International Journal of Electrical, Electronic and Communication Sciences
ISSN: 2517-9438
Vol:4, No:3, 2010

Text-independent Speaker Identification Based on MAP Channel Compensation and Pitch-dependent Features

Jiqing Han, Rongchun Gao

***Abstract*—**One major source of performance decline in speaker recognition system is channel mismatch between training and testing. This paper focuses on improving channel robustness of speaker recognition system in two aspects of channel compensation technique and channel robust features. The system is text-independent speaker identification system based on two-stage recognition. In the aspect of channel compensation technique, this paper applies MAP (Maximum A Posterior Probability) channel compensation technique, which was used in speech recognition, to speaker recognition system. In the aspect of channel robust features, this paper introduces pitch-dependent features and pitch-dependent speaker model for the second stage recognition. Based on the first stage recognition to testing speech using GMM (Gaussian Mixture Model), the system uses GMM scores to decide if it needs to be recognized again. If it needs to, the system selects a few speakers from all of the speakers who participate in the first stage recognition for the second stage recognition. For each selected speaker, the system obtains 3 pitch-dependent results from his pitch-dependent speaker model, and then uses ANN (Artificial Neural Network) to unite the 3 pitch-dependent results and 1 GMM score for getting a fused result. The system makes the second stage recognition based on these fused results. The experiments show that the correct rate of two-stage recognition system based on MAP channel compensation technique and pitch-dependent features is 41.7% better than the baseline system for closed-set test.

***Keywords*—**Channel Compensation, Channel Robustness, MAP, Speaker Identification

I. Introduction

CHANNEL mismatch happens when we enroll a speaker's speech using one microphone or handset and then identify him using a different microphone or handset, namely the channel environment of testing speech changes. In the real world, the type of the training channel and testing channel are usually different, and the acoustic parameters are different for the same type of channels. Recently, the main measures to improving channel robustness of speaker recognition system are channel compensation and channel robust features [1].

The acoustic parameters of speech signal change because of channel, so they cannot reflect the primary information, Channel compensation technique improves the acoustic parameters to make them match the acoustic parameters of training speech signal. Usually, channel compensation includes feature domain compensation, model domain compensation and score domain compensation.

Feature domain compensation aims to remove channel mismatch when feature vectors are being extracted. These include well-known and widely used techniques such as cepstral mean subtraction [2], RASTA filtering [3] and cepstral subtraction [4]. MAP (Maximum A Posterior Probability) [5] channel compensation technique has been used in speech recognition. Recently, Reynolds proposed feature mapping technique [1] which had a good performance on NIST 2002 corpora.

Model domain compensation modifies models to minimize channel mismatch. An example is SMS (Speaker Model Synthesis) [6], which learns how model parameters change between different channels and applies a transformation to synthesize speaker models under unseen enrollment conditions.

Score domain compensation attempts to remove model score scales and shifts caused by channel mismatch. Examples of score domain compensation technique are Hnorm [7] and Tnorm [8].

Human rely on several different types or levels of information in the speech signal to recognize a person from the others. These information can be the deep bass and timber of a voice, a friend's unique laugh, or the particular repeated word usage of a colleague, which have the character of channel robustness, so they are high-level information related to low-level information such as LPCC (Linear Predictive Cepstral Coefficients) and MFCC (Mel Frequency Cepstral Coefficients) which have a big distortion when channels are mismatching [9].

Researchers propose so many approaches to extract high-level features, and prove that these features can improve the system through experiments and have the potentiality of channel robustness. Some popular high-level features and application approaches [9] are enumerated as follows.

(1) Prosodic Features

Including pitch and energy distributions, pitch and energy

Jiqing Han is with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin, 150001 China (corresponding author, e-mail: jqhan@ hit.edu.cn).

Rongchun Gao was with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin, 150001 China. He is now with the Huawei Company, Shenzhen, China.

Sponsored by the National Basic Research Program of China (973 Program, No. 2007CB311100), National High Technology Research and Development Program of China (863 Program, No. 2006AA010103)

track dynamics, prosodic statistics, and so on.

(2) Phone Features

Including phone N-grams, phone binary trees, cross-stream phone modeling, pronunciation modeling, and so on.

