Speaker Vector-Based Speaker Recognition with Phonetic Modeling

Tetsuo Kosaka, Tatsuya Akatsu, Masaharu Kato and Masaki Kohda
Yamagata University
Japan

1. Introduction

This chapter describes anchor model-based speaker recognition with phonetic modeling. Gaussian Mixture Models (GMMs) have been successfully applied to characterize speakers in speaker identification and verification when a large amount of enrolment data to build acoustic models of target speakers is available. However, a small amount of enrolment data of as short as 5 sec. might be preferred for some tasks. A conventional GMM-based system does not perform well if the amount of enrolment data is limited. In general, 1-minute or more of enrolment data are required in the conventional system.

In order to solve this problem, a speaker characterization method based on anchor models has been proposed. The first application of the method was proposed for speaker indexing (Sturim et al., 2001). And the method has been also used for speaker identification (Mami & Charlet, 2003) and speaker verification (Collet et al., 2005).

In the anchor model-based system, the location of each speaker is represented by a speaker vector. The speaker vector consists of the set of the likelihood between a target utterance and the anchor models. It can be considered as a projection of the target utterance in a speaker space. One of the merits of this approach is that it is not necessary to train a model for a new target speaker, because the set of anchor models does not include the model of target speaker. It can save users time to utter iteratively for model training.

However, there is a significant disadvantage in the system because the recognition performance is insufficient. It has been reported that an identification rate of 76.6% was obtained on a 50-speaker identification task with 16-mixture GMMs as anchor models (Mami & Charlet, 2003). Also, an equal error rate (EER) of 11.3% has been reported on speaker verification task with 256-mixture GMMs (Collet et al., 2005). Compared with the conventional GMM approach, the performance of anchor model-based system is remarkably insufficient.

The aim of this work is to improve the performance of the method by using phonetic modeling instead of the GMM scheme as anchor models and to develop text-independent speaker recognition system that can perform accurately with very short reference speech. A GMM-based acoustic model covers all phonetic events for each speaker. It can represent an overall difference in acoustic features between speakers, however, it cannot represent a difference in pronunciation. Consequently, we propose the method to detect the detailed difference in phonetic features and try to use it as information for speaker recognition. In order to detect the phonetic features, a set of speaker-dependent phonetic HMMs is used as

the anchor models. The likelihood calculation between the target utterance and the anchor models is performed with an HMM-based phone recognizer with a phone-pair grammar.

In order to evaluate the proposed method, we compare the phonetic-based system with the GMM-based system on the framework of the anchor model. The number of dimensions in speaker space is also investigated. For this purpose, a large-size speech corpus is used for training the anchor models (Nakamura et al., 1996). Furthermore, another anchor model-based system in which phonetically structured GMMs (ps_GMMs) are used is compared to show the reason why phonetic modeling is effective in this method. Phonetically structured GMMs have been proposed by Faltlhauser (Faltlhauser & Ruske, 2001) to improve speaker recognition performance. In the method, 'phonetic' mixture components in a single state are weighted in order to improve speaker recognition performance.

The rest of this introduction reviews some related work. Recently, some phonetic based methods have been proposed. Hebert et al. have proposed the speaker verification method based on a tree structure of phonetic classes (Hebert & Heck, 2003). This paper reported that the proposed phonetic class-based system overcame a conventional GMM approach. Park et al. have proposed a speaker identification method in which phonetic class GMMs for each speaker were used (Park & Hazen, 2002). Both two methods differ from our approach in that a model for a target speaker is needed. Kohler et al. have developed a speaker-recognition system based only on phonetic sequences instead of the method based on acoustic feature vectors (Kohler et al., 2001). In this method, a test speaker model is not an acoustic model and it is generated by using n-phone frequency counts. The work which has a similar motivation of reducing enrolment data has been proposed by Thyes (Thyes et al., 2000). In this work, the concept of 'eigenvoice' was used for representing a speaker space. The method of 'eigenvoice' was proposed for speaker adaptation on earlier work (Kuhn et al., 1998). A phonetic information was not used for speaker discrimination in that work.

This chapter is organized as follows. Section 2 describes the method of speaker recognition. Section 3 shows the results of speaker identification experiments. Finally, we conclude the paper and suggest future research in Section 4.

