Phonetic Subspace Mixture Model for Speaker Diarization

I-Fan Chen[1], *Shih-Sian Cheng*[2], *Hsin-Min Wang*[1,2]

[1] Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan
[2] Institute of Information Science, Academia Sinica, Taipei, Taiwan
{ifanchen,sscheng,whm}@iis.sinica.edu.tw

Abstract

This paper presents an improved distance measure for speaker clustering in speaker diarization systems. The proposed phonetic subspace mixture (PSM) model introduces phonetic information to the ΔBIC distance measure. Therefore, the new PSM model-based ΔBIC distance measure can remove the effect of phonetic content on the diarization results. The typical ΔBIC distance measure can be seen as a special case of the new ΔBIC distance measure. Our experiment results show that the new distance measurement consistently improves the speaker diarization performance on three datasets.

Index Terms: BIC, phonetic information, speaker diarization

1. Introduction

The objective of speaker diarization is to group together speech segments produced by the same speaker in an audio stream [1]. The technique is a vital processing step for automatic audio transcription/indexing [1] and spoken document retrieval [2]. It has been studied in various data domains, e.g., conversational telephone speech [3], broadcast news data [4][5], and meeting data [6].

Speaker diarization systems usually comprise two core components: speaker segmentation, which cuts the audio stream into homogeneous segments; and speaker clustering, which groups the homogeneous segments into speaker clusters. Currently, leading systems usually apply agglomerative hierarchical clustering (AHC) to perform speaker clustering [4][5] after segmentation. When performing AHC for speaker clustering, each speech segment derived by speaker segmentation is considered a cluster initially; then, in each merging step, the two clusters with the smallest distance measure are merged into a new cluster. The two major tasks of AHC are 1) computing the inter-cluster distances; and 2) determining the number of clusters.

However, the popular distance measures used in AHC, such as ΔBIC [7], the cross-likelihood ratio [4] and the information crossing rate (ICR) [8], are usually derived with cepstral coefficients which have both phonetic and speaker characteristics. As a result, the obtained distance measures might reflect both the speaker variation and the phonetic variation in the speech content of two speech segments. In other words, performing speaker diarization without considering the background phonetic content could reduce the sensitivity of a diarization system to different speakers. For example, Fig. 1 shows the distribution of the acoustic features of five phones, /a/, /i/, /u/, /e/, and /o/, uttered by two speakers. From the figure, it is clear that, without considering the phones, the acoustic features of these two speakers' speech have a large overlap in the distribution; hence, it is difficult to separate these two speakers on the feature plane. However, there is hardly any overlap between the acoustic features of phones /u/ and /o/ of the two speakers, which means we can easily separate the two speakers using the features corresponding to the phonetic content /u/ and /o/. Therefore, it is expected that the speaker diarization performance would be improved to some extent if the phonetic information could be introduced to speaker diarization systems.

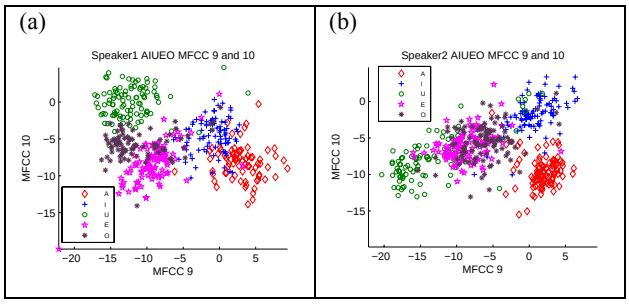

Figure 1: The distribution of the 9th and 10th elements of the MFCC acoustic features of five vowels (/a/, /i/, /u/, /e/, and /o/) uttered by two speakers.

