Far-Field Speaker Recognition

Qin Jin, Tanja Schultz, and Alex Waibel

Abstract—**In this paper, we study robust speaker recognition in far-field microphone situations. Two approaches are investigated to improve the robustness of speaker recognition in such scenarios. The first approach applies traditional techniques based on acoustic features. We introduce reverberation compensation as well as feature warping and gain significant improvements, even under mismatched training–testing conditions. In addition, we performed multiple channel combination experiments to make use of information from multiple distant microphones. Overall, we achieved up to 87.1% relative improvements on our Distant Microphone database and found that the gains hold across different data conditions and microphone settings. The second approach makes use of higher-level linguistic features. To capture speaker idiosyncrasies, we apply n-gram models trained on multilingual phone strings and show that higher-level features are more robust under mismatching conditions. Furthermore, we compared the performances between multilingual and multiengine systems, and examined the impact of a number of involved languages on recognition results. Our findings confirm the usefulness of language variety and indicate a language independent nature of this approach, which suggests that speaker recognition using multilingual phone strings could be successfully applied to any given language.**

Index Terms—**Far-field microphones, mismatched conditions, multilingual phone strings, robust speaker recognition.**

I. Introduction

AUTOMATIC speaker recognition has developed into an important technology and is lately perceived to be crucial to various speech-aided applications. Traditional approaches, such as Gaussian mixture models (GMMs) [4] achieve very high accuracies for speaker identification and verification tasks on high-quality data when training and testing conditions are well controlled. However, real-world speech-aided applications require handling a large variety of speech signals, corrupted by adverse environmental conditions (noise, background, channel), and mismatched training–testing conditions, i.e., scenarios in which speaker models were trained in one signal condition but deployed in different conditions.

GMM-based systems are known to degrade significantly under adverse and mismatched conditions, often below the threshold of being useful to real-world applications. This degradation becomes even more severe when the speech signals are captured from the distance [5]. However, while far-field *speech* recognition has been investigated for some time, far-field *speaker* recognition has not received much attention yet.

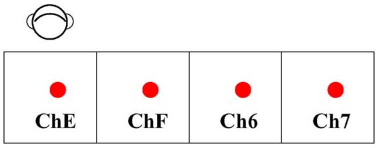


Fig. 1. Distant table microphone setup in ICSI meetings.

This paper describes our efforts to improve the performance of speaker recognition in far-field situations, focusing on meeting scenarios that are recorded with multiple distant microphones. We tackle this problem by two approaches that differ in the level of applied linguistic knowledge. Our low-level approach follows the traditional GMM techniques, applying acoustic features. Here, we introduced reverberation compensation, feature warping, and multiple channel combination techniques to alleviate the issues of acoustic mismatches. In the high-level approach, we explore the potential of phonetic speaker information and speaker pronunciation idiosyncrasy.

II. Databases

We use three distant microphone databases that differ in microphone positioning, room characteristics, and speaking style for evaluation, in order to demonstrate the robustness of our approaches and to compare results across sites.

A. ICSI Meeting Database

The ICSI Meeting Database [3] contains 75 meetings with simultaneous multichannel audio recordings collected at the International Computer Science Institute (ICSI) in Berkeley, CA. We selected 24 out of a total of 53 speakers for training and testing based on the speakers' position with respect to the microphone and their total speaking time. Fig. 1 illustrates the distant table microphone arrangement in the ICSI meeting room and indicates the position of the selected speakers. The table microphones are desktop omnidirectional pressure zone microphones (PZMs). They were arranged in a staggered line along the table center. We randomly selected 90 s of speech per speaker from the meetings as training data. The remaining speech was used for testing, which leads to 397 test trials in total for all speakers. Test segments longer than 20 s were split into 20-s chunks.

B. 2-D Distant Microphone Database

The 2-D Distant Microphone database (2-D DMD) collected at the Interactive Systems Labs in 2000 [14] contains recordings from microphones at various distances. Similar to the ICSI setup, the microphones were arranged in a staggered line along the table center. However, there were eight microphones used in total, one close-talking Sennheiser headset microphone (Dis0), one Lapel microphone (DisL) attached to the speaker's lapel,

Manuscript received February 16, 2007; revised May 15, 2007. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jean-François Bonastre.

The authors are with the Interactive Systems Laboratories, Language Technologies Institute, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213 USA (e-mail: qjin@cs.cmu.edu; tanja@cs.cmu.edu; ahw@cs.cmu.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2007.902876

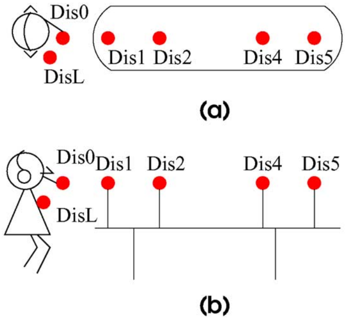


Fig. 2. Microphone setup in 2-D DMD collection.

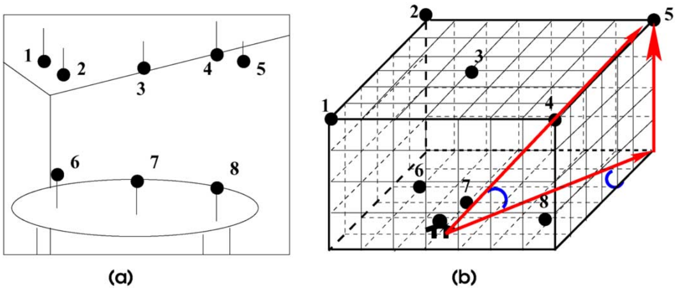


Fig. 3. Microphone setup in 3-D DMD collection.

and six additional Lapel microphones, mounted on microphone stands positioned at distances of 1 ft (Dis 1), 2 ft (Dis 2), 4 ft (Dis 4), 5 ft (Dis 5), 6 ft (Dis 6), and 8 ft (Dis 8) to the speaker, respectively.

Fig. 2 displays the microphone distance arrangement with respect to the speaker (top) and that all microphones are positioned in the same vertical space, but not the same horizontal one (bottom), hence the name "2-D" database.

The database consists of 30 speakers (16 female, 14 male). Each speaker recorded five sessions using all eight microphones in parallel. The speakers sit at the table and read articles. The articles were different for each session. For each speaker we randomly select 60 s from the first session as training data. The remaining data was split into 20-s segments and used as test trials, summing to 60 test trials.

C. 3-D Distant Microphone Database

We collected a second Distant Microphone Database (3-D DMD) [14] to investigate robust speaker recognition with distant microphones arranged in the 3-D space. Five microphones (labeled as 1 to 5) are hanging from the ceiling, while three microphones (6, 7, and 8) are mounted on the meeting table, as depicted in Fig. 3 (left-hand side). The right hand-side of Fig. 3 illustrates the positioning of these eight microphones with respect to the speaker. The cubical grid defines the distances of the microphones to the speaker. A grid unit corresponds to 0.5 m. The recording room of the 3-D DMD was quieter and slightly smaller than in the 2-D DMD setup.