(3) Lexical Features

(4) Conversational Features.

This paper applies MAP channel compensation technique, which was used in speech recognition, to speaker recognition system and introduces pitch-dependent features and pitch-dependent speaker model to the second stage recognition. Based on the first stage recognition to testing speech using GMM (Gaussian Mixture Model), the system uses GMM scores to decide if it needs to be recognized again. If it needs to, the system selects a few speakers from all of the speakers who participate in the first stage recognition for the second stage recognition. For each selected speaker, the system obtains 3 pitch-dependent results from his pitch-dependent speaker model, and then uses ANN (Artificial Neural Network) to unite the 3 pitch-dependent results and 1 GMM score for getting a fusion result. The system makes the second stage recognition based on these fusion results.

II. System Framework

The system is a two-stage recognition and text-independent speaker identification system. Fig. 1 shows the system diagram. For testing speech, the system applies MAP channel compensation technique to compensate LPCC in feature domain, and then recognizes by GMM in the first stage recognition. Based on selection strategy, the system uses GMM scores to decide if it needs to be recognized again. If it does not need to, the result of the first stage recognition is the final result. If it needs to, the system selects a few speakers from all of the speakers who participate in the first stage recognition and get their pitch-dependent speaker models from speaker model library. The system extracts three dimensions pitch-dependent features from the testing speech, and then does matching computation on these selected pitch-dependent speaker models. For each selected speaker, the system obtains 3 pitch-dependent results on his pitch-dependent speaker model, and then uses ANN to unite the three pitch-dependent results and one GMM score for getting the fusion results. The speaker who is relevant to the best fusion results is the final result.