This paper reports on our first attempt to provide today's speaker diarization paradigms with the ability to exploit the phonetic content information in speech. We focus on the speaker clustering component only, i.e., the speaker segmentation component does not consider phonetic content information currently. Because the ΔBIC distance measure has a mathematical closed form solution, which ensures that the evaluation is efficient and yields a good performance, it is one of the most popular distance measures used in today's speaker clustering implementations. In a typical AHC-based speaker clustering implementation using the ΔBIC distance measure, to determine whether two clusters A and B should be merged or not, two hypotheses are made: H_0 – the two clusters should be merged, and all samples in these two clusters can be represented by a single Gaussian model; H_1 – the two clusters should not be merged, and their samples should be modeled by their respective Gaussian models. The difference between the Bayesian Inference Criterion (BIC) values of these two hypotheses, namely ΔBIC, provides a measurement to determine which hypothesis should be adopted. Mathematically, the computation of ΔBIC has the following simple closed form solution:

$$\begin{aligned}\Delta BIC &= BIC(H_1) - BIC(H_0) \\ &= \frac{(N_A+N_B)}{2}\log(|\boldsymbol{\Sigma}|) - \frac{N_A}{2}\log(|\boldsymbol{\Sigma}_A|) \\ &\quad - \frac{N_B}{2}\log(|\boldsymbol{\Sigma}_B|) - \lambda\#(H_0)\log(N_A+N_B),\end{aligned} \quad (1)$$

where N_A and N_B denote the number of samples in cluster A and B respectively; $\boldsymbol{\Sigma}_A$, $\boldsymbol{\Sigma}_B$, and $\boldsymbol{\Sigma}$ are the covariance matrices of the samples in clusters A, B, and their union; $\#(H_0)$ is the number of model parameters used in hypothesis H_0, i.e., the number of parameters of a single Gaussian model; and λ is a tunable parameter. Obviously, the typical ΔBIC evaluation does not conceal the effect of phonetic variation on the

distance measure. If multiple subspaces for different kinds of phonetic content, such as /u/ and /o/ phones, are created so that Gaussian model estimation and ΔBIC distance measurement can be applied on these phonetic subspaces respectively, we might be able to handle the phonetic variation in the ΔBIC distance measure. To this end, the acoustic feature space is divided into a set of phonetic subspaces according to the corresponding phonetic content. Then, a phonetic subspace mixture (PSM) model is used to replace the single Gaussian model in the ΔBIC distance measurement.

The remainder of this paper is organized as follows. Section 2 introduces the proposed approach, including how to generate a PSM model, how to evaluate ΔBIC of two PSM models, and how to design the phonetic subspaces. Section 3 reports the results of speaker diarization experiments on several datasets. Section 4 contains some concluding remarks.

2. Phonetic subspace mixture model

To enable the ΔBIC distance measurement to exploit phonetic information, we propose using the phonetic subspace mixture (PSM) model. The basic idea of the phonetic subspace is illustrated in Fig. 2(b). Unlike the typical ΔBIC distance measurement, which uses a single Gaussian to model the samples in a cluster, the PSM model-based ΔBIC distance measurement divides the sample space into several subspaces, each corresponding to a phonetic category. For each subspace, a Gaussian is used to model the samples in the subspace.

However, three major problems must be solved before applying the PSM model in the ΔBIC distance measurement: 1) How to estimate the PSM model from the speech samples? 2) How to evaluate the ΔBIC distance measure between PSM models? 3) How to design the phonetic subspaces for generating the PSM model?

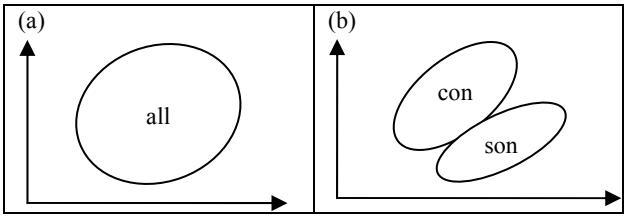

Figure 2: (a) The typical ΔBIC distance measurement, which uses a single Gaussian to model the samples in a cluster. (b) The PSM model-based ΔBIC distance measurement uses the PSM model to model the samples in a cluster. In (b), the sample space is divided into two phonetic subspaces – consonant and sonorant; and, for each phonetic subspace, a Gaussian is derived from the samples belonging to the subspace.

2.1. Estimation of the PSM model

In the typical ΔBIC distance measure in Eq. (1), a Gaussian estimated according to the maximum likelihood (ML) criterion is used to model the distribution of the acoustic feature samples $\mathbf{O}=\{\mathbf{o}_1,\cdots,\mathbf{o}_N\}$ in a cluster, as shown in Fig. 2(a). In the PSM model-based ΔBIC distance measure, the acoustic feature sample space is divided into L phonetic subspaces based on the phonetic knowledge; and, for each phonetic subspace, a Gaussian is used to model the samples in the subspace. The PSM model is similar to a Gaussian mixture model (GMM); however, each Gaussian in the PSM model has a phonetic meaning.