For the recordings, we used miniature cardioid condenser microphones that are very similar to omnidirectional microphones. Since the speaker (sound source) is not omnidirectional, the microphones will receive different signals even when they have the same Euclidean distance to the speaker. Therefore, we defined the distance between the microphone (receiver) and the speaker (sound source) to be the Euclidean grid distance, which is penalized by both the horizontal and vertical angles between the speaker and the microphone. For example, the distance of channel 5 is computed as

$$D(5) = \frac{\sqrt{4^2+5^2+4^2}}{\cos\left(\arctan\left(\frac{4}{5}\right)\right)\cos\left(\arctan\left(\frac{4}{\sqrt{4^2+5^2}}\right)\right)} = 11.4 \qquad (1)$$

which is the Euclidean distance in both horizontal and vertical planes divided by the cosine values of the angle in horizontal plane and vertical plane, respectively. For Dis 6, Dis 7, and Dis 8, no vertical angle penalty was applied since the speaker is positioned in the same horizontal plane as the table microphones. The distance calculations gave $D(1) = 14.5, D(2) = D(3) = D(8) = 10, D(4) = 12, D(5) = 11.4, D(6) = 4.3$, and $D(7) = 2$.

The 3-D DMD consists of 24 speakers (4 female, 20 male). We recorded one session per speaker, in which they were asked to talk about ten given topics in a spontaneous speaking style. The duration varies from 8 to 20 min. As training data, we randomly picked 2 min of speech from the first 80% of a speaker's recording. The remaining 20% of speech was split into 20-s segments, leading to a total of 183 test trials over all speakers. Although the fact that test and training data are taken from the same session may result in optimistic estimates, such experimental design allows us to focus on far-field effects. All test speakers are assumed to be enrolled in the system, i.e., we perform a closed-set speaker recognition task.

III. Feature Processing For Far-Field

A. Reverberation Compensation

Speech signals recorded by a distant microphone are more prone to be degraded by additive background noise and reverberation. Considering room acoustics as a linear shift-invariant system, the receiving signal $y(t)$ can be written as

$$y[t] = x[t] * h[t] + n[t] \qquad (2)$$

where the source signal $x[t]$ is the clean speech, $h[t]$ is the impulse response of the room reverberation, and $n[t]$ is the room noise. Cepstral mean subtraction (CMS) [2] has been successfully used to compensate the convolution distortion. In order for CMS to be effective, the length of the channel impulse response has to be shorter than the short-time spectral analysis window, which is usually 16–32 ms. Unfortunately, the duration of the impulse response of reverberation usually has a much longer tail, more than 50 ms. Therefore, traditional CMS will not be as effective under these conditions.

Following the work of Pan [19], we separate the impulse response $h[t]$ into two parts $h_1[t]$ and $h_2[t]$, where $h[t] = h_1[t] + \delta(t-T)h_2[t]$

$$h_1[t] = \begin{cases} h[t], & t < T \\ 0, & \text{otherwise} \end{cases}$$

$$h_2[t], = \begin{cases} h[t+T], & t \geq 0 \\ 0, & \text{otherwise.} \end{cases}$$

We rewrite formula (2) as $y[t] = x[t] * h_1[t] + x[t-T] * h_2[t] + n[t]$. $h_1[t]$is a much shorter impulse response with length smaller than the DFT analysis window; thus, it can be compensated by the conventional CMS. We treat $x[t-T]*h_2[t]$ similar to additive noise $n[t]$ and apply the noise reduction technique based on spectrum subtraction. Assuming the noise $x[t-T]*h_2[t]+n[t]$ could be estimated from $y[t-T]$, the spectrum subtraction is performed as

$$\hat{X}[t,\omega] = \max(Y[t,\omega] - a\cdot g(\omega)Y[t-T,\omega], b\cdot Y[t,\omega])$$

where a is the noise overestimation factor, and b is the spectral floor parameter to avoid negative or underflow values. We can empirically estimate the optimum a, b, and $g(\omega)$ on a development dataset. We found that the system performance is not sensitive to T. Within the range of 20–40 ms there is no significant difference on the effect of the spectra subtraction. However, outside that range, there is obvious performance degradation. For the recording setup, we found $a = 1.0, b = 0.1$, and $g(\omega) = |1-0.9e^{j\omega}|$ to be optimal in most changing conditions, based on development data described in [19]. Standard CMS was applied after spectrum subtraction to eliminate the effect of $h_1[t]$.

B. Feature Warping

The feature warping method applied here [1] warps the distribution of a cepstral feature stream to a standardized distribution over a specified time interval. The warping can be considered as a nonlinear transformation $\mathcal{T}$, which transforms the original feature X to a warped feature $\hat{X}$, i.e., $\hat{X} = \mathcal{T}(X)$. This can be implemented by cumulative distribution function (CDF) matching as described in [6], which warps a given feature such that its CDF matches a desired distribution, for example the normal distribution. This method assumes that the dimensions of the Mel frequency cepstral coefficient (MFCC) vectors are independent. Therefore, each dimension is processed as a separate stream. The CDF matching is performed over short time intervals by shifting a window, and only warping the central frame of that window. The warping executes as follows:

- For $i = 1, \ldots, d$ where d is the feature dimension.
- Sort features of dimension i in ascending order in a given window.
- Warp the raw feature value x in dimension i of the central frame to its warped value $\hat{x}$, which satisfies $\phi = \int_{-\infty}^{\hat{x}} f(y)dy$, where $f(y)$ is the probability density function (PDF) of the standard normal distribution, i.e., $f(y) = (1/\sqrt{2\pi})\exp(-(y^2/2))$ and ϕ is its corresponding CDF value. Suppose x has a rank r and the window size is N. Then, the CDF value can be approximated as $\phi = ((r-(1/2))/N)$.
- Find $\hat{x}$ by lookup in a standard normal CDF table.

In our experiments, we set the window size to 300 frames and the window shift to one frame. Zeros are padded at the beginning and at the end of the raw feature stream.

C. Experiments on Noise Compensation

Throughout this paper, the system performance is measured using closed-set speaker identification accuracy, which corresponds to the percentage of correctly recognized test trials over all test trials. This GMM-based system was evaluated in NIST CLEAR 06 and CLEAR 07 evaluations and achieved very competitive results.

TABLE I
BASELINE SYSTEM (BASELINE-CMS) PERFORMANCE (%) ON 3-D DMD

Test Train	Ch1	Ch2	Ch3	Ch4	Ch5	Ch6	Ch7	Ch8
Ch1	**95.6**	94.0	76.0	83.6	72.7	77.6	71.6	83.1
Ch2	61.2	**100.0**	86.3	70.0	84.2	94.0	89.1	88.0
Ch3	38.3	63.4	**98.4**	49.2	59.0	71.6	78.7	78.7
Ch4	71.0	83.1	70.5	**87.4**	59.6	83.1	77.6	84.2
Ch5	54.1	86.9	76.0	59.6	**91.8**	85.3	84.7	84.7
Ch6	49.2	77.1	78.1	47.0	76.5	**90.7**	90.7	76.0
Ch7	38.8	68.9	75.4	52.5	72.1	86.3	**92.9**	80.9
Ch8	62.8	85.3	78.1	65.0	86.9	85.3	89.6	**95.1**

The front-end processing relies on 13-dimensional MFCC signal analysis every 10 ms. Noninformative frames are removed by performing speech detection based on normalized energy, with an empirically set threshold that is applied to all microphone channels. The mean feature vector in CMS is computed on the informative frames only, all noninformative frames are discarded from training as well as from testing. The *baseline system (baseline-CMS)* consists of the following components: 1) energy-based speech detection, 2) 13-dimensional MFCC front-end processing and CMS, and 3) 128-mixture GMMs per speaker model trained with the EM algorithm. The "*improved baseline system* (RC + Warp + CMS)" varies from the baseline system only in applying reverberation compensation (RC) and feature warping (Warp) in addition to CMS in the front-end processing.

