

The results obtained can be listed as follows:

1) The feature selection method is very important in ICA-based feature extraction.
2) The ICA-based features improve the robustness of speech recognition system.
3) When the number of Gaussian components is increased, the recognition rate of ICA-based features is degraded in comparison to MFCC features.

4.2. Experiments on AURORA 2 speech task

In this section, we briefly review the result of experiments on AURORA 2 task [9], adopting global reordering method and various initial number of ICA-based filters. It should be noted that ICA training phase was carried out using 22 files from 22 selected speakers in this task.

The results are reported in Table 3 and the following conclusions can be listed:

1) The global reordering is adequate in this smaller size task.
2) The ICA-based features are more robust than MFCC features.
3) The initial number of ICA-derived filters is effective in obtained recognition rate.

5. Conclusions

In this paper, we tried to emphasize on the significant role of the feature selection algorithm on the overall performance of ICA-based speech recognition systems. The ICA-derived filters are adapted to processing data, specifically with data used in training ICA basis functions. Therefore, they are able to extract maximum higher-order information from data. The insufficient feature selection algorithm can remove some of the essential filters and lead to degradation of recognition rate. However, if an appropriate algorithm is used for selecting dominant filters in ICA feature extraction block, the recognition rate in noisy conditions shows great improvements in comparison to the baseline. This effect also originates from adapting of ICA filters with data. It seems that the ICA-based features contain the maximum amount of data information, while extract minimum information of noise data. The initial number of filters is also an important parameter in obtaining filters with appropriate frequency resolutions and then better recognition rate.

6. Acknowledgement

This work was in part supported by a grant from the Iran Telecommunications Research Center (ITRC).

7. References

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

[2] M. Kotani, Y. Shirata, S. Maekawa, S. Ozawa, K. Akazawa, "Application of independent component analysis to feature extraction of speech," in *Proc. IJCNN, Vol. 5, pp. 2981-2985, 1999.*

[3] J.H Lee, H.Y Jung, T.W Lee, S.Y Lee "Speech feature extraction using independent component analysis," in *Proc. ICASSP .Vol. 3, pp. 1631-1634, 2000.*

[4] E.K. Ekenel, N.Sankur, "Feature selection in the independent component subspace for face recognition," in *Pattern Recognition Letters*, *Vol. 25, pp. 1377-1388, June 2004.*

[5] N. Malayath, H. Hermansky, "Data-driven spectral basis functions for automatic speech recognition," in *Speech Communication*, *Vol. 40, Issue 4, pp. 449-466, June 2003.*

[6] D.Gharavian, S.M.Ahadi, " Evaluation of the effect of stress on formants in Farsi vowels," in *Proc. ICASSP*, *vol. 1, pp. 661-664, Montreal, May 2004.*

[7] The hidden Markov model toolkit available from http://htk.eng.cam.ac.uk.

[8] Available from http://spib.rice.edu/spib/select_noise.html

[9] H.G. Hirsch, D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. ASRU*, *pp. 181-188, September 2000.*