2. Speaker recognition with phonetic-based modeling

2.1 Conventional speaker recognition

The technology of the speaker recognition can be categorized into two fields: one is speaker identification, and the other is speaker verification. Speaker identification is a technique for assigning the input utterance to one person of a known speaker set, while speaker verification is a technique for confirming the identity of the input speaker. Although the anchor model-based method can be applied to both techniques, a speaker identification method is described mainly in the following sections. Since it is difficult to separate the speaker information from the phonetic one, many speaker recognition systems perform in a text dependent way. In those systems, users must utter a predefined key sentence. However, sometimes that is not acceptable to users. In this work, we have developed the speaker recognition system which behaves in a text independent way. In that system, users can utter an arbitrary sentence.

In conventional speaker recognition systems, GMMs have been successfully used to characterize speakers. The characteristics of a reference speaker are modeled by GMM,

$$p(o_t \mid \lambda) = \sum_k w_k b_k(o_t) \,, \tag{1}$$

with mixture weights w_k and Gaussian densities $b_k(o_t)$. The average log-likelihood of a model given an utterance $o = \{o_1, \cdots, o_T\}$ is calculated as

$$L(o \mid \lambda) = \frac{1}{T} \sum_{t=1}^{T} \log p(o_t \mid \lambda) . \tag{2}$$

The average log-likelihood scores are compared to determine an input speaker in speaker identification system. For speaker verification system, those scores are normalized to reduce the variation of utterances,

$$\widetilde{L}(o \mid \lambda) = L(o \mid \lambda) - L(o \mid \lambda_{UBM}) , \tag{3}$$

where λ_{UBM} is the Universal Background Model which is derived from training data of all or selected speakers to normalize the variation. For the tasks of speaker identification or verification, a reference speaker model λ must be trained by using 60 sec. or more of enrolment speech in advance. It causes users the loss of taking time to utter iteratively. Since acoustic models of reference speakers are not required in anchor model-based system, the user has only to utter just one sentence in advance.

2.2 Speaker space representation using anchor models

In the anchor model-based system, the speaker is characterized by a vector consisted of the set of the likelihood between the target utterance and the anchor models. The speech utterance is represented by the following vector:

$$v = \begin{bmatrix} \dfrac{\log p(o \mid M_1) - \mu}{\sigma} \\ \dfrac{\log p(o \mid M_2) - \mu}{\sigma} \\ \vdots \\ \dfrac{\log p(o \mid M_N) - \mu}{\sigma} \end{bmatrix} , \tag{4}$$

where

$$\mu = \frac{1}{N} \sum_{n=1}^{N} \log p(o \mid M_n) , \tag{5}$$

and

$$\sigma = \sqrt{\frac{1}{N} \sum_{n=1}^{N} \left(\log p(o \mid M_n) - \mu \right)^2 } . \tag{6}$$

$\log P(o \mid M_n)$ is the log-likelihood of the input utterance o for the anchor model M_n. The vector is normalized to have zero mean and unit variance to reduce the likelihood variation among utterances (Akita & Kawahara, 2003). N is the number of anchor models and denotes the number of dimensions of speaker space. In the identification step, a measure

between each reference speaker vector r_i and the input vector is calculated in speaker space. We used Euclidean metric for distance calculation. Input speaker is identified by:

$$\hat{i} = \arg\min_i D(v, r_i) \ , \tag{7}$$

where i is a reference speaker index and input vector v is determined to be uttered by speaker $\hat{i}$. Note that the target speaker is not included in N speakers of anchor models. Since the method doesn't require the model training for the target speaker, only about single utterance is needed for reference vector.

2.3 Phonetic representation of anchor models

In the previous works, GMMs were used as the anchor models (Sturim et al., 2001; Mami & Charlet, 2003). A GMM covers all phonetic events for each speaker, however it does not directly consider phonetic information. It can cause the degradation in the performance of speaker discrimination. For example, the vowel /a/ of speaker A is sometimes confused with the vowel /o/ of speaker B in a phoneme recognition task. The GMM-based system cannot represent such a difference. Consequently, we propose the method to detect the detailed difference in phonetic features and try to use it as information for speaker recognition.