1) Results of the Baseline System: Table I presents the baseline speaker recognition accuracies on the 3-D DMD. Rows refer to training channels and columns refer to test channels. Bold numbers indicate that accuracies of matched training-testing conditions (i.e., speaker models trained and tested on the same channel) are much higher than accuracies of mismatched conditions (off the diagonal cells). From this table, we also observe large differences in recognition results among training and test pairs. For example, the performance for training on CH1 and testing on CH2 (94%) differs a lot from the performance when training on CH2 and testing on CH1 (61.2%). This is expected as the quality of test speech has a large impact on the system performance. Since microphone 1 is hanging from the ceiling behind the speaker (see Fig. 3), it receives more reverberations, therefore the CH 1 signal is more distorted. Also, CH 1 has the largest distance according to our distance definition in Section II-C. Consequently, the average performance on CH 1 is the worst among all the channels as shown in Fig. 4.

Fig. 4 shows the correlation between recognition accuracy and channel distances on the 3-D DMD. Apparently, the performance is a function of the distance value: after surpassing a critical distance (mic 1, 4, 5) the performance decreases significantly. The worst performance is achieved for the two ceiling microphones, which are mounted in the back of the speaker.

Fig. 5 summarizes for each test channel the performance of the baseline system (baseline-CMS) under matched and mismatched conditions. The curve for the matched condition corresponds to the bold numbers in Table I. The curve for mismatched conditions gives the average performance, calculated over each but the diagonal cell of a column in Table I. The bars of the curve for mismatched conditions in Fig. 5 refer to

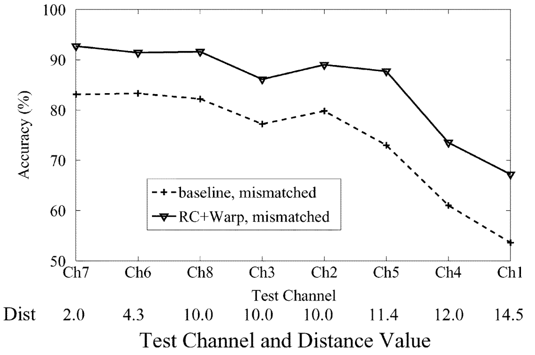


Fig. 4. Correlation between performance and distance on the 3-D DMD.

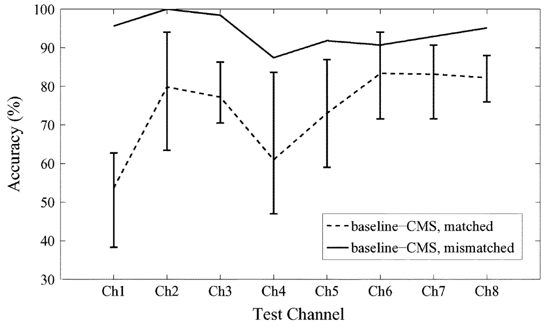


Fig. 5. Baseline performance (matched versus mismatched) on 3-D DMD.

TABLE II
RC AND WARP IMPACT ON 3-D DMD

System	Matched	Mismatched
baseline-CMS	94.0	74.2
RC+CMS (relative gain)	94.8 (13.3%)	78.1 (15.1%)
Warp+CMS (relative gain)	96.4 (40.0%)	79.1 (19.0%)
RC+Warp+CMS (relative gain)	96.7 (45.5%)	84.9 (41.6%)

the performance range. The average accuracy under matched conditions is 94.0% and compares to 74.2% for mismatched conditions, which proves that the system performance degrades severely under mismatched conditions. Furthermore, the range varies strongly between the different channels and tends to be smaller for the closer microphones.

2) Results From the Improved Baseline System: Table II shows the performance and relative improvements over the baseline results on the 3-D DMD when applying reverberation compensation, feature warping, and their combination. Each approach outperforms the baseline under both matched and mismatched conditions. Furthermore, the combination of both gives additional gains, indicating that the two techniques take care of different signal degradation effects.

Fig. 6 summarizes the gains achieved by reverberation compensation and feature warping (RC + Warp) on all three data sets. Significant improvements were achieved under both matched and mismatched conditions on all three data sets. On average, we achieved 45.5% and 41.6% relative improvements under matched and mismatched conditions on 3-D DMD, 20.0% and 17.7% on 2-D DMD, and 31.9% and 34.1% on the ICSI Meeting Database. These results confirm that the applied methods are robust under different microphone positionings and distances, as well as for different speaking styles. Consequently, we will use both techniques in all following experiments and refer to this improved system as "improved baseline (RC + Warp + CMS)." In the next section, we will investigate the concept of multiple channel combination for further improvements.

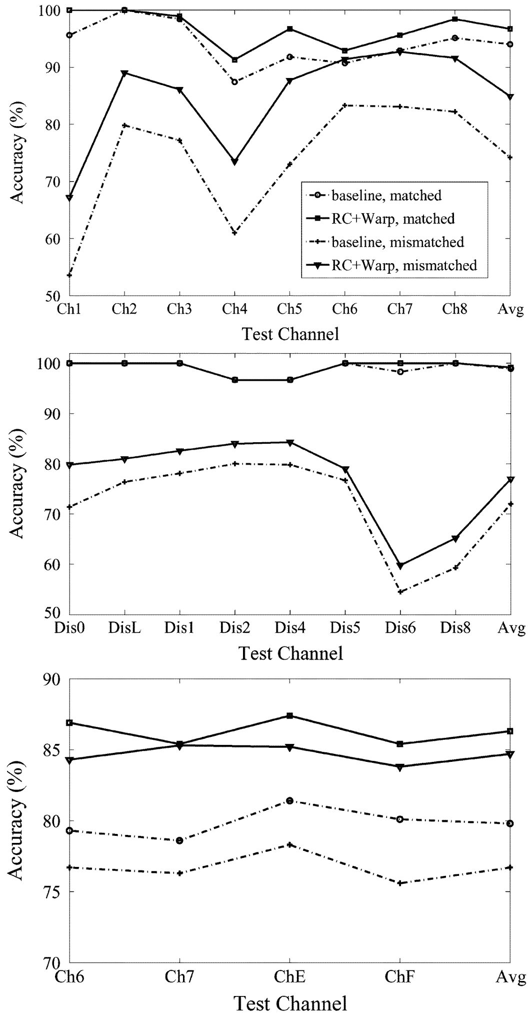


Fig. 6. Gains from RC + Warp on 3-D (top), 2-D (middle), and ICSI (bottom); all three figures share the same legend as shown in the top figure.

