Analysis of Speaker Variability

Chao Huang, Tao Chen, Stan Li, Eric Chang and Jianlai Zhou*

Microsoft Research, China
5F, Beijing Sigma Center, No. 49, Zhichun Road Haidian District
*Department of Automation, Tsinghua University
Beijing 100080, P.R.C
chaoh@microsoft.com

Abstract

Analysis and modeling of speaker variability, such as gender, accent, age, speech rate, and phones realizations, are important issues in speech recognition. It is known that existing feature representations describing speaker variations can be of very high dimension. In this paper, we introduce two powerful multivariate statistical analysis methods, namely, principal component analysis (PCA) and independent component analysis (ICA), as tools for analysis of such variability and extraction of low dimensional feature representation. Our findings are the following: (1) the first two principal components correspond to the gender and accent, respectively. The result that the second component corresponding to the accent has never been reported before, to the best of our knowledge. (2) It is shown that ICA based features yield better classification performance than PCA ones. Using 2-dimensional ICA representation, we achieved about 6.1% and 13.3% error rate in gender and accent classification, respectively, for 980 speakers.

1. Introduction

Speaker variability, such as gender, accent, age, speech rate, and phones realizations, is one of the main difficulties in speech signals. How they correlate each other and what the key factors are in speech realization are real concerns in speech research. As we know, performance of speaker-independent (SI) recognition systems is generally 2~3 times worse than that of speaker-dependent ones. As an alternative, different adaptation techniques, such as MAP and MLLR, have been used. The basic idea is to adjust the SI model and make it reflect intrinsic characteristics about specific speakers by re-training the system using appropriate corpora. Another method to deal with the speaker variability problem is to build multiple models of smaller variances, such as gender dependent model and accent dependent model, and then use a proper model selection scheme for the adaptation. SI system and speaker adaptation can be facilitated if the principal variances can be modeled and corresponding compensations can be made.

Another difficulty in speech recognition is the complexity of speech models. There can be a huge number of free parameters associated with a set of models. In other words, a representation of a speaker has to be high-dimensional when different phones are taken into account. How to analyze such data is a challenge.

Fortunately, several powerful tools, such as principal component analysis (PCA) [2] and more recently independent component analysis (ICA) [1], are available for high dimension multivariate statistical analysis. They have been applied widely and successfully in many research fields such as pattern recognition, learning and image analysis. Recent years have seen some applications in speech analysis [5] [6] [7].

PCA decorrelates second order moments corresponding to low frequency property and extracts orthogonal principal components of variations. ICA is a linear, not necessarily orthogonal, transform which makes unknown linear mixtures of multi-dimensional random variables as statistically independent as possible. It not only decorrelates the second order statistics but also reduces higher-order statistical dependencies. It extracts independent components even if their magnitudes are small whereas PCA extracts components having largest magnitudes. ICA representation seems to capture the essential structure of the data in many applications including feature extraction and signal separation

In this paper, we present a subspace analysis method for the analysis of speaker variability and for the extraction of low-dimensional speech features. The transformation matrix obtained by using maximum likelihood linear regression (MLLR) is adopted as the original representation of the speaker characteristics. Generally each speaker is a super-vector which includes different regression classes (65 classes at most), with each class being a vector. Important components in a low-dimensional space are extracted as the result of PCA or ICA. We find that the first two principal components clearly present the characteristics about the gender and accent, respectively. That the second component corresponds to accent has never been reported before, while it has been shown that the first component corresponds to gender [6] [7]. Furthermore, using ICA features can improve classification performance than using PCA ones. Using the ICA representation and a simple threshold method, we achieve gender classification accuracy of 93.9% and accent accuracy of 86.7% for a data set of 980 speakers.

The paper is organized as follow. In section 2, we will highlight the basic ideas of PCA and ICA and some related work. The original and efficient speaker representation will also be discussed here. Detailed experiments setups and result analysis will be given in section 3. Section 4 concluded with our findings and possible applications discussions.