To estimate the PSM model, we need to identify the acoustic feature samples that belong to each phonetic subspace. Let $\mathbf{Y}=\{y_1,\cdots,y_N\}$, where $y_n\in\{1,\cdots,L\}$, denotes the phonetic subspace indices for a set of acoustic feature samples $\mathbf{O}=\{\mathbf{o}_1,\cdots,\mathbf{o}_N\}$. Suppose there is a phonetic knowledge source $\mathbf{\Lambda}$ that determines the probability $P(y_n\,|\,\mathbf{O},\mathbf{\Lambda})$. Then, the ML-estimated PSM model can be obtained by

$$\mathbf{\Theta}^*=\arg\max_{\mathbf{\Theta}} E\left[\log P(\mathbf{O},\mathbf{Y}\,|\,\mathbf{\Theta})\,|\,\mathbf{O},\mathbf{\Lambda}\right], \tag{2}$$

where $\mathbf{\Theta}=\{\alpha_1,\boldsymbol{\mu}_1,\mathbf{\Sigma}_1,...,\alpha_L,\boldsymbol{\mu}_L,\mathbf{\Sigma}_L\}$ is the parameter set of the PSM model, α_l denotes the weight of the Gaussian for the l th phonetic subspace in the cluster, and $\boldsymbol{\mu}_l$ and $\mathbf{\Sigma}_l$ are the associated Gaussian's mean vector and covariance matrix respectively. According to the M step of the expectation maximization (EM) algorithm, the solutions of the parameters in Eq. (2) are as follows:

$$\alpha_l^{ML}=\frac{\sum_{n=1}^{N}P(y_n=l\,|\,\mathbf{O},\mathbf{\Lambda})}{N}, \tag{3}$$

$$\boldsymbol{\mu}_l^{ML}=\frac{\sum_{n=1}^{N}P(y_n=l\,|\,\mathbf{O},\mathbf{\Lambda})\cdot\mathbf{o}_n}{\sum_{n=1}^{N}P(y_n=l\,|\,\mathbf{O},\mathbf{\Lambda})}, \tag{4}$$

and

$$\mathbf{\Sigma}_l^{ML}=\frac{\sum_{n=1}^{N}P(y_n=l\,|\,\mathbf{O},\mathbf{\Lambda})\cdot(\mathbf{o}_n-\boldsymbol{\mu}_l)(\mathbf{o}_n-\boldsymbol{\mu}_l)^T}{\sum_{n=1}^{N}P(y_n=l\,|\,\mathbf{O},\mathbf{\Lambda})}. \tag{5}$$

In our current implementation, a speaker independent free phone decoder with context independent (CI) hidden Markov models (HMMs) trained on the TIMIT corpus is used as the phonetic knowledge source $\mathbf{\Lambda}$ to provide $P(y_n=l\,|\,\mathbf{O},\mathbf{\Lambda})$.

2.2. Computation of PSM model-based ΔBIC

Since ΔBIC is the difference between the BIC values of two hypotheses, we first revisit the definition of BIC to facilitate the deduction of the ΔBIC distance measure between two PSM models. For a set of data samples $\mathbf{O}=\{\mathbf{o}_1,\cdots,\mathbf{o}_N\}$, the BIC value of a model hypothesis H is defined as

$$BIC(H)=\log P(\mathbf{O}\,|\,\mathbf{\Theta})-\frac{1}{2}\lambda\#(H)\ln N, \tag{6}$$

where N is the number of data points in $\mathbf{O}$; $\mathbf{\Theta}$ is the parameter set of the model hypothesis H; and #(H) is the number of parameters in $\mathbf{\Theta}$. To evaluate the BIC value of a PSM model, we use the ML-estimated PSM model derived in Section 2.1. Therefore, $\log P(\mathbf{O}\,|\,\mathbf{\Theta})$ in Eq. (6) can be expressed as

$$\begin{aligned}\log P(\mathbf{O}\,|\,\mathbf{\Theta}^{ML})&=\sum_{n=1}^{N}\log P(\mathbf{o}_n\,|\,\mathbf{\Theta}^{ML})\\&=\sum_{n=1}^{N}\log\left(\sum_{l=1}^{L}\alpha_l^{ML}\cdot P(\mathbf{o}_n\,|\,\boldsymbol{\mu}_l^{ML},\mathbf{\Sigma}_l^{ML})\right).\end{aligned} \tag{7}$$