IV. MULTIPLE CHANNEL COMBINATION

The setup of hands-free multiple distant microphones is cheap and easy compared to arrangements such as microphone arrays and becomes common practice in applications like meetings and lectures. In order to benefit from such a multiple channel setup, we investigated four approaches to combine information from multiple channels.

A. Data Combination (DC)

In the DC approach, speaker models are trained by combining data from multiple channels, here from all mismatching channels, i.e., for the evaluation on channel 1, the speaker models are trained on all but channel 1 data. In the 3-D DMD case this, would lead to an increase of training data by a factor of 7 (3-D DMD has eight channels). However, since we aim to study the effect of data variety rather than data volume, we limited the training data to the same size as in the matched condition. In case of 3-D DMD, the training data for the speaker on channel 1 (CH 1) is formed by randomly selecting a fraction of $(1/7)$th data from the mismatched channels (CH 2 to CH 8).

B. Frame-Based Score Competition (FSC)

In the FSC approach, we compute the likelihood of observing a speaker based on a set of GMM models, where the term "score" refers to the log likelihood. Before explaining the FSC approach, we briefly review the likelihood calculation and decision rules in a GMM-based system. The GMM Θ^k of speaker k is defined as a mixture of M Gaussian distributions $\Theta^k = (\lambda_m, \mathrm{N}(\mu_m, \Sigma_m)), m = 1, 2, \ldots, M$ and the identification decision is given by $s^* = \arg\max_k(\mathrm{LL}(X \mid \Theta^k)), k = 1, 2, \ldots, S$, where s^* is the recognized speaker identity, S is the total number of enrolled speakers, and $\mathrm{LL}(X \mid \Theta^k)$ is the log likelihood score that the entire test feature set X was generated by the GMM Θ^k.

The likelihood of an observation (here for feature vector x_n) given a GMM model Θ^k of speaker k is estimated as

$$p(x_n \mid \Theta^k) = \sum_{i=1}^{M} \frac{\lambda_i}{\sqrt{2\pi|\Sigma_i|}} \exp\left\{\frac{-(x_n - \mu_i)^T \Sigma_i^{-1} (x_n - \mu_i)}{2}\right\}. \tag{3}$$

The feature vectors X are assumed to be independent and identically distributed (i.i.d.). Accordingly, the likelihood of observation sequence X given Θ^k is estimated as

$$p(X \mid \Theta^k) = \prod_{n=1}^{N} p(x_n \mid \Theta^k) \tag{4}$$

$$\mathrm{LL}(X \mid \Theta^k) = \sum_{n=1}^{N} \log p(x_n \mid \Theta^k) = \sum_{n=1}^{N} \mathrm{LL}(x_n \mid \Theta^k). \tag{5}$$

Since a multiple microphone setup emits speech samples from multiple channels, we can build multiple GMM models for each speaker k, one for each channel i and refer to them as Θ^{k,Ch_i}. For a total number of C channels we get $\{\Theta^k = \Theta^{k,\mathrm{Ch}_1}, \ldots, \Theta^{k,\mathrm{Ch}_C}\}$ models for speaker k.

The idea of the FSC approach is to use the set of multiple GMM models rather than a single GMM model. In each frame, we compare the incoming feature vector of channel Ch_h to all GMMs $\{\Theta^{k,\mathrm{Ch}_1}, \ldots, \Theta^{k,\mathrm{Ch}_{h-1}}, \ldots, \Theta^{k,\mathrm{Ch}_{h+1}}, \ldots, \Theta^{k,\mathrm{Ch}_C}\}$ of speaker k but the GMM of the test channel Ch_h. The highest log likelihood score of all GMM models is chosen to be the frame score. Finally, the log likelihood score of the entire test feature vector set X from channel h is estimated as

$$\mathrm{LL}(X \mid \Theta^k) = \sum_{n=1}^{N} \mathrm{LL}(x_n \mid \Theta^k) = \sum_{n=1}^{N} \max\{\mathrm{LL}(x_n \mid \Theta^{k,\mathrm{Ch}_j})\}_{j=1, j\neq h}^{C}. \tag{6}$$

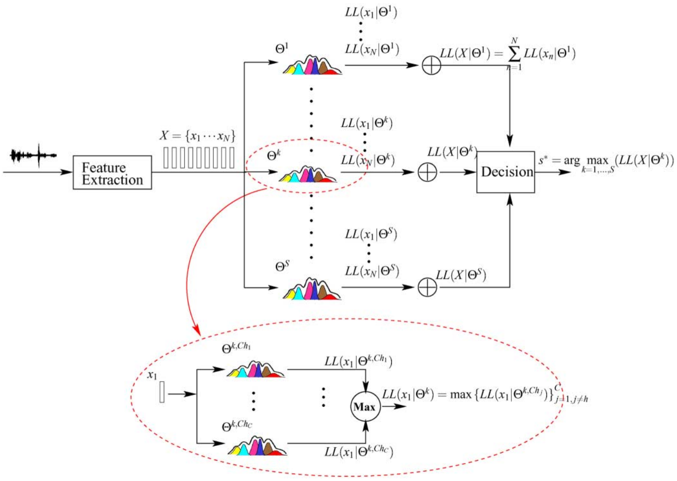


Fig. 7. Speaker recognition procedure with FSC.

This multiple-channel competition process differs from the one-channel standard scoring process in that the per-frame log likelihood scores are not necessarily derived from the same microphone. Fig. 7 illustrates how the standard procedure gets replaced by the circled portion to form the "frame-based score competition" speaker recognition approach.

C. Segment-Based Score Fusion (SSF)

The term "segment" refers to the entire test utterances, i.e., the complete set of test feature vectors X. In the "segment-based score fusion" approach, we compute the score of a test utterance from channel Ch_h given the set of C GMM models $\Theta^k = \{\Theta^{k,\mathrm{Ch}_1} \ldots \Theta^{k,\mathrm{Ch}_C}\}$ for speaker k by a fusion of all GMM models scores but the one from Θ^{k,Ch_h}

$$\mathrm{LL}(X \mid \Theta^k) = \sum_{j=1, j\neq h}^{C} w_j \cdot \mathrm{LL}(X \mid \Theta^{k,\mathrm{Ch}_j}). \tag{7}$$

The fusion weights w_j are simply set to be equal across channels.

D. Segment-Based Decision Voting (SDV)

In the "segment-based decision voting" approach, the entire set of feature vectors X is evaluated multiple times by particular speaker models that were trained on one mismatched channel. As a consequence, the speaker identity decision rule delivers $(C-1)$ identity values, one for each mismatched channel. The final decision for the speaker's identity is made by picking the identity value which appears the most often among the $C-1$ values. In case of a tie, we pick the one with the highest log-likelihood score.