2. Speaker Variance Investigations

2.1. Related work

PCA and ICA have been widely used in image processing, especially in face recognition, identification and tracing.

However, their application in speech field is comparatively rare. Like linear discriminant analysis (LDA), most speech researchers use PCA to extract or select the acoustic features. [5]. Kuhn et al. applied PCA at the level of speaker representation and proposed eigenvoices in analog to eigenfaces and further apply it in the rapid speaker adaptation [6]. Hu applied PCA to vowel classification. [7].

All above work are based on representing speakers with concatenate the mean feature vector of vowels [7] or put one line of all the means from the Gaussian model that specifically trained for a certain speaker [6]. We have adopted the speaker adaptation model, specifically; we use the transformation matrix and offset that are adapted from the speaker independent model to represent the speaker. Here, maximum likelihood linear regression (MLLR) [3] was used in our experiments.

In addition, all above work only use PCA to pursue the projection of speaker in low dimension space in order to classify the vowels or construct the speaker space efficiently. As we know, PCA uses only second-order statistics and emphasize the dimension reduction, while ICA depends on the high-order statistics other than second order. PCA is mainly aim to the Gaussian data and ICA aiming to the Non-Gaussian data. Therefore, based on PCA, we introduce ICA to analysis the variability of speaker further because we have no clear sense on the statistical characteristics of speaker variability initially.

2.2. Speaker representation

2.2.1. MLLR matrices vs. Gaussian models

As mentioned in section. 2.1, we have used the MLLR transformation matrix (including offset) to represent all the characteristics of a speaker, instead of using the means of the Gaussian models. The main advantage is such a representation provides a flexible means to control the model parameters according to the available adaptation corpora. The baseline system and setups can be found in [4]. To reflect the speaker in detail, we have tried to use multiple regression classes, at most 65 according to the phonetic structures of Mandarin.

2.2.2. Supporting regression classes selection

We have used two different strategies to remove undesirable effects brought about by different phones. The first is to use the matrices of all regression classes. However, this increases the number of parameters that have to be estimated and hence increases the burden on the requirements on the adaptation corpora. In the second strategy, we choose empirically several supporting regression classes among all. This leads to significant decrease in the number of parameters to be estimated; and when the regression classes are chosen properly, there is little sacrifice in accuracy; as will be shown in Tables 4 and 5 in Section 3. The benefit is mainly due to that a proper set of support regression classes are good representatives of speakers in the sense that they provide good discriminative feature for the classification between speakers. Furthermore, fewer classes mean lower degree of freedom and increase in the reliability of parameters.

2.2.3. Diagonal matrix vs. offsets

Both diagonal matrix and offset are considered when making the MLLR adaptation. We have experimented with three combinations to represent speakers in this level: only diagonal matrix (with tag d), only offset (with tag b) and both of them (with tag bd). The only offset item of MLLR transformation matrix achieved much better result in gender classification, as will be shown in Table 3.

2.2.4. Acoustic feature pruning

The state of art speech recognition systems often apply multiple order dynamic features, such as first-order difference and second-order one, in addition to the cepstrum and energy. However, the main purpose of doing so is to build the speaker independent system. Usually, the less speaker-dependent information is involved in the training process, the better the final result will be. In contrast to such a feature selection strategy, we choose to extract the speaker-dependent features and use them to effectively represent speaker variability. We have applied several pruning strategies in the acoustic features level. We have also integrated pitch related features into our feature streams. Therefore, there are the six feature pruning methods as summarized in Table 1.

Table 1: Different feature pruning methods (number in each cell mean the finally kept dimensions used to represent the speaker)

Dynamic features	0-order (static)	1 order	2 order
w/o pitch	13	26	33
w/ pitch	14	28	36

3. Experiments

3.1. Data corpora and SI model

The whole corpora contain 980 speakers, 200 utterances per speaker. They are from two accent areas in China, Beijing (EW) and Shanghai (SH). The gender and accent distributions are summarized in Table 2.