However, the summation over all phonetic subspaces inside the log function in Eq. (7) makes the deduction very complicated. To solve such problems, the variable $\mathbf{Y}=\{y_1,\cdots,y_N\}$ and the phonetic knowledge source $\mathbf{\Lambda}$ are introduced again so that Eq. (7) can be rewritten as

$$\begin{aligned}&\log P(\mathbf{O}\,|\,\mathbf{\Theta}^{ML})=E_{\mathbf{Y}}\left[\log P(\mathbf{O},\mathbf{Y}\,|\,\mathbf{\Theta}^{ML})\,|\,\mathbf{O},\mathbf{\Lambda}\right]+Ent[\mathbf{Y}]\\&=\sum_{l=1}^{L}\sum_{n=1}^{N}\left\{\log\left[\alpha_l^{ML}\cdot P(\mathbf{o}_n\,|\,\boldsymbol{\mu}_l^{ML},\mathbf{\Sigma}_l^{ML})\right]\cdot P(y_n=l\,|\,\mathbf{O},\mathbf{\Lambda})\right\}+Ent[\mathbf{Y}]\\&=\sum_{l=1}^{L}\left\{\begin{array}{l}\sum_{n=1}^{N}\log\alpha_l^{ML}\cdot P(y_n=l\,|\,\mathbf{O},\mathbf{\Lambda})+\\\sum_{n=1}^{N}\log P(\mathbf{o}_n\,|\,\boldsymbol{\mu}_l^{ML},\mathbf{\Sigma}_l^{ML})\cdot P(y_n=l\,|\,\mathbf{O},\mathbf{\Lambda})\end{array}\right\}+Ent[\mathbf{Y}]\end{aligned}, \tag{8}$$

where

$$\sum_{n=1}^{N}\log\alpha_l^{ML}\cdot P(y_n=l\,|\,\mathbf{O},\mathbf{\Lambda}) = \log(\alpha_l^{ML})\sum_{n=1}^{N}P(y_n=l\,|\,\mathbf{O},\mathbf{\Lambda}) = \log(\alpha_l^{ML})\cdot N\cdot\alpha_l^{ML}\,; \quad (9)$$

$$\sum_{n=1}^{N}\log P(\mathbf{o}_n\,|\,\boldsymbol{\mu}_l^{ML},\boldsymbol{\Sigma}_l^{ML})\cdot P(y_n=l\,|\,\mathbf{O},\mathbf{\Lambda}) = -\frac{\left(\sum_{n=1}^{N}P(y_n=l\,|\,\mathbf{O},\mathbf{\Lambda})\right)\cdot D}{2}\log(2\pi) - \frac{\sum_{n=1}^{N}P(y_n=l\,|\,\mathbf{O},\mathbf{\Lambda})}{2}\log\left(|\boldsymbol{\Sigma}_l^{ML}|\right) - \frac{1}{2}\sum_{n=1}^{N}P(y_n=l\,|\,\mathbf{O},\mathbf{\Lambda})\cdot D\,, \quad (10)$$

where D is the dimension of the acoustic feature vector $\mathbf{o}_n$; and $Ent[\mathbf{Y}] = -\sum_{n=1}^{N}\sum_{l=1}^{L}\{P(y_n=l\,|\,\mathbf{O},\mathbf{\Lambda})\cdot\log P(y_n=l\,|\,\mathbf{O},\mathbf{\Lambda})\}$ is the entropy of the variable $\mathbf{Y}$.