E. Experimental Results on Mismatched Conditions

We investigated the performance of all proposed multiple-channel combination approaches on the three described data sets. Fig. 8 presents the performance improvements under mismatched conditions for 3-D DMD (top), 2-D DMD (middle) and

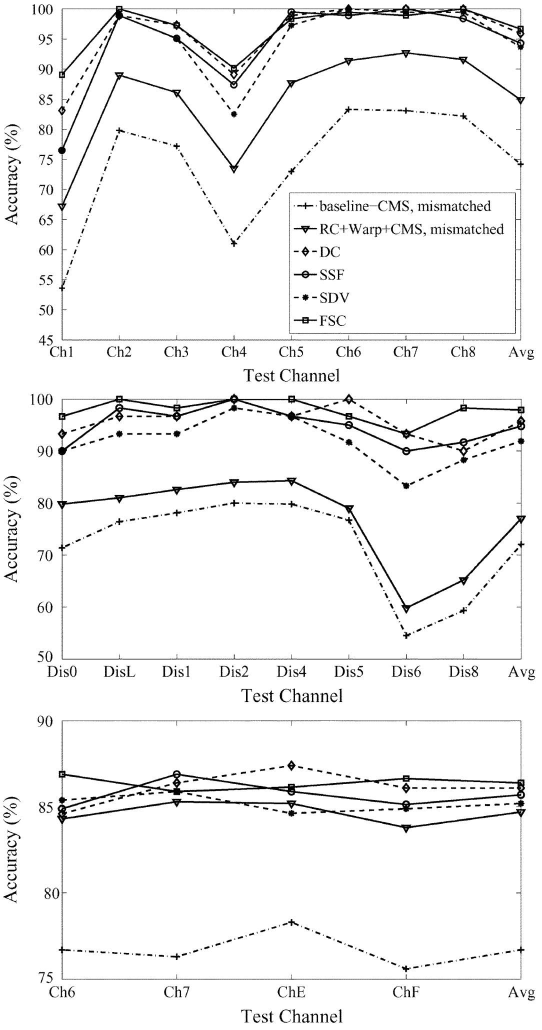


Fig. 8. Multiple-channel combination on 3-D (top), 2-D (middle), ICSI (bottom); all three figures share the same legend as shown in the top figure.

ICSI (bottom) and shows that significant improvements were achieved for all combination approaches.

TABLE III
RELATIVE IMPROVEMENT BY MULTIPLE-CHANNEL COMBINATION

Approach \ Data Combination	3D DMD	2D DMD	ICSI
Data Combination	72.8%	81.9%	9.7%
Frame based Score Competition	**77.8%**	**91.0%**	**11.4%**
Segment based Score Fusion	62.4%	77.4%	6.8%
Segment based Decision Voting	57.9%	64.7%	3.5%

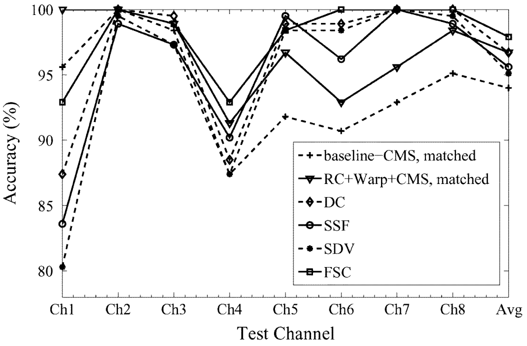


Fig. 9. Multiple-channel combination for all channels (3-D DMD).

On average, "Data Combination" gives 72.8% relative improvement over the improved baseline $(\mathrm{RC}+\mathrm{Warp}+\mathrm{CMS})$ and 84.1% relative improvement over the baseline (baseline-CMS), i.e., DC achieves additional gains over reverberation compensation and feature warping. Since we controlled the amount of data in the DC approach to be the same as in the baseline, the gains prove that more variability in training data significantly improves the robustness of speaker recognition. 77.8% relative improvement was achieved by the "frame-based score competition," 62.4% by the "segment-based score fusion" and 57.9% by "segment-based decision voting" over the improved baseline $(\mathrm{RC}+\mathrm{Warp}+\mathrm{CMS})$. This indicates that it is beneficial to use information from multiple sources, even though each single one is not very powerful. We also tested the DC approach using all data from all channels and observed 76.9% relative improvement over the improved baseline $(\mathrm{RC}+\mathrm{Warp}+\mathrm{CMS})$. Compared to the 72.8% above from the controlled data amount, this testifies that the significant improvement is mostly due to the larger variability rather than the larger amount in training data. Even higher relative gains are achieved on the 2-D DMD. We see 81.9%, 91.0%, 77.4%, and 64.7% relative improvements by DC, FSC, SSF, and SDV, respectively. On the ICSI Meeting Database, we got 9.7%, 11.4%, 6.8%, 3.5% relative improvement by DC, FSC, SSF, and SDC over the improved baseline $(\mathrm{RC}+\mathrm{Warp}+\mathrm{CMS})$ under mismatched conditions. The improvement differences among the three databases might be a result of the range of speaking styles (2-D is read speech, 3-D is spontaneous but monologue, ICSI is meeting style). The significantly smaller gains for ICSI might be due to the smaller number of channels, which makes channel combination less powerful. Table III summarizes these findings for all four multiple-channel combination approaches on the three databases and shows that the "frame-based score competition" approach achieves the highest improvements, while "segment-based decision voting" gives the lowest gains. We think the reason why the FSC approach works best among the multiple channel combination approaches is that FSC combines multiple sources at the finest granularity, i.e., at the frame level, while the other approaches combine at the segment level.

F. Matched and Mismatched Conditions

In the experiments above, we focused on the performance impact of multiple-channel combination approaches under mismatched-channel conditions. For this purpose, we use only data of mismatched channels. However, this does not imply that we always assume to have prior knowledge about the channel origin of the test signal. The intention was rather to prove that very high speaker recognition accuracies can be achieved, even when no data of particular microphone positions are available. In this section, we investigate if the performance can be further improved by including training material of the matched channel. Fig. 9 compares the performance of multiple-channel combination with the improved baseline $(\mathrm{RC}+\mathrm{Warp}+\mathrm{CMS})$ when all channel data including the matched channel are used for training. The results show that, on average, DC and FSC outperform the improved baseline, while SSF and SDV do not gain but stay within close range.

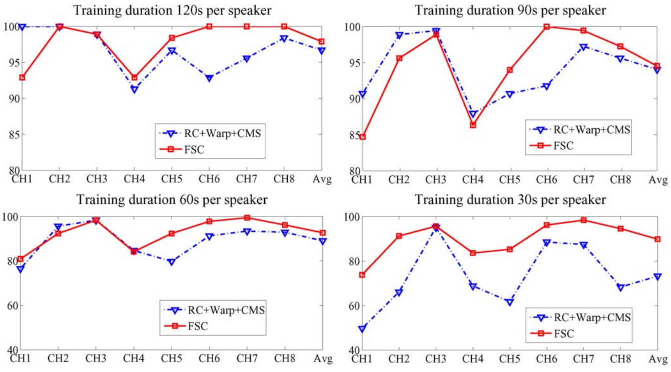


Fig. 10. Performance of FSC versus number of training samples (3-D DMD).