Table 2: Distribution of speakers in corpora

	Beijing	Shanghai
Female	250 (EW-f)	190 (SH-f)
Male	250 (EW-m)	290 (SH-m)

The speaker-independent model we used to extract the MLLR matrix is trained according to all corpora from EW. It is also gender–independent, unlike the baseline system.

3.2. Efficient speaker representation

Figure 1 show the component contribution and cumulative contribution of top N principal components on variances, where N=1, 2…156. The PCA algorithm used in these and the following experiments is based on the covariance matrix. The dynamic range for each dimension has been normalized for each sample. This way, covariance matrix becomes the same as the correlation matrix.

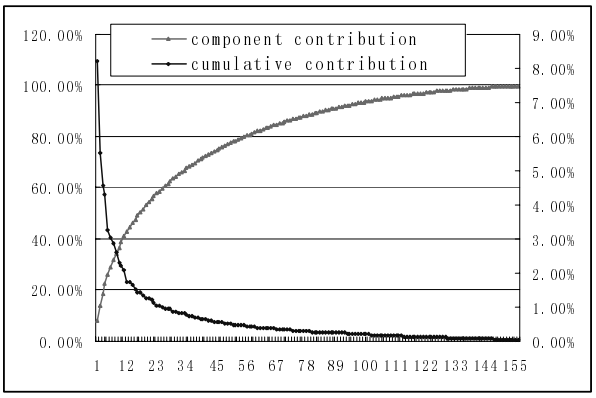

Figure 1: Single and cumulative variance contribution of top N components in PCA (horizontal axis means the eigenvalues order, left vertical axis mean cumulative contribution and right vertical axis mean single contribution for each eigenvalue)

To find the efficient and typical representation about speaker characteristics, we have applied strategies at several levels from supporting regression classes to acoustic features. Table 3 shows the gender classification results based on EW and SH corpora for various methods. Tags of -b,-d and –bd in the first column are according to the definition in section 2.2.3. Here the number of supporting regression classes is 6. From Table 3, we can conclude that the offset item in the MLLR matrix gives the best result.

Table 3: Gender classifications errors based on different speaker representation methods (The result is according to the projection of PCA, the total number for EW and SH are 500 and 480 respectively)

Dims	13	26	33	14	28	36
SH-b	22	14	24	22	20	30
SH-d	58	78	80	62	82	86
SH-bd	34	42	46	38	40	46
EW-b	52	38	66	52	56	78
EW-d	76	124	100	108	140	118
EW-bd	48	92	128	88	82	122

Furthermore, among all the acoustic feature combinations, the combination of the static features, first order of cepstrum and energy gives the best result for both EW and SH sets. It can be explained that these dimensions carry the most of the speaker specific information. However, it is very interesting to note that the addition of the pitch related dimensions leads to a slight decrease in the accuracy. It contradicts to the common conclusion that the pitch itself is the most significant feature of gender. This may be due to the following two reasons: First, pitch used here is in the model transformation level instead of the feature level. Secondly, multiple–order cepstrum feature dimensions have already included speaker gender information.

To evaluate the proposed strategy for the selection of supporting regression classes, we made the following experiments. There are a total of 65 classes. Here only the offset of MLLR transformation matrix and the 26 dimensions in feature stream are used according to the results demonstrated in Table 3. The selections of different regression classes are defined in Table 4, and the corresponding gender classification results are shown in Table 5.

Obviously, the combination of the 6 regression classes is a proper choice to balance the classification accuracy and the number of model parameters. Therefore, in the following experiments where the physical meaning of the top projections is investigated, we optimize the input speaker representation with the following setups:

- Supporting regression classes: 6 single vowels (/a/, /i/, /o/, /e/, /u/, /v/)
- Offset item in MLLR transformation matrix;
- 26 dimensions in acoustic feature level

As a result, a speaker is typically represented with a supervector of 6*1*26=156 dimension.