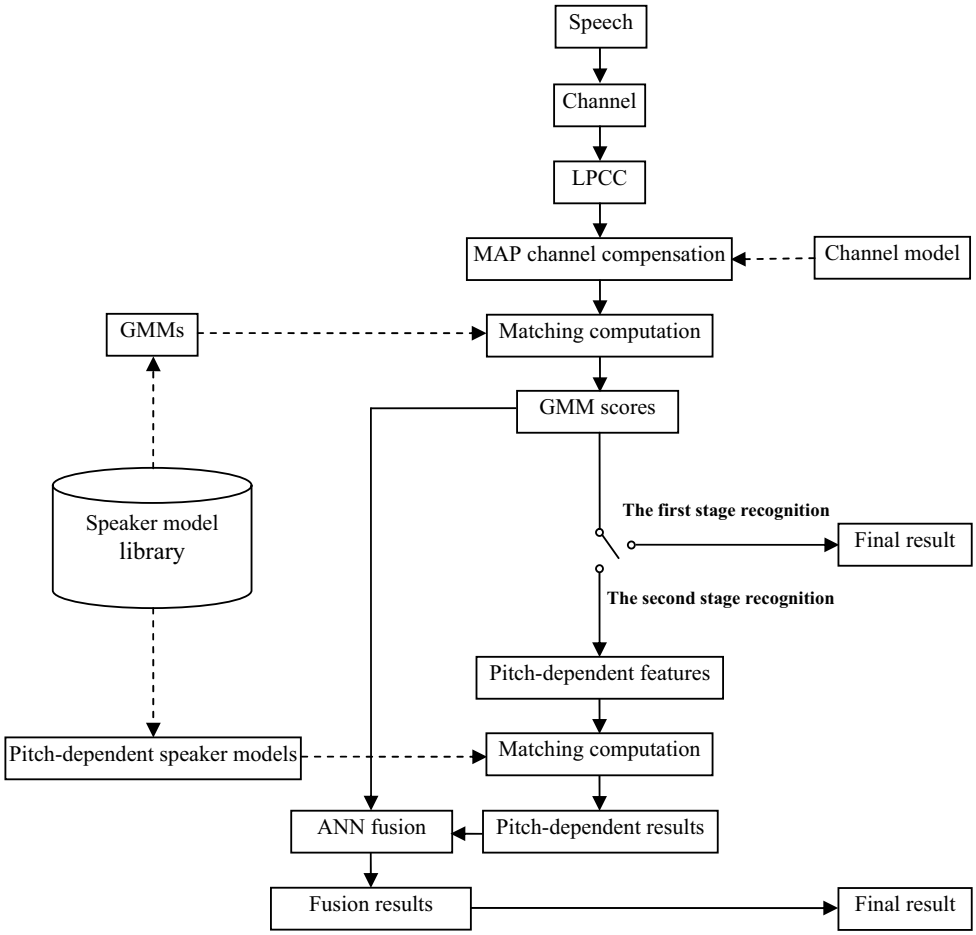

Fig. 1 System diagram

III. MAP CHANNEL COMPENSATION TECHNIQUE

A. Principle

MAP channel compensation technique is based on two hypotheses [5], the first, h denotes the channel bias between training speech and testing speech. The second, h can be modeled by a multivariate Gaussian distribution with mean vector μ_h and covariance matrix Σ_h.

Based on MAP criterion, the channel bias h is estimated as follows

$$\hat{h}_{MAP} = \arg\max_h \{P(h \mid X, \lambda)\} \quad (1)$$

λ denotes speaker model, $X = \{x_1, ..., x_T\}$ is the feature vectors of testing speech. Equation (1) is equivalent to

$$\hat{h}_{MAP} = \arg\max_h \{\log P(X \mid h, \lambda) + \log P(h)\} \quad (2)$$

$P(h)$ represents the prior pdf of h. In order to evaluate the weights of the two terms, we introduce a scale factor α into (2) showed in (3).

$$\hat{h}_{MAP} = \arg\max_h \{\alpha \log P(X \mid h, \lambda) + (1-\alpha) \log P(h)\} \quad (3)$$

We use EM (Expectation Maximum) algorithm to solve (3). Q function can be written as (4).

$$Q(h, \bar{h}) = \alpha \sum_{t=1}^{T} \sum_{i=1}^{M} \frac{p(x_t, i \mid h, \lambda)}{p(x_t \mid h, \lambda)} \log p(x_t, i \mid \bar{h}, \lambda) + (1-\alpha) T \log p(\bar{h}) \quad (4)$$

In which, *T* represents the length of testing speech, *M* is the number of mixture components of GMM, $p(x_t, i \mid h, \lambda)$ denotes the probability of $(x_t - h)$ on the *i*-th mixture component of λ, $p(x_t \mid h, \lambda)$ denotes the probability of $(x_t - h)$ on all mixture component of λ. The method is iterative, h is the iterative result of last time and $\bar{h}$ is the result of this time.

To maximize $Q(h, \bar{h})$, differentiating (4) with respect to $\bar{h}_j$ for each dimension and equating to zero, we obtain (5).

$$\bar{h}_j = \frac{\alpha \sum_{t=1}^{T} [\sum_{i=1}^{M} \frac{p(x_t, i \mid h)}{p(x_t \mid h)} \times \frac{(x_{tj} - \mu_{ij})}{\sigma_{ij}^2}] + (1-\alpha) \frac{T \mu_{hj}}{\sigma_{hj}^2}}{\alpha \sum_{t=1}^{T} [\sum_{i=1}^{M} \frac{p(x_t, i \mid h)}{p(x_t \mid h)} \times \frac{1}{\sigma_{ij}^2}] + (1-\alpha) \frac{T}{\sigma_{hj}^2}} \quad (5)$$

$\bar{h}_j$ is the *i*-th dimension of $\bar{h}$, $j = 1, 2, ..., L$, L is the dimension number of features. x_{tj} is the *j*-th dimension of x_t. μ_{ij} and σ_{ij}^2 are the *j*-th dimension of μ_i and σ_i^2 of GMM.

When we obtain the estimate value of channel bias h, the compensation method is

$$\hat{X} = X - \hat{h}_{MAP} \quad (6)$$

$\hat{X}$ are the feature vectors after channel compensation. Observing from (5), we can find that μ_h and Σ_h are unknown, so we should obtain the two parameters before channel compensation.