The strength of the multiple-channel combination shows when training data gets scarce. Fig. 10 compares the performance of the improved baseline (RC + Warp + CMS) with FSC for varying number of training samples (in seconds). Obviously, the performance differences increase with smaller amount of training data. With 120 s of training material, the relative improvements of FSC over the improved baseline (RC + Warp + CMS) is 24.2%, while with 30 s, the gain increases to 62%.

V. Far-Field Phonetic Speaker Recognition

Recently, the focus in automatic speaker recognition has extended from utilizing low-level acoustic features to "higher-level" features, originally proposed by [12] in the context of speaker recognition. Most prominent example of this phonetic speaker recognition are phonetic n-gram models, which are supposed to capture speaker idiosyncrasies and other phonetic and lexical speaker patterns [7]–[11], [13].The basic idea is to decode speech by various phone recognizers and to use the relative frequencies of phone n-grams as features for training and evaluating speaker models. In this section, we describe our extension of phonetic speaker recognition to the far-field speaker recognition task. We enrich existing algorithms by applying a larger number of language independent phone recognizers, compare multilingual to multiengine system combinations, and study the number of languages involved in decoding the speech.

A. *Phonetic Speaker Model Training and Evaluation*

To decode the speaker's speech into phone sequences, we used phone recognizers from the GlobalPhone project [17] available in eight different languages: Mandarin Chinese (CH), German (DE), French (FR), Japanese (JA), Croatian (KR), Portuguese (PO), Spanish (SP), and Turkish (TU). Fig. 11 shows the phoneme error rate in relation with number of phonemes modeled in each phone recognizer. Phone recognition is performed with a Viterbi search using a fully connected null-grammar network of mono-phones with equal-probable language model, i.e., no prior knowledge about phone statistics is applied to the decoding. For further details on the phone recognizers, we refer the reader to [16]. Silence labels of duration greater than 0.5 s in the obtained phonetic sequences were wrapped together as an end of utterance to capture information about how a speaker interacts with others by for example pausing frequently.

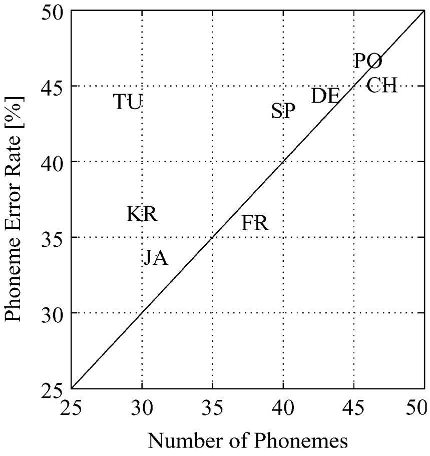


Fig. 11. Phoneme error rate versus modeled phonemes.

Based on the postprocessed phonetic sequences, we generated language-dependent speaker phonetic models (LSPM) by applying n-grams modeling. Bigram models were estimated from the CMU-Cambridge Statistical Language Modeling Toolkit (CMU-SLM) [15]. In the following, LSPM_i^k refers to the language-dependent phonetic bigram model of speaker k in language i. To train a speaker model for speaker k, each of the $M = 8$ phone recognizers $(\mathrm{PR}_1, \ldots, \mathrm{PR}_M)$ decodes the training data of speaker k to produce M phonetic sequences. Based on these phonetic sequences, M LSPMs are created for speaker k. No transcriptions of speech data are required at any step of model training. To recognize a speaker, each of the M phone recognizers PR_i of language i (the very same as used for generating the phonetic sequences) decodes the test speech. The resulting M phonetic strings are scored against k LSPMs in language i that matches the language of the phone recognizer. The scoring is performed by calculating the perplexity PP_i^k of the test sequence on LSPM_i^k. Finally, the perplexity scores from all M languages are combined into a final decision score IDS^k for speaker k, with an (equal) fusion weight w_i for each language

$$\mathrm{IDS}^k = \sum_{i=1}^{M} w_i \cdot \mathrm{PP}_i^k.$$

An unknown speaker is identified as $s*$ by

$$s^* = \arg\min_{k=1}^{S}\{\mathrm{IDS}^k\}$$

where S refers to the total number of enrolled speakers. We refer to this approach as LSPM-pp and illustrate the recognition process in Fig. 12 for $M = 5$.

For the LSPM-pp approach, the data amount per speaker is crucial in order to estimate reliable probabilities for the bigram speaker models. Therefore, we tested this approach on the 2-D DMD as this database has the most data per speaker. The first four sessions in 2-D DMD are used for training (about 7 min per speaker, which corresponds to roughly 5000 phone tokens). The evaluation is carried out on the remaining fifth session, resulting

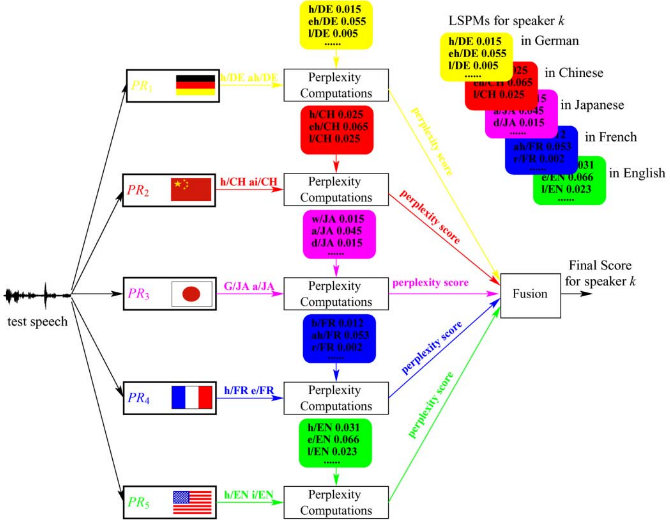


Fig. 12. Decision score computation with LSPM-pp (one enrolled speaker).

TABLE IV
PERFORMANCE PER LANGUAGE ON DIS0 UNDER MATCHED CONDITION (%)

Test Duration Language	60s	40s	10s	5s
CH	100	100	56.7	40
DE	80	76.7	50	33.3
FR	70	56.7	46.7	16.7
JA	30	30	36.7	26.7
KR	40	33.3	30	26.7
PO	76.7	66.7	33.3	20
SP	70	56.7	30	20
TU	53.3	50	30	16.7
fusion of all languages	**96.7**	**96.7**	**96.7**	**93.3**

in about 1 min of speech per speaker, corresponding to 1000 phone tokens.

We developed two systems for phonetic speaker recognition. The first uses the eight phone recognizers from the eight GlobalPhone languages, refered to as the "multilingual system." The second system called "multiengine system" was produced by English phone recognizers trained on multiple conditions (conversational telephone speech—SWB, spontaneous wide-band dialogues—Verbmobil, and planned broadcast news monologues—BN). Since the first system draws information from complementary phone recognizers, we anticipate greater robustness. Furthermore, the multilingual system is expected to be somewhat language independent.