Table 4: Different supporting regression classes selection

# of regression classes	Descriptions
65	All classes
38	All classes of finals
27	All classes of initials
6	/a/, /i/, /o/, /e/, /u/, /v/
3	/a/, /i/, /u/
2	/a/, /i/
1	/a/

Table 5: Gender classifications errors of EW based on different supporting regression classes (The relative size of feature vector length is indicated as Parameters.)

Number of Regression Classes	65	38	27	6	3	2	1
Errors	32	36	56	38	98	150	140
Parameter	--	0.58	0.42	0.09	0.046	0.03	0.015

3.3. Speaker space and physical interpretations

The experiments here are performed with the mixed corpora sets of EW and SH. In this case, the PCA is performed with 980 samples of 156 dimensions each. Then, all speakers are projected into the top 6 components. A matrix of 980* 6 is obtained and is used as the input to ICA (The ICA is implemented according to the algorithm of FastICA proposed by Hyvarinen [1]). Figure 2 and Figure 3 show the projections of all the data onto the first two independent components. In the horizontal direction is the speaker index for the two sets. The alignment is: EW-f (1-250), SH-f (251-440), EW-m (441-690) and SH-m (691-980).

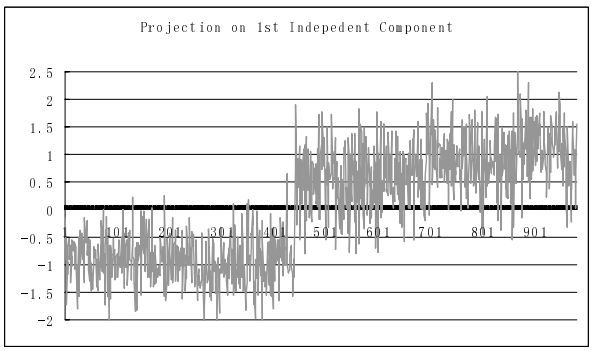

Figure 2: Projection of all speakers on the first independent component (The first block corresponds to the speaker sets of EW-f and SH-f, and the second block corresponds to the EW-m and SH-m)

From Figure 2, we can make a clear conclusion that the independent component corresponds to the gender characteristics of speaker. Projections on this component almost separate all speakers into two categories: male and female.

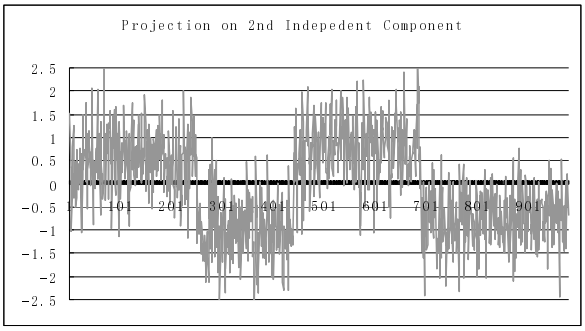

Figure 3: Projections of all speakers on the second independent component. (The four blocks correspond to the speaker sets of EW-f, SH-f, EW-m, SH-m from left to right)

According to Figure 3, four subsets occupy four blocks. The first and the third one together correspond the accent set EW (with Beijing accent) while the second and the fourth one together correspond to another accent set SH. They are separated in the vertical direction. It is obvious that this component has strong correlation with accents.

To illustrate the projection of the four different subsets onto the top two components, we draw each speaker with a point in Figure 4. The distribution spans a 2-d speaker space. It can be concluded that the gender and accent are the two main components that constitute the speaker space.