Setting the scale factor $\alpha = 1$, we get the iterative (7).

$$\bar{h}_j = \frac{\sum_{t=1}^{T} \sum_{i=1}^{M} [\frac{p(x_t, i \mid h)}{p(x_t \mid h)} \times \frac{(x_{tj} - \mu_{ij})}{\sigma_{ij}^2}]}{\sum_{t=1}^{T} \sum_{i=1}^{M} [\frac{p(x_t, i \mid h)}{p(x_t \mid h)} \times \frac{1}{\sigma_{ij}^2}]} \quad (7)$$

If we have *H* channels, we can obtain *H* estimate values of channel bias h using (7), denote as $\{\hat{h}_{M1}, \hat{h}_{M2}, ..., \hat{h}_{MH}\}$, and then get the estimate values of μ_h and Σ_h using (8) and (9).

$$\mu_h = \frac{1}{H} \sum_{k=1}^{H} \hat{h}_{Mk} \quad (8)$$

$$\Sigma_h = \frac{1}{H} \sum_{k=1}^{H} (\hat{h}_{Mk} - \mu_h)^2 \quad (9)$$

We define $N(\mu_h, \Sigma_h)$ as channel model.

Besides, it is ML (Maximum Likelihood) compensation technique when we use (7) to get channel bias h directly during testing.

B. Experiments and Discussion

Channels are represented by different microphones. Speech is recorded in the environment of office. It is quiet comparing with the real world, so that the system can ignore the effect of environment noise and be absorbed in the effect of channels, which lowers the complexity. But, it is not quiet absolutely in office, because there is also little noise, such as talking, the sound of computer fans, ring of telephone, and so on. These kinds of noise are ignored.

TABLE I
THE INFORMATION OF MICROPHONES

Type of Microphone	NO.
Carbon-button, special	1
Electric,special	2
Carbon-button, common	3, 4, 7
Electric, common	5, 6, 8

There are eight different microphones which are numbered 1~8. Microphone No.1 is the standard microphone which is used to record training speech of speaker, all of the channel biases are obtained from non-standard microphones' speech comparing with standard microphone's speech. Microphone

No.7 and No.8 are treated as open-set microphones, so they are not used to record speech for training channel model. Table I is the information of eight microphones.

The speech using to train channel model: we record 10 minutes speech from one speaker using recorder, then play by the recorder, and record it again through microphone No.1~6 respectively. In this way, the last speech we obtained is the same in content and speaker, the only difference is the microphone we used.

There are 20 speakers who are all males and their ages are between 20 and 40. The training speech and testing speech of speakers are showed in Table II.

TABLE II
THE TRAINING SPEECH AND TESTING SPEECH OF SPEAKERS

NO.	Training Speech	Testing Speech
1	20 speakers 5 minutes/speaker	20 speakers, 1.5 minutes/speaker
2~6		20 speakers, 1.5 minutes/speaker
7~8		11 speakers, 2 minutes/speaker

The window function is Hamming window. The frame length is 25 milliseconds and frame overlapping length is 12.5 milliseconds. The number of mixture components in GMM is 32 and the length of training speech is 5 minutes. All the speaker models are trained by the speech recorded through microphone NO.1.

Each testing file is cut into pieces, and the length of each piece is about 10 seconds. There is only one speaker in a piece and the system will give one result for a piece. The system selects the beginning 1~6 seconds speech of each testing file (the whole testing file, not a piece) to get the channel bias for MAP, ML technique and the cepstral mean for CMS technique, so we regard these 1~6 seconds speech as adapted speech. The experiments are closed-set test among 20 speakers. Setting $\alpha = 0.5$.

Table III is the comparison of correct rate of baseline system and these systems using CMS, ML or MAP when channels are matching. The testing speech is recorded by microphone NO.1 too. There are 180 results altogether.

TABLE III
COMPARISON OF CORRECT RATE OF BASELINE SYSTEM AND THESE SYSTEMS USING CMS, ML OR MAP WHEN CHANNELS ARE MATCHING

	The Length of Adapted Speech					
	1 sec	2 secs	3 secs	4 secs	5 secs	6 secs
Baseline	88.9%					
CMS	97.2%	98.3%	98.9%	98.9%	98.9%	98.9%
ML	97.8%	98.3%	98.9%	98.9%	98.9%	98.9%
MAP	97.8%	98.3%	98.9%	98.9%	98.9%	98.9%

Table III shows that the three techniques can all improve the system when channels are matching. MAP and ML have the same performance, they are both better than CMS. As the length of adapted speech growing, the system correct rate is not varying.