B. Multilingual LSPM-pp Speaker Identification Results

Table IV gives a breakdown for language-dependent speaker recognition accuracy of the LSPM-pp approach at different test length under matched conditions for distance Dis0. It shows that the performance decreases significantly with shorter test duration on single languages, but that the fusion of multilingual information from all eight languages can overcome this shortfall. Table V compares the multilingual LSPM-pp recognition results for all distances on varying test durations under matched and mismatched conditions. In case of mismatched conditions, we apply all $D \times M$ phonetic models $(\mathrm{LSPM}_{i,d}^{k})$ for speaker k, where D is the total number of distant channels, i.e., we do not assume to know the test channel. For this scenario, final decision score is computed as

TABLE V
LSPM-PP PERFORMANCE UNDER MATCHED AND MISMATCHED CONDITIONS (%)

Test Length	Matched				Mismatched			
Channel	60s	40s	10s	5s	60s	40s	10s	5s
Dis0	96.7	96.7	96.7	93.3	96.7	96.7	96.7	90
DisL	96.7	96.7	86.7	70.0	96.7	100	90.0	66.7
Dis1	90.0	90.0	76.7	70.0	93.3	93.3	80.0	70.0
Dis2	96.7	96.7	93.3	83.3	96.7	96.7	86.7	80.0
Dis4	96.7	93.3	80.0	76.7	96.7	96.7	93.3	80.0
Dis5	93.3	93.3	90.0	76.7	93.3	93.3	86.7	70.0
Dis6	83.3	86.7	83.3	80.0	93.3	86.7	83.3	60.0
Dis8	93.3	93.3	86.7	66.7	93.3	93.3	86.7	70.0

$$\mathrm{IDS}^{k} = \sum_{i=1}^{M} w_i \cdot \min_{d=1}^{D}\{\mathrm{PP}_{i,d}^{k}\}$$

where $\mathrm{PP}_{i,d}^{k}$ is the decision score in language i on the distant channel d. The decision rule is given as

$$s^{*} = \arg\min_{k=1}^{S}\{\mathrm{IDS}^{k}\}$$

where k is the index of enrolled speakers, and S is the total number of enrolled speakers. Since we used all the channels under mismatched conditions, we sometimes see better performance than under matched condition.

C. Multiengine LSPM-pp Speaker Identification Results

Implicit to our strategy is the assumption that phone strings originating from different language-dependent phone recognizers yield crucial complementary information. In the following experiment, we explore if the success of this approach indeed results from language diversity or from simply using different recognizers. If the latter is the case, a multiengine approach, in which phone recognizers are trained on the same language but different conditions, might perform equally well. To test this hypothesis, we trained three English phone recognizers on different channel conditions (telephone, channel-mix, clean) and speaking styles (highly conversational, spontaneous, planned) using data from Switchboard, Broadcast News, and Verbmobil.

The experiments were carried out on matched conditions on all distances for 60-s chunks. For a fair comparison, we generated all possible language triples out of the set of eight languages ($\binom{8}{3} = 56$ triples) and calculated the average, minimum, and maximum performance over all triples. The results are given in Table VI.

The results show that the multiple-engine approach lies in all but two cases within the range of the multiple-language approach. However, the best performance of the multiple-language approach mostly outperforms the multiple-engine approach, i.e., most of the language triples achieve better results than the single language multiple-engines. We conclude that multiple English language recognizers provide less useful information than multiple language phone recognizers, at least for the given choice of multiple engines in the context of speaker identification. The fact that the multiple engines were trained on English, i.e., the same language which is spoken in

TABLE VI
PERFORMANCE COMPARISON OF MULTILINGUAL VERSUS MULTIENGINE (%)

System Test Channel	Multilingual Avg (Min - Max)	Multi-Engine
Dis0	87.92 (66.7 - 100)	93.3
DisL	88.21 (63.3 - 96.7)	86.7
Dis1	83.57 (66.7 - 93.3)	86.7
Dis2	93.63 (86.7 - 96.7)	76.7
Dis4	81.43 (56.7 - 96.7)	86.7
Dis5	86.07 (66.7 - 96.7)	83.3
Dis6	81.96 (66.7 - 93.3)	63.3
Dis8	87.14 (63.3 - 93.3)	63.3

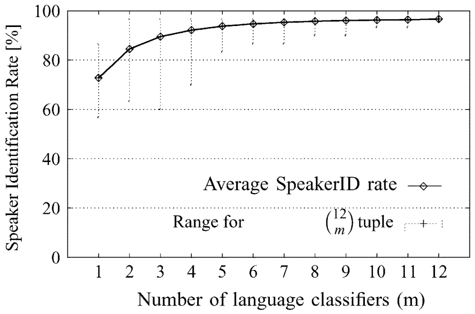


Fig. 13. SID performance over number of languages.

the speaker identification task, whereas the multiple languages were trained on eight languages but English, makes the multiple-language approach even more appealing as it indicates language independence and thus great potential for portability to speaker recognition in any language. Further experiments in which the multiengine recognizers were combined with the multilingual recognizers did not give additional improvements.

D. Number of Languages Versus Identification Performance

We investigated the impact of the number of languages, i.e., the number of phone recognizers on speaker recognition performance. Fig. 13 plots the speaker identification rate over the number k of languages on matched conditions on 60 s of data. For these experiments, we applied phone recognizers from GlobalPhone available in 12 languages, the eight as described above plus Arabic (AR), Korean (KO), Russian (RU), and Swedish (SW). The performance is given in average over the m out of 12 language m-tupel $\binom{12}{m}$ for Dis0. The average speaker identification rate increases for all distances with the number of languages. For some distances, a saturation effect takes place after six languages involved (such as Dis0 and Dis1), while for others distances even adding the 12th language has a positive effect on the average performance (such as Dis4, Dis6, and DisL). It also shows that the maximum performance of 96.7% can already be achieved using two languages. Among the total of $\binom{12}{m} = 66$ language pairs, CH-KO and CH-SP gave the best results. However, we were not able to derive an appropriate strategy to predict the best language tupels. Therefore, it is comforting that the chances of finding suitable language tupels get better with the number of applied languages. While only 4.5% of all 2-tupels achieved highest performance, 35% of 4-tupels, 60% of all 6-tupels, and 88% of all 10-tupels gave optimal performance. We furthermore analyzed whether the performance gain is related to the total number of phones rather than the number of different engines, but did not find evidence for such a correlation.

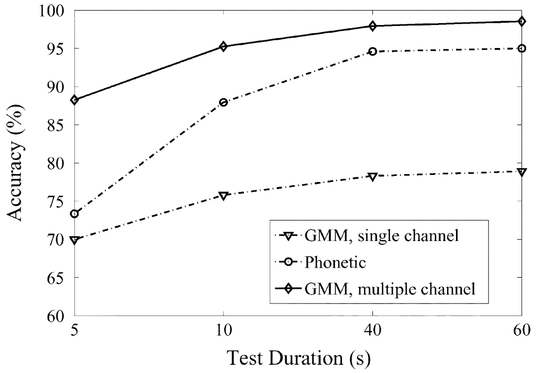


Fig. 14. Low-level system (GMMs) versus high-level system (phonetic).