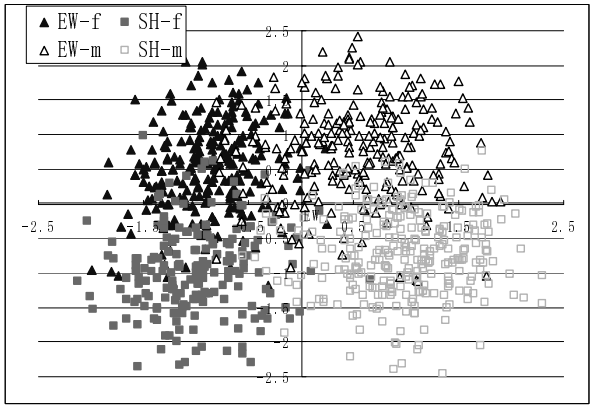

Figure 4: Projection of all speakers on the first and second independent components, horizontal direction is the projection on first independent component; vertical direction is projection on second independent component.

To illustrate accurately the performance of ICA, we compute the classification errors on the gender and accent classification through proper choice of projection threshold on each dimension shown in Figure 4. There are 60 and 130 errors for gender and accent, respectively. The corresponding error rates are 6.1% and 13.3%.

3.4. ICA vs. PCA

When applying PCA and ICA to gender classification on EW corpus, we received the error rate of 13.6% and 8.4% respectively. The results are achieved with the following setups to represent each speaker:

- 6 supporting regression classes;
- Diagonal matrix (-d)
- Static cepstrum and energy (13)

The similar results are achieved with other settings. It is shown that ICA based features yield better classification performance than PCA ones.

Unlike PCA where the components can be ranked according to the eigenvalues, ranking of the positions of the ICA components representing variations in gender and accent can not be done. However, we can always identify them in some way (e.g. from plots). Once they are determined, the projection matrix is fixed.

4. Conclusion

In this paper, we have investigated the variability between speakers through two powerful multivariate statistical analysis methods, PCA and ICA. It is found that strong correlations between gender and accent exist in two ICA components. While strong correlation between gender and the first PCA component is well known, we give the first physical interpretation for the second component: it is strongly related with accent.

We propose to do a proper selection of supporting regression classes, to obtain an efficient speaker representation. This is beneficial for speaker adaptation with limited corpus available. Through gender classification experiments combined with MLLR and PCA, we concluded that the static and first–order cepstrum and energy carry most information about speakers.

The features extracted by using PCA and ICA analysis can be directly applied to speaker clustering. Further work of its application in speech recognition is undergoing.

Acknowledgement
Thanks to Xiaoguang Lv in Microsoft Research China, for suggestive discussions about PCA and ICA.

5. References

[1] A. Hyvarinen and E. Oja, “Independent component analysis: algorithms and application” *Neural Networks* 13. pp.411~430, 2000.

[2] H. Hotellings, “Analysis of a complex of statistical variables into principle components”, *J. Educ. Psychol.*, 24, pp.417-441, 498-520, 1933.

[3] C. J. Leggetter, P. C. Woodland, “Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models”, pp. 171-185, *Computer Speech and Language*, Vol. 9, No. 2, April, 1995.

[4] E. Chang, J. L. Zhou, C. Huang, S. Di, K. F. Lee, “Large Vocabulary Mandarin Speech Recognition with Different Approaches in Modeling Tones”. *Proc. of ICSLP 2000*, Beijing, Oct. 2000.

[5] N. Malayath, H. Hermansky, and A. Kain , “Towards decomposing the sources of variability in speech”, *Eurospeech’ 97*, Vol. 1, pp. 497-500, Sept. 1997

[6] R. Kuhn, J. C. Junqua, P. Nguyen and N. Niedzielski, “Rapid Speaker Adaptation in Eigenvoice Space”, *IEEE Trans. on Speech and Audio Processing*, Vol. 8, No. 6, Nov. 2000.

[7] Z. H. Hu, “Understanding and adapting to speaker variability using correlation-based principal component analysis”, *Dissertation of OGI*. Oct. 10, 1999.