Table IV is the comparison of correct rate of baseline system and these systems using CMS, ML or MAP when channels are mismatching. The testing speech is recorded by microphone NO.2~6. There is 1007 results altogether.

TABLE IV
COMPARISON OF CORRECT RATE OF BASELINE SYSTEM AND THESE SYSTEMS USING CMS, ML OR MAP WHEN CHANNELS ARE MISMATCHING

	The Length of Adapted Speech					
	1 sec	2 secs	3 secs	4 secs	5 secs	6 secs
Baseline	50.8%					
CMS	84.9%	87.5%	85.9%	87.9%	88.2%	87.4%
ML	88.1%	88.5%	88.6%	89.3%	89.6%	89.2%
MAP	89.1%	89.0%	89.3%	89.1%	89.8%	89.4%

When channels are not matching, MAP is better than CMS and ML because of the prior pdf of channel bias. The effect is obvious when the length of adapted speech is 1 second, MAP is 4.2% better than CMS and 1% better than ML. As the length of adapted speech growing, the effect of MAP is approaching to ML and still better than CMS. The system has the best performance of 89.8% when the length of adapted speech is 5 seconds.

Table V is the comparison of correct rate of baseline system and these systems using CMS, ML or MAP when testing under open-set microphones. The testing speech is recorded by microphone NO.7 and NO.8. There are 296 results altogether of 11 speakers.

When these systems testing under open-set microphones, the channels are mismatching and the open-set microphones are not used to record speech for training channel model. In Table V, MAP is much better than CMS. MAP is also little better than ML, because microphone NO.7, NO.1, NO.3, NO.4 are carbon-button microphone, microphone NO.8, NO.2, NO.5 and NO.6 are electric microphone, the channel model which is trained by speech from microphones NO.1~6 includes partly the information of microphone NO.7 and NO.8. But there is less information, so the result is worse than the result in Table 4. ML technique is the best choice when we have not channel model.

TABLE V
COMPARISON OF CORRECT RATE OF THESE SYSTEMS USING CMS, ML OR MAP WHEN TESTING UNDER OPEN-SET MICROPHONES

	The Length of Adapted Speech					
	1 sec	2 secs	3 secs	4 secs	5 secs	6 secs
Baseline	40.2%					
CMS	71.6%	80.4%	82.8%	83.4%	87.2%	88.2%
ML	86.8%	88.5%	88.9%	89.2%	88.5%	90.2%
MAP	87.5%	89.5%	88.5%	88.2%	89.5%	90.5%

IV. THE SECOND STAGE RECOGNITION BASED ON PITCH-DEPENDENT FEATURES

A. The Selection Strategy Based on GMM Scores

After the first stage recognition based on MAP channel compensation technique, the system has a good performance already. Considering from recognition speed and performance, the system needn't and can't recognize again for all of the testing speech (especially, the result of the first stage recognition is already right). So it needs to apply some strategy to select the testing speech which will be recognized again. The principle is selecting more testing speech whose result is wrong and less testing speech whose result is already right.

Experiments show that the real speaker of testing speech is one of the top 5 candidates, and its GMM score is closed to the GMM score of the first candidate, when the result of the testing speech is wrong. So the system selects testing speech and a few speaker models for the second stage recognition by score threshold and score order threshold. Firstly, the system uses score threshold to decide if one testing speech needs to be recognized again, namely the testing speech needs the second stage recognition if the difference of GMM scores between the first candidate and the second candidate is lower than score threshold. Secondly, the system allows the top N candidates to take part in the second stage recognition, N is the score order threshold. In experiments, score threshold is 30 and score order threshold is 5.

It is important that the selection strategy cannot ensure the selected testing speech is not doubly wrong and the remainder testing speech is not doubly right.

B. Piecewise Linear Model

The second stage recognition is based on pitch which represents the characters of glottis and is robust for channel. But it is difficult to get the exact value of pitch, so we expect to capture the dynamic information of pitch. The pitch value is not change suddenly when a person is talking, so the pitch values of neighboring frames are closed to each other. Sönmez stylized pitch contour by piecewise linear model [10].