When determining whether two clusters A and B should be merged using the PSM model-based ΔBIC distance measure, two hypotheses are made: H_0 – the two clusters should be merged, and all samples in the clusters can be represented by a PSM model; H_1 – the two clusters should not be merged, and their samples should be modeled by their respective PSM models. The ΔBIC value is computed by

$$\Delta BIC = BIC(H_1) - BIC(H_0) = \left\{\log P(\mathbf{O}_A\,|\,\mathbf{\Theta}_A^{ML}) + \log P(\mathbf{O}_B\,|\,\mathbf{\Theta}_B^{ML}) - \frac{1}{2}\lambda\#(H_1)\ln(N_A+N_B)\right\} - \left\{\log P(\mathbf{O}_{A+B}\,|\,\mathbf{\Theta}^{ML}) - \frac{1}{2}\lambda\#(H_0)\ln(N_A+N_B)\right\}. \quad (11)$$

Since two PSM models are used in hypothesis H_1, but only one is used in H_0, we know that $\#(H_1) = 2\cdot\#(H_0)$. By substituting Eq. (8) into Eq. (11), we obtain

$$\Delta BIC = \sum_{l=1}^{L}\left\{\begin{array}{l} N_A\alpha_{lA}^{ML}\log(\alpha_{lA}^{ML}) + N_B\alpha_{lB}^{ML}\log(\alpha_{lB}^{ML}) - (N_A+N_B)\alpha_l^{ML}\log(\alpha_l^{ML}) \\ + \frac{(N_A+N_B)\alpha_l^{ML}}{2}\log(|\boldsymbol{\Sigma}_l^{ML}|) - \frac{N_A\alpha_{lA}^{ML}}{2}\log(|\boldsymbol{\Sigma}_{lA}^{ML}|) - \frac{N_B\alpha_{lB}^{ML}}{2}\log(|\boldsymbol{\Sigma}_{lB}^{ML}|) \end{array}\right\} - \lambda\#(H_0)\log(N_A+N_B) \quad (12)$$

where $\alpha_l = (N_A\alpha_{lA}^{ML} + N_B\alpha_{lB}^{ML})/(N_A+N_B)$. Eq. (12) shows that the PSM model-based ΔBIC value is the summation of the ΔBIC values computed on individual phonetic subspaces. It is clear that Eq. (1) could be deduced from Eq. (12) by setting the number of phonetic subspaces L to one. Since the PSM model-based ΔBIC value also has a closed form, the computation is as efficient as the typical ΔBIC value.

2.3. Design of phonetic subspaces

Since the phonetic knowledge source used in this paper is a free phone decoder trained on the TIMIT corpus, the most intuitive way to design the phonetic subspaces for the PSM model is based on the TIMIT 61 phones, i.e., the sample space is divided into 61 subspaces. However, such a scheme has two apparent drawbacks: 1) There is no guarantee that the phone decoder can provide perfect phone recognition accuracy. The recognition errors generated by the decoder could introduce serious noise to the phonetic variation removal process, and thereby affect the speaker diarization performance. 2) At the beginning of the clustering process, some clusters may only contain a few phones. In other words, for such clusters, many phonetic subspaces might be empty. This would affect the PSM model-based ΔBIC distance measurement.

To avoid the above two drawbacks, instead of using 61 phonetic subspaces, we reduce the number of phonetic subspaces by dividing the 61 TIMIT phones into several categories. The categorization is implemented by performing AHC clustering on the HMM models of the 61 phones used in the phone decoder. The distance measure between two HMM models is computed by

$$dist(hmm_1, hmm_2) = \sum_{state=1}^{stateNum} KL2(hmm_1.GMM_{state}, hmm_2.GMM_{state}), \quad (13)$$

where the KL2 distance between two GMMs is computed by the method in [9].

Table 1 shows the clustering results of TIMIT's 61 phones. The first six columns in the table correspond to the configurations of the phonetic subspace setting evaluated in this paper, which yield from one to six phonetic subspace mixtures, respectively. The last column shows the TIMIT phones that belong to each phonetic subspace. The phonetic subspace setting derived by the data driven approach shows a hierarchical structure that roughly matches the structure discovered in phonetics.

Table 1: AHC results for TIMIT's 61 phone HMM models.