In order to improve the performance of the method, phonetic HMMs or phonetically structured GMMs are used. In this section, phonetic HMM based system is described. This approach requires a phonetic speech recognizer in order to calculate the log-likelihood $\log P(o \mid M_n)$ shown in Eq. (4). The log-likelihood for the speaker n is obtained by a phoneme recognizer with speaker dependent phonetic HMMs. Since the recognizer can decode an unknown utterance, the identification system can be performed in a text independent way.

Figure 1 indicates the examples of speaker vector composed of 1000 dimensions with both HMMs and GMMs. The horizontal axis represents the values of speaker vector for the utterance number 26, and the vertical axis represents the values for the utterance number 27. Both utterances are given by the same speaker (female), however, the contents of utterances are different. Each point represents the values of one of the 1000 anchor models. The left figure shows the vector values with HMMs as anchor models, and the right one shows the vector values with GMMs as anchor models. Even though contents are not same between horizontal axis and vertical axis, two sets of values indicate a similar tendency in these figures. This observation suggests that the speaker identification in text independent way can be performed with this method. In the left figure, the correlation between two is smaller than that of the right one. This suggests that speaker recognition performs well by using HMMs as anchor models.

Figure 2 also indicates the examples of speaker vector composed of 1000 dimensions with both HMMs and GMMs. In this figure, the horizontal axis and the vertical axis represent the different speakers (F.AIFU and F.HAHZ are speaker IDs. Both speakers are female). However, the contents of utterances are same.

In contrast to the results of Fig. 1, a correlation between two axes is small. This means that the values of speaker vector differ too much between different speakers even if the contents of utterances are same.
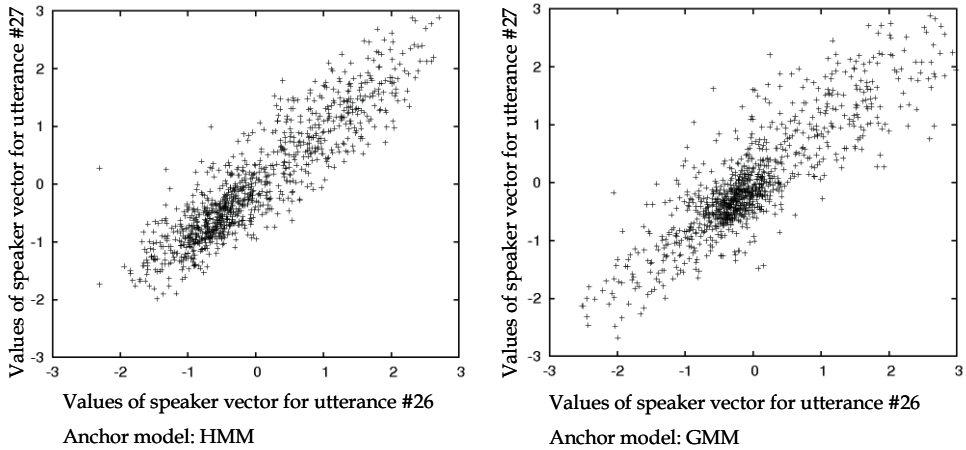
Fig. 1. Example of values of speaker vector. The horizontal and vertical axes represent the same speaker. (the left figure: HMMs are used as anchor models, the right figure: GMMs are used as anchor models)
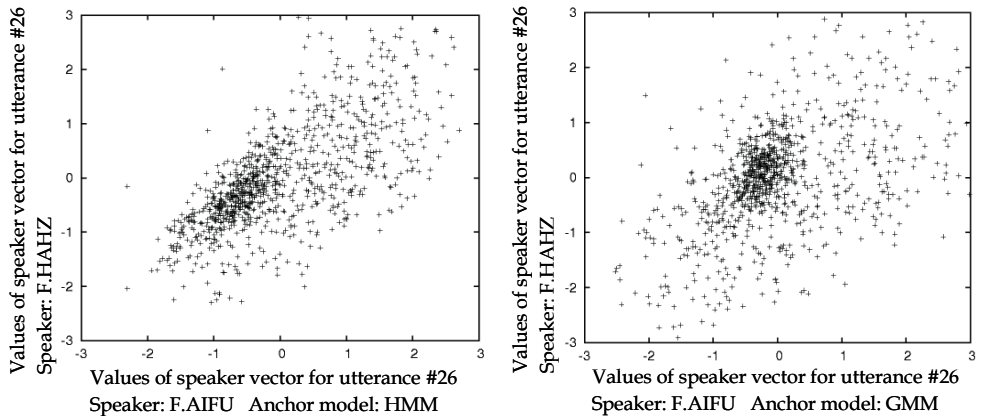