In order to obtain the piecewise linear model, the system divides the speech into segments. There are a lot of zero values in the raw pitch values, so the system considers in two aspects to get the segments. The first aspect is zero values segment, if there are three continuous zero values, it should be segmented here. The second aspect is the difference of pitch values between neighboring frames, if the difference is higher than the threshold, it should be segmented here too, setting the threshold 20 in experiments.

The system stylizes each segment using linear model respectively. There are K segments in one speech and the *k*-th segment is from the T_{1k} -th to T_{2k} -th frame, so the piecewise linear model is

$$g(t) = \sum_{k=1}^{K} (a_k t + b_k) \tag{10}$$

Where, t is the serial number of frames. a_k and b_k are the slope and intercept of the *k*-th linear model and estimated by minimizing the MSE (Mean Square Error), namely

$$(a_k, b_k) = \arg\min \{ \sum_{t=T_{1k}}^{T_{2k}} (f_0(t) - g(t))^2 \} \tag{11}$$

C. Pitch-Dependent Features and Pitch-Dependent Speaker Model

The system uses the new pitch values computing from the piecewise linear model to replace the raw pitch values.

For each segment, the system extracts three dimensions pitch-dependent features including segment median, segment slope and segment duration.

The system uses three dimensions pitch-dependent features to train pitch-dependent speaker model.

(1) segment median $\log(\bar{f}_0)$

The former researchers assume the distribution of each speaker pitch is normal distribution when recognize speaker using pitch[10], but Sönmez found that the distribution of log pitch is closed to normal distribution comparing to the distribution of pitch [11], so the system lets $\log(\bar{f}_0) \sim N(\mu_0, \sigma_0^2)$.

(2) segment slope $\bar{f}_0'$

The system lets segment slope $\bar{f}_0'$, which is the slope of the *k*-th linear model, be modeled by normal distribution [12], namely $\bar{f}_0' \sim N(\mu_1, \sigma_1^2)$.

(3) segment duration T_s

The system lets segment duration T_s be modeled by shifted exponential distribution [12], namely $(T_s - \tau_0) \sim E(\theta)$.

The pitch-dependent speaker model is defined as

$\{\mu_0, \sigma_0^2, \mu_1, \sigma_1^2, \tau_0, \theta\}$

The system trains one pitch-dependent speaker model for each speaker during training, so the new speaker model includes one pitch-dependent speaker model and one GMM.

D. The Second Stage Recognition Based on Fusion

In the second stage recognition, the system extracts three dimensions pitch-dependent features for each testing speech to get the sequence of pitch-dependent feature vectors. The system makes matching computation between pitch-dependent features and pitch-dependent model, so that the system can obtain three pitch-dependent results for each selected speaker who will take part in the second stage recognition. The second stage recognition is based on the four results including three pitch-dependent results and one GMM score.

(1) Matching Computation

The system applies two ways of matching computation to judge how close the pitch-dependent feature vectors to one pitch-dependent model.

Score

The system computes three likelihood scores on pitch-dependent model for the pitch-dependent feature vectors. The bigger the likelihood score is, the closer the pitch-dependent feature vectors to the pitch-dependent model.

Divergence

Divergence is used to judge how close one distribution to the other. The smaller the divergence is, the closer one distribution to the other. There are two distributions $p(x)$ and $q(x)$, the standard formula to compute divergence is

$$d(p,q) = \int p(x)\log\frac{p(x)}{q(x)}dx + \int q(x)\log\frac{q(x)}{p(x)}dx \quad (12)$$

For normal distribution, we can get another formula (13)

$$d(p,q) = 0.5tr(\Sigma_p^{-1}\Sigma_q + \Sigma_q^{-1}\Sigma_p - 2I) + 0.5(\mu_p - \mu_q)^T(\Sigma_p^{-1} + \Sigma_q^{-1})(\mu_p - \mu_q) \quad (13)$$

In (13), $p \in N(\mu_p, \Sigma_p)$, $q \in N(\mu_q, \Sigma_q)$, $tr()$ is the trace of a matrix. I is the unit matrix. It's important that (13) is only applied to normal distribution.