1 PS	2 PS	3 PS	4 PS	5 PS	6 PS	TIMIT phones
All	PS2-1	PS3-1	PS4-1	PS5-1	PS6-1	s z ch jh sh zh
				PS5-2	PS6-2	b d g f th epi k t p hh v dh
					PS6-3	h# pau bcl dcl tcl gcl kcl pcl
	PS2-2	PS3-2	PS4-2	PS5-3	PS6-4	l el w oy r er axr ax uh ih ix uw ux aa ao ah ow aw ae eh ay
			PS4-3	PS5-4	PS6-5	m em n en ng axh q hv dx nx y ey iy
		PS3-3	PS4-4	PS5-5	PS6-6	eng

3. Experiments

3.1. Experiment setup

Our speaker diarization experiments were conducted on three broadcast news corpora, namely, NIST-RT02 [10], NIST-RT03 [10], and MATBN [11]. NIST-RT02 is a small English broadcast news corpus, which consists of six 10-minute shows. We used the first and second shows as the development set to tune the parameters and the remaining four shows as the evaluation set. NIST-RT03 is also an English broadcast news corpus. There are also 6 recordings; however, the average length is about 30 minutes, which is much longer than that of NIST-RT02. The corpus was also divided into a development set and an evaluation set. MATBN is a Mandarin broadcast news corpus. The development and evaluation sets have five 45-minute shows respectively [11].

We used 12 Mel-frequency cepstrum coefficients (MFCCs) and energy plus their first and second order delta coefficients as the 39-dimensional acoustic feature vector for our free phone decoder and speaker diarization system.

The free phone decoder used to generate the posterior probability of the phonetic subspace index $P(y_n\,|\,\mathbf{O},\mathbf{\Lambda})$ is a mono-phone decoder trained on the TIMIT English corpus. The phone recognition accuracy on the TIMIT standard test set is 55.40% (for 61 phones), and 63.45% when the recognized results are mapped into the MIT/CMU 39 phone set. In the speaker diarization experiments, for each audio recording, a phone graph is first generated by the decoder, and

then a forward-backward algorithm is applied to the graph to estimate the posterior probability of the 61 phones for each frame. Finally, the phone posterior probabilities are mapped to the phonetic subspace posterior probabilities according to Table 1.

We adopted the speaker diarization system that uses AHC speaker clustering based on the typical ΔBIC distance measure as our baseline, which was in fact equivalent to the proposed system that used a single phonetic subspace. In the experiments, all the speaker diarization systems used the same speaker segmentation results produced by the DACDec3 method proposed in [12]. The segmentation results were tuned to yield the best speaker diarization performance for the baseline system. The flow chart of the proposed PSM model-based speaker diarization system is shown in Fig. 3.

For the performance evaluation, we used the diarization evaluation tool (md-eval-v21.pl) released by NIST [13] to evaluate the diarization error rate.

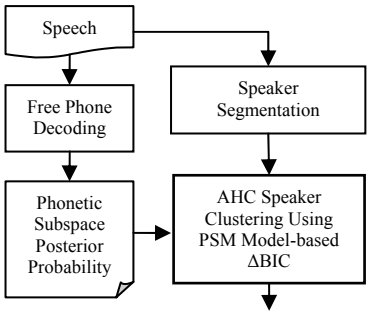

Figure 3: The flow chart of the proposed PSM model-based speaker diarization system.

3.2. Experiment results

Table 2 shows the speaker diarization error rates of the baseline and the proposed methods on the three corpora mentioned in the previous section. On the NIST-RT02 English broadcast news corpus, the proposed method has a 2.44% absolute error reduction compared to the baseline method. The improvement can also be found in the results of experiments on the NIST-RT03 and MATBN corpora. Recall that, as the number of phonetic subspaces increases, the recognition errors of the decoder and the empty subspace problem start to affect the performance of the PSM model-based clustering, and thus increase the speaker diarization error rate. This phenomenon is consistent across all three corpora. Generally speaking, when a simple free phone decoder is used as the phonetic knowledge source, it is appropriate to use two to three phonetic subspaces.

Table 2: Diarizatoin error rates (%) of the baseline and proposed methods on three different corpora.