Fig. 2. Example of values of speaker vector. The horizontal and vertical axes represent the different speakers. (the left figure: HMMs are used as anchor models, the right figure: GMMs are used as anchor models)

2.4 Phonetically structured GMMs as anchor models

In our approach, phonetically structured GMMs (ps_GMMs) (Faltlhauser & Ruske, 2001) are also used as anchor models to find the reason why phonetic modeling is effective in the proposed method. In the method, 'phonetic' mixture components in a single state are weighted in order to create a GMM based on phonetic classes. In our work, PDFs obtained from monophone HMMs are used for 'phonetic' mixture components.

Assume that S-state and K-mixture monophone HMM for phoneme j is trained in advance. Total number of PDFs in monophone HMMs is $J \times S \times K$ when the number of phonemes is J. Those PDFs are gathered to make a single state model λ_{ps},

$$p(o_t \mid \lambda_{ps}) = \sum_{j=1}^{J} \sum_{s=1}^{S} \sum_{k=1}^{K} w_{jsk} N(o_t, \mu_{jsk}, \Sigma_{jsk}),$$ (8)

where $N(o_t, \mu_{jsk}, \Sigma_{jsk})$ is a Gaussian distribution of the k th density in the mixture at state s of phoneme j and w_{jsk} is a mixture weight.

After producing λ_{ps}, parameters of λ_{ps} are re-estimated. Three types of methods were compared in a speaker identification task in order to find the best re-estimation method.

No re-estimation   Re-estimation is not conducted. All PDFs from monophone HMMs are used as they are. New mixture weight $\hat{w}_{jsk}$ is simply given as follows:

$$\hat{w}_{jsk} = \frac{w_{jsk}}{JS}.$$ (9)

Weight re-estimation   Only mixture weights are re-estimated to obey the probabilistic constraint as follows:

$$\sum_{j=1}^{J} \sum_{s=1}^{S} \sum_{k=1}^{K} w_{jsk} = 1.$$ (10)

PDF and weight re-estimation   Both all of PDFs and weights are re-estimated. In this case, phonetic information is only used as an initial model. Then the method is not exactly a phonetic modeling.

In the experiment of a 30-speaker identification task, 'weight re-estimation' and 'PDF and weight re-estimation' obtained similar identification rates. The identification rates were 89.88% and 89.99%, respectively. The performance of 'no re-estimation' was insufficient and the identification rate was 82.64%. The method of 'weight re-estimation' is used in the following experiments.

In terms of model topology, ps_GMM consists of a single state just like the conventional GMM, however, each group of PDFs in ps_GMM is trained with the data of each phoneme class. Then comparing conventional GMM with ps_GMM, the model topology is same but PDFs are different. Also comparing ps_GMM with phonetic HMM, PDFs are same but the model topology is different.

3. Speaker identification experiments

3.1 Speaker identification system

An experimental system of speaker identification has been developed and used for evaluation. In this section, we describe the overview of the system. Fig. 3 shows the block diagram of the proposed speaker identification system.