The system trains the testing pitch-dependent model using the testing pitch-dependent feature vectors, and then computes divergence to judge how close the testing pitch-dependent model to each selected pitch-dependent speaker model which will be used to the second stage recognition. The system uses (13) to compute divergence of $\log(\bar{f}_0)$ and $\bar{f}_0'$. For segment duration T_s, its divergence can be computed using (14).

$$d(p,q) = \sum_x p(x)\log\frac{p(x)}{q(x)} + \sum_x q(x)\log\frac{q(x)}{p(x)} \quad (14)$$

(2) Results Fusion

To obtain better results, the system uses ANN to combine the four results including one GMM score and three pitch-dependent results. The ANN is three-layer BP (Back Propagation) network which includes 20 input nodes, 10 hidden nodes and 5 output nodes. The information of input and output as follows.

Input

20 input nodes, namely the top 5 candidates of the first stage recognition, each candidate have 4 results, 20 results altogether. Because of the big difference among the 4 results for each candidate, the system makes the 4 results to unit respectively before inputting to ANN.

The divergence results should be treated specially, because the smaller the divergence is, the closer one distribution to the other. So the system should get their opposite numbers.

Output

5 output nodes represent the top 5 candidates of the first stage recognition respectively.

The training process is a supervised training process. During training, if the testing speech belongs to some candidate, the value of output node relevant to this candidate is +1, contrariwise, the value is –1, and in this way there is one node whose value is +1 in all of the top 5 candidates at best. If the testing speech belongs to the speaker who is not in the top 5 candidates, the values of the 5 output nodes are all –1, the system won't use this kind of testing speech to train ANN.

The system uses 1/3 samples of all samples to train ANN and all samples to test. In order to avoid local minimum, the system uses 1/5 samples of all testing samples to validate. The training process stops when the performance of ANN for validation samples declines obviously. The best training result is the final training result. So the real testing samples are 4/15 samples of all samples.

During testing, the system uses the training result of ANN to combine the results, and the result of the second stage recognition is the candidates relevant to the biggest value of output node.

E. Experiments and Discussion

The speaker model is GMM which has 32 mixture components plus pitch-dependent speaker model. The length of training speech is 5 minutes. Closed-set test includes 20 speakers. The testing speech is recorded by microphone NO.2~6. There is 1007 results altogether. Setting $\alpha = 0.5$.

Table VI is the comparison of correct rate of one-stage recognition system based on MAP or two-stage recognition system based on MAP and pitch-dependent features when channels are mismatching.

TABLE VI
COMPARISON OF CORRECT RATE OF ONE-STAGE RECOGNITION SYSTEM BASED ON MAP OR TWO-STAGE RECOGNITION SYSTEM BASED ON MAP AND PITCH-DEPENDENT FEATURES WHEN CHANNELS ARE MISMATCHING

The Length of Adapted Speech	One-stage recognition	Matching computation	Two-stage recognition
1 second	89.1%	Score	91.5%
		Divergence	90.9%
3 seconds	89.3%	Score	91.8%
		Divergence	91.5%
6 seconds	89.8%	Score	92.5%
		Divergence	92.1%

Table VI shows that the improvement of two-stage recognition system based on pitch-dependent speaker model and ANN fusion is obvious, because the second stage recognition reduces the number of wrong results which are brought by channel nonlinear mismatch, this compensates the disadvantage of MAP only compensating channel linear mismatch. The matching computation of score is better than divergence, because the number of segments in testing speech (about 30) is much less than the number in training speech (about 2000). When the system recognizes by the matching computation of divergence, the testing pitch-dependent speaker model can't represent the real model, so its effect is worse than score's.

Using the matching computation of score, when the length of adapted speech is 1 second, 3 seconds, 5 seconds, the performance are 2.4%, 2.5%, 2.7% better than one-stage recognition system respectively. The system has the best performance of 92.5% when the length of adapted speech is 5 seconds.

V. CONCLUSION

This paper focuses on improving channel robustness of

speaker recognition system in two aspects of channel compensation technique and channel robust features. In the first aspect, this paper applies MAP channel compensation technique, which was used in speech recognition, to speaker recognition system. In the second aspect, this paper introduces pitch-dependent features and pitch-dependent speaker model to recognize again, and then uses ANN to combine the three pitch-dependent results and one GMM score for getting a fusion result. The system makes the second stage recognition based on these fusion results.