Corpus	NIST-RT02		NIST-RT03		MATBN	
Phonetic Subspace	DEV	EVAL	DEV	EVAL	DEV	EVAL
1 PS (baseline)	10.33	17.67	**15.43**	10.76	18.25	20.03
2 PS	9.86	**15.23**	15.61	10.03	**17.19**	19.45
3 PS	9.86	16.13	15.61	10.03	17.22	**19.30**
4 PS	9.68	16.23	16.11	**10.01**	17.91	19.47
5 PS	**9.40**	15.76	24.69	13.67	25.20	23.73
6 PS	9.85	16.94	25.23	16.83	30.69	27.87

4. Conclusion

We have investigated the use of phonetic information in speaker diarization. The proposed phonetic subspace mixture (PSM) model separates speaker variations from phonetic variations in speech. As a result, speaker diarizaiton systems can focus on the characteristics of speakers when performing speaker clustering. The typical ΔBIC distance measurement can be seen as a special case of the proposed PSM model-based ΔBIC distance measurement. Like the typical ΔBIC distance measurement, the new ΔBIC distance measurement also has a closed form solution; hence, the computation is efficient. We used a data driven method to construct a hierarchical structure of phones for designing the phonetic subspaces. Our experiment results show that the speaker diarization performance could be improved to some extent even if the phonetic information was provided by a very simple free phone decoder.

There is still much room for improvement. For example, we could improve the phone decoder by using language models to help determine phonetic subspaces. We can adjust the number of phonetic subspaces in a cluster dynamically based on the cluster's size and the hierarchical structure of phonetic subspaces. Moreover, we can apply the PSM model-based ΔBIC distance measure in speaker segmentation.

5. Acknowledgement

This work was supported in part by the National Science Council of Taiwan under Grant: NSC96-2628-E-001-024-MY3.

6. References

[1] S. E. Tranter and D. A. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Trans. on Audio, Speech and Language Processing*, 14(5), pp. 1557-1565, 2006.

[2] H. M. Wang, S. S. Cheng, and Y. C. Chen, "The SoVideo Mandarin Chinese broadcast news retrieval system," *International Journal of Speech Technology*, 7(2-3), pp. 189-202, 2004.

[3] X. Zhong, M. Clements, and S. Lim, "Acoustic change detection and segment clustering of two-way telephone conversation," in *Proc. Eurospeech*, 2003.

[4] C. Barras, X. Zhu, S. Meignier, and J.-L. Gauvain, "Multistage speaker diarization of broadcast news," *IEEE Trans. on Audio, Speech and Language Processing*, 14(5), pp. 1505-1512, 2006.

[5] S. Meignier, D. Moraru, C. Fredouille, J.-F. Bonastre, and L. Besacier, "Step-by-step and integrated approaches in broadcast news speaker diarization," *Computer Speech & Language*, 20, pp. 303-330, 2006.

[6] J. M. Pardo, X. Anguera, and C. Wooters, "Speaker diarization for multiple-distant-microphone meetings using several sources of information," *IEEE Trans. on Computers*, 56(9), pp. 1212-1224, 2007.

[7] S. Chen and P. S. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the Bayesian information criterion," in *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, 1998.

[8] K. J. Han and S. S. Narayanan, "A robust stopping criterion for agglomerative hierarchical clustering in a speaker diarization system," in *Proc. Interspeech*, 2007.

[9] J. Goldberger and H. Aronowitz, "A distance measure between GMMs based on the unscented transform and its application to speaker recognition," in Proc. Interspeech, 2005

[10] NIST, *Rich Transcription Evaluation Project*, http://www.itl.nist.gov/iad/mig//tests/rt/

[11] H. M. Wang, B. Chen, J. W. Kuo, and S. S. Cheng, "MATBN: A Mandarin Chinese broadcast news corpus," *Int. J. Comput. Linguist. Chinese Lang. Process.*, 10(2), pp. 219-236, 2005.

[12] S. S. Cheng, H. M. Wang, and H. C. Fu, "BIC-based speaker segmentation using divide-and-conquer strategies with application to speaker diarization," *IEEE Trans. on Audio, Speech and Language Processing*, 18(1), pp. 141-157, 2010.

[13] NIST, *Rich Transcription Spring 2006 Evaluation*, http://www.nist.gov/speech/tests/rt/2006-spring/index.html.