Since we are planning to develop a speaker identification method in noisy conditions, the analysis is processed by the ETSI advanced front-end (AFE-WI008) in which noise robust

algorithms are used (ETSI, 2002). In this front-end, noise reduction for additive noise and blind equalization for channel distortion are applied. The blind equalization process is omitted for our experiments, because we found that it had a bad influence on the performance of speaker identification from the results of comparative experiments. In this front-end, a speech signal is digitized at a sampling frequency of 16kHz and at a quantization size of 16bits. The length of the analysis frame is 25ms and the frame period is set to 10ms. The 13-dimensional feature (12-dimensional MFCC and log power) is derived from the digitized samples for each frame. Additionally, the delta and delta-delta features are calculated from MFCC feature and the log power. Then the total number of dimensions is 39. The delta and delta-delta coefficients are useful for the system of HMM-based anchor models. After speech analysis is carried out, an input features are transformed into a speaker vector by Eq. (4). A value of log-likelihood in Eq. (4) is obtained by a phoneme recognizer with a phone-pair grammar. In the recognizer, one-pass frame-synchronous search algorithm with beam searching has been adopted. The speaker vector derived from the input utterance is used for calculating distances from reference speaker vectors, which are computed in advance. The reference speaker of minimal distance is determined to be an identified speaker by Eq. (7).
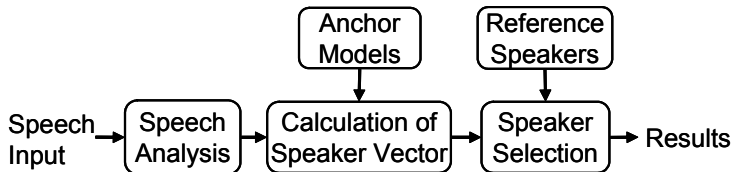

Fig. 3. Block diagram of speaker identification system

3.2 Experimental set-up

For evaluation of the proposed method, we used ATR SDB-I as a speech corpus (Nakamura et al., 2001). This corpus was designed to cover speaker variations with a large number of speakers' read speech and dialogue speech. For representing the anchor models, phonetically-balanced speech data uttered by 2032 speakers composed of 744 male and 1288 female were used. Then the maximum number of dimensions of speaker space was 2032.

Since a limited amount of speech data from each speaker was available, we used the maximum a *posteriori* (MAP) estimation instead of the ML (Maximum Likelihood) estimation for training of the anchor models. MAP estimation is successfully used for adaptation of CHMM parameters (Lee & Gauvain, 1993). It uses information from an initial model as *a priori* knowledge to complement the training data. This *a priori* knowledge is statistically combined with *a posteriori* knowledge derived from the training data. When the amount of training data is small, the estimates are tightly constrained by the *a priori* knowledge, and the estimation error is reduced.

The evaluation data sets consisted of 30 speakers, each of which contains 25 utterances. The average length of utterances in test set was 5.5 sec.

In order to avoid variations in identification performance with reference speech, the following evaluation method was adopted. 24 out of the 25 utterances by each evaluation speaker were tested, and the rest of one utterance was used as a reference speech. Then the 25 different references were used in the same way and the average identification rate of those results was calculated. Thus, the average length of test speech and that of reference

speech were same and were only 5.5 sec. In general, 1-minute or more of enrolment data are required in the conventional speaker recognition systems. Compared with those systems, the proposed system can be performed with a very short utterance. It can save users time to utter iteratively.
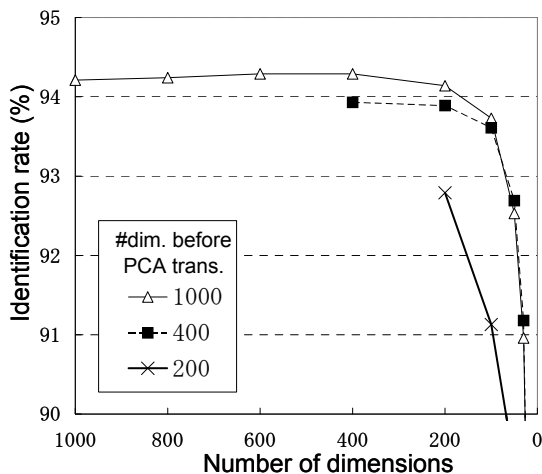

Fig. 4. Reduction of number of dimensions with PCA trans.