This paper can get the conclusions as follows.

(1) MAP channel compensation technique compensates channel linear mismatch better. MAP, ML and CMS all improve the system when channels are matching. Because of the prior pdf of channel bias, MAP is better than CMS and ML, especially under less adapted speech, which is useful in real-time recognition system.

(2) When the prior pdf of channel bias does not exist, ML compensates channel linear mismatch better than CMS because MAP can't be used.

(3) Three dimensions pitch-dependent features represent the speaker-dependent information and have the advantage of channel robustness.

(4) ANN unites the three pitch-dependent results and GMM score effectively. The improvement of two-stage recognition system based on pitch-dependent speaker model and ANN fusion is obvious, because the second stage recognition reduces the number of wrong results which are brought by channel nonlinear mismatch, this compensates the disadvantage of MAP only compensating channel linear mismatch.

Although the system this paper used have a good performance, there are lots of work should be researched in depth. In order to keep on improving the system, we should obtain more speech for training channel model and use more better approach to compensate channel nonlinear mismatch. Although the results of this paper can't be compared to other techniques except ML and CMS, because they aren't be validated by standard speech library, these results are still worthy to be referenced.

REFERENCES

[1] D. A. Reynolds, "Channel Robust Speaker Verification via Feature Mapping," in Proc. of ICASSP'03, Hong Kong, 2003,pp.53-56.
[2] B. S. Atal, "Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identification and Verification," Journal of the Acoustical Society of America. Vol. 55, no.6, pp.1304-1312, 1974.
[3] H. Hermansky, N. Morgan, "RASTA Processing of Speech," IEEE Speech And Audio Processing, Vol.2, no.4, pp.578-589, 1994.
[4] S. Furui, "Cepstral Analysis Technique for Automatic Speaker Verification," IEEE, ASSP, Vol.29, no.2, pp.254-72, 1981.
[5] J. Chien, H. Wang, L. Lee, "Estimation of Channel Bias for Telephone Speech Recognition," in Proc. of ICSLP, 1996, pp.1840-1843.
[6] Teunen R, Shahshahani B, Heck L, "A Model-based Transformational Approach to Robust Speaker Recognition," in Proc. of ICSLP, 2000, pp.495-498.
[7] D. A. Reynolds, "The Effect of Handset Variability on Speaker Recognition Performance: Experiments on the Switchboard Corpus," in Proc. of ICASSP, 1996, pp.113-116.
[8] R. Auckenthaler, M. Carey, H. Lloyd-Thomas, "Score Normalization for Text-independent Speaker Verification System," Digital Signal Processing, vol.10, no.1, 2000.
[9] D. A. Reynolds, W. Andrews, J. Campbell, J. Navratil, B. Peskin, A. Adami, Q. Jin, D. Klusacek, J. Abramson, R. Mihaescu, J. Godfrey, D. Jones, B. Xiang, "The SuperSID Project: Exploiting High-level Information for High-accuracy Speaker Recognition," in Proc. of ICASSP'03, Hong Kong, 2003, pp. 784-787.
[10] K. Sönmez, E. Shriberg, L. Heck, M. Weintraub, "Modeling Dynamic Prosodic Variation for Speaker Verification," in Proc. of ICSLP, 1998, pp.3189-3192.
[11] M. J. Carey, E. S. Parris, H. Lloyd-Thomas, S. Bennett, "Robust Prosodic Features for Speaker Identification," in Proc. of ICSLP, 1996, pp.1800-1803.
[12] M. K. Sönmez, L. Heck, M. Weintraub, E. Shriberg, "A Lognormal Tied Mixture Model of Pitch for Prosodybased Speaker Recognition," in Proc. of Eurospeech, 1997, pp.1391-1394.

Jiqing Han (M'04) was born in August 1964, and received the Ph.D. degree in Computer Science and Engineering, Harbin Institute of Technology, Harbin, China, in 1998. He became a Member of **IEEE** in 2004.

He is a professor in School of Computer Science and Technology, Harbin Institute of Technology. His interests include speech recognition, signal processing and pattern recognition.

Prof. Han is both a senior member of Acoustics Society of China and China Computer Federation. He is also a member of editorial board of Journal of Chinese Information Processing.