3.3 Influence of the number of dimensions

First, we evaluated the influence of the number of dimensions on a speaker space. In (Mami & Charlet, 2003), the speaker space was composed of 500 dimensions each of which was calculated with 16-mixture GMM. Since the detailed study is not carried out until now, an adequate number of parameters for representing speaker space is not clear. In the experiments, various numbers of dimensions from 200 to 2032 were compared. The result shows that the performance is saturated around 1000-dim. In order to study an influence of the number of dimensions further, we apply a technique of PCA (Principal Component Analysis) to reduce a redundancy of a speaker space. Since speakers for anchor models are selected randomly from speaker list, features of some speakers are similar and may be redundant. Fig. 4 shows the results of reduction of number of dimensions on a 30-speaker identification task. Three types of experiments were conducted. The number of dimensions before PCA was 200, 400 and 1000, respectively. The results show that the degradation of recognition performance was not observed at the range of 400 to 1000-dim. This means that there is some redundancy in representing the speaker space.

3.4 Comparison of anchor models

In this section, three types of anchor models are compared on 30-speaker identification task: 3-state 10-mixture phonetic HMMs (the number of phonemes is 34), 1020-mixture phonetically structured GMMs (ps_GMMs) and 1024-mixture conventional GMMs. For the phonetic HMMs, total number of PDFs is 3state × 10mixture × 34phonemes = 1020. Then the similar number of PDFs is used for three types of models, and they are comparable. The number of model parameters was determined experimentally. The details of the relation

between the number of the mixture components and the identification rate have been reported in (Kosaka et al., 2007). The ps_GMMs in this experiment are composed as follows. All PDFs except those in silence model are extracted from the phonetic HMMs to form a single state GMM. After producing the GMM, only mixture weights are re-estimated.

Table 1 shows the speaker identification result. The number of dimensions was 1000 and the number of test speakers was 30. The HMM-based system showed significant improvement over the GMM-based system, although the number of PDFs was nearly same in those systems. The performance of ps_GMMs is in between two. Comparing conventional GMM with ps_GMM, the model topology is same but PDFs are different. Also comparing ps_GMM with phonetic HMM, PDFs are same but the model topology is different. This means that both the model topology and the PDFs derived from phonetic models contributed to improve the performance of speaker identification. Finally, the identification rate of 94.21% could be obtained with 3-state 10-mixture HMM system in 30 speaker identification task. By comparison with the GMM-based system, the relative improvement of 62.9% was achieved.

We also investigated the comparison between an anchor model-based system and a conventional GMM-based system described in Sect. 2.1. In our experiments, the average length of reference speech was only 5.5 sec. and it is difficult to train GMMs accurately by using the ML (Maximum Likelihood). Thus, we used the maximum a *posteriori* (MAP) estimation instead of the ML estimation for training of conventional GMMs. The number of mixture components was varied to find the most appropriate one. The speaker identification rate of 77.14% was obtained with 8-mixture GMMs. This indicates that conventional GMM-base system does not work well with such a small amount of enrolment data.

Anchor model	HMM	ps_GMM	GMM
Identification rate (%)	94.21	89.88	84.41

Table 1. Performance comparison of three types of anchor models (#test speakers = 30)

4. Conclusions

This chapter proposed the method of anchor model-based speaker recognition in text-independent way with phonetic modeling. Since the method doesn't require model training for the target speaker, only about single utterance is needed for reference speech. In order to improve the recognition performance, phonetic modeling was used instead of Gaussian Mixture Model (GMM) scheme as anchor models. The proposed method was evaluated on Japanese speaker identification task. Compared with the performance of GMM-based system, significant improvement could be achieved. The identification rate of 94.21% could be obtained with 3-state 10-mixture HMMs in 30-speaker identification task. In the experiments, the average length of reference speech was only 5.5 sec. By comparison with the GMM-based system, the relative improvement of 62.9% was achieved. The results show that the phonetic modeling is effective for anchor model-based speaker recognition.

We are now conducting the evaluation of the method on speaker verification task. We are also conducting the evaluation of speaker identification in noisy conditions. Some results in noisy conditions have been reported in (Goto et al., 2008). The merit of this method is that the system can detect speaker characteristics with a very short utterance as short as 5 sec. Then the method can be used in the tasks of speaker indexing or tracking.

5. References

Akita, Y. & Kawahara, T. (2003), Unsupervised speaker indexing using anchor models and automatic transcription of discussions, *Proceedings of Eurospeech2003,* pp.2985-2988, Geneva, Switzerland, Sept. 2003

Collet, M.; Mami, Y.; Charlet, D. & Bimbot, F. (2005), Probabilistic anchor models approach for speaker verification, *Proceedings of INTERSPEECH2005,* pp.2005-2008, Lisbon, Portugal, Sept. 2005

ETSI, (2002), ETSI ES 202 050 V1.1.1, *STQ; Distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms,* European Telecommunications Standards Institute, France

Faltlhauser, R. & Ruske, G. (2001), Improving speaker recognition performance using phonetically structured Gaussian mixture models, *Proceedings of Eurospeech2001,* pp.751-754, Aalborg, Denmark, Sept. 2001

Goto, Y.; Akatsu, T.; Katoh, M.; Kosaka, T. & Kohda, M. (2008), An investigation on speaker vector-based speaker identification under noisy conditions, *Proceedings of ICALIP2008,* pp.1430-1435, Shanghai, China, Jul. 2008

Hebert, M. & Heck, L.P. (2003), Phonetic class-based speaker verification, *Proceedings of INTERSPEECH2003,* pp.1665-1668, Geneva, Switzerland, Sept. 2003

Kohler, M.A.; Andrews, W.D. & Campbell, J.P. (2001), Phonetic speaker recognition, *Proceedings of EUROSPEECH2001,* pp.149-153, Aalborg, Denmark, Sept. 2001

Kosaka, T.; Akatsu, T.; Katoh, M. & Kohda, M. (2007), Speaker Vector-Based Speaker Identification with Phonetic Modeling, *IEICE Transactions (Japanese)*, Vol. J90-D, No. 12, Dec. 2007, pp. 3201-3209

Kuhn, R,; Nguyen, P.; Junqua, J.-C.; Goldwasser, L.; Niedzielski, N.; Fincke, S.; Field, K. & Contolini, M. (1998), Eigenvoices for speaker adaptation, *Proceedings of ICSLP98,* pp. 1771-1774, Sydney, Australia, Dec. 1998

Lee, C.-H. & Gauvain, J.-L. (1993), Speaker adaptation based on MAP estimation of HMM parameters, *Proceedings of ICASSP93*, pp.558-561, Minneapolis, USA, Apr. 1993, IEEE

Mami, Y. & Charlet, D. (2003), Speaker identification by anchor models with PCA/LDA post-processing, *Proceedings of ICASSP2003,* pp.180-183, Hong Kong, China, Apr. 2003, IEEE

Nakamura, A.; Matsunaga, S.; Shimizu, T.; Tonomura, M. & Sagisaka, Y. (1996), Japanese speech databases for robust speech recognition, *Proceedings of ICSLP1996,* pp.2199-2202, Philadelphia, USA, Oct. 1996

Park, A & Hazen, T.J. (2002), ASR dependent techniques for speaker identification, *Proceedings of ICSLP2002,* pp.1337-1340, Denver, USA, Sept. 2002

Sturim, D.; Reynolds, D. ; Singer, E. & Campbell, J. (2001), Speaker indexing in large audio databases using anchor models, *Proceedings of ICASSP2001,* pp.429-432, Salt Lake City, USA, May. 2001, IEEE

Thyes, O.; Kuhn, R.; Nguyen, P. & Junqua, J.-C. (2000), Speaker identification and verification using eigenvoices, *Proceedings of ICSLP2000,* pp. 242-246, Beijing, China, Oct. 2000

Speech Recognition

Edited by France Mihelic and Janez Zibert

ISBN 978-953-7619-29-9

Hard cover, 550 pages

Publisher InTech

Published online 01, November, 2008

Published in print edition November, 2008

Chapters in the first part of the book cover all the essential speech processing techniques for building robust, automatic speech recognition systems: the representation for speech signals and the methods for speech-features extraction, acoustic and language modeling, efficient algorithms for searching the hypothesis space, and multimodal approaches to speech recognition. The last part of the book is devoted to other speech processing applications that can use the information from automatic speech recognition for speaker identification and tracking, for prosody modeling in emotion-detection systems and in other speech processing applications that are able to operate in real-world environments, like mobile communication services and smart homes.

INTECH
open science | open minds