E. GMM Versus Phonetic Approach

Finally, we compared the traditional GMM-based approach incorporating reverberation compensation and feature warping techniques to the phonetic approach. Fig. 14 shows the comparison between three systems for mismatched conditions on 2-D DMD, the GMM system with and without channel combination, and the phonetic system. As can be seen, the phonetic approach outperforms the GMM system in the single-channel condition. Overall, the GMM system with multiple channel combination achieves the best performance. Additionally, we used simple linear fusion to combine the GMM and the phonetic systems but did not see any additional gains. More elaborated fusion strategies may be investigated in the future.

VI. CONCLUSION

We studied robust speaker recognition in far-field microphone situations and investigated two approaches to improve the robustness of speaker recognition: the traditional GMM technique based on low-level acoustic features and the phonetic speaker recognition technique using high-level phonetic information. For the low-level approach, we introduced reverberation compensation and feature warping to the feature preprocessing step. Both methods lead to significant gains, even under mismatched training-testing conditions. On mismatched conditions, 41.6% relative improvement is achieved on the 3-D Distant Microphone Database, 17.1% on the 2-D Distant Microphone Database, and a 34.1% on the ICSI Meeting Database. In addition, we performed multiple-channel combination experiments to make use of information from multiple distant microphones. Four different combination methods were introduced "data combination," "frame-based score competition," "segment-based score fusion," and "segment-based decision voting." All four achieve additional gains on mismatched conditions. Overall, we observed 72.8%, 77.8%, 62.4%, and 57.9% relative improvements over the improved baseline $(\text{RC} + \text{Warp} + \text{CMS})$ on the 3-D DMD. We also found that the gains hold across the three different databases. The results indicate that variability in the training data combined with supplementary information from multiple sources improves the

TABLE VII
RELATIVE GAINS BY RC, WARP, AND MULTIPLE-CHANNEL COMBINATION OVER BASELINE-CMS

Database Approach	3D DMD	2D DMD	ICSI
RC+Warp+CMS+DC	84.1%	85.1%	40.5%
RC+Warp+CMS+FSC	87.1%	92.6%	41.6%
RC+Warp+CMS+SSF	78.1%	81.4%	38.6%
RC+Warp+CMS+SDV	75.4%	71.0%	36.4%

overall system robustness. Table VII summarizes the findings for mismatched conditions.

The second approach makes use of higher-level linguistic features. To capture speaker idiosyncrasies, we apply n-gram models trained on multilingual phone strings and show that higher-level features are more robust under mismatching conditions. Our results show that the phonetic approach is also very robust against channel mismatch; however, one major limitation of phonetic speaker recognition is the requirement for large amounts of training data to reliably estimate phonetic n-gram models. Furthermore, we compared the performances between multilingual and multiengine systems, and examined the impact of number of involved languages on recognition results. Our findings confirm the usefulness of language variety and indicate a language-independent nature of this approach, which suggests that speaker recognition using multilingual phone strings could be successfully ported to any language.

REFERENCES

[1] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. Speaker Odyssey 2001 Conf.*, Jun. 2001, pp. 213–218.

[2] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-19, no. 2, pp. 254–272, Apr. 1981.

[3] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "The ICSI meeting corpus," in *Proc. ICASSP*, 2003, pp. I-364–I-367.

[4] D. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Commun.*, vol. 17, no. 1–2, pp. 91–108, Aug. 1995.

[5] Q. Jin, Y. Pan, and T. Schultz, "Far-field speaker recognition," in *Proc. ICASSP*, 2006, pp. I-937–I-940.

[6] B. Xiang, U. Chaudhari, J. Navratil, G. Ramaswamy, and R. Gopinath, "Short-time Gaussianization for robust speaker verification," in *Proc. ICASSP*, 2002, pp. I-681–I-684.

[7] F. Weber, L. Manganaro, B. Peskin, and E. Shriberg, "Using prosodic and lexical information for speaker identification," in *Proc. ICASSP*, 2002, pp. I-141–I-144.

[8] Q. Jin, T. Schultz, and A. Waibel, "Phonetic speaker recognition," in *Proc. ICSLP*, 2002, pp. 1345–1348.

[9] Q. Jin, T. Schultz, and A. Waibel, "Speaker identification using multilingual phone strings," in *Proc. ICASSP*, 2002, pp. I-145–I-148.

[10] W. Andrews, M. Kohler, and J. Compbell, "Phonetic Speaker Recognition," in *Proc. Eurospeech*, 2001, pp. 2517–2520.

[11] W. Andrews, M. Kohler, J. Campbell, J. Godfrey, and J. Hernandez-Cordero, "Gender-dependent phonetic refraction for speaker recognition," in *Proc. ICASSP*, 2002, pp. I-149–I-152.

[12] G. Doddington, "Speaker recognition based on idiolectal differences between speakers," in *Proc. Eurospeech*, 2001, pp. 2521–2524.

[13] Q. Jin, J. Navratil, D. Reynolds, W. Andrews, J. Campbell, and J. Abramson, "Combining cross-stream and time dimensions in phonetic speaker recognition," in *Proc. ICASSP*, 2003, pp. IV-800–IV-803.

[14] Q. Jin, "Robust speaker recognition," Ph.D. dissertation, School of Computer Science, Carnegie Mellon Univ., Pittsburgh, PA, 2007.

[15] P. Clarkson and R. Rosenfeld, "Statistical language modeling using the CMU-Cambridge toolkit," in *Proc. Eurospeech*, 1997, pp. 2707–2710.

[16] T. Schultz and K. Kirchhoff, *Multilingual Speech Processing*, 1st ed. New York: Elsevier/Academic, 2006.

[17] T. Schultz and A. Waibel, "Language independent and language adaptive acoustic modeling for speech recognition," *Speech Commun.*, vol. 35, pp. 31–51, 2001.

[18] T. Schultz, Q. Jin, K. Laskowski, A. Tribble, and A. Waibel, "Improvements in non-verbal cue identification using multilingual phone strings," in *Proc. Speech-to-Speech Translation Workshop, ACL*, 2002, pp. 101–108.

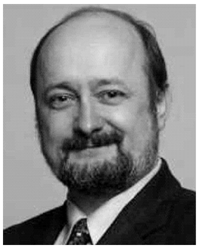
[19] Y. Pan, "Robust speech recognition on distant microphones," Ph.D. dissertation, Lang. Technol. Inst., Carnegie Mellon Univ., Pittsburgh, PA, 2007, thesis in submission.

Qin Jin received the B.S. and M.E. degrees in computer science and technology from Tsinghua University, Beijing, China, in 1996 and 1999, and the M.S. and Ph.D. degrees in computer and information science in 2001 and 2007, respectively.

She is now a Postdoctoral Fellow with the Interactive Systems Laboratories, Language Technologies Institute, Carnegie Mellon University (CMU), Pittsburgh, PA.

Tanja Schultz received the M.S. and Ph.D. degrees in computer science from the University of Karlsruhe, Karlsruhe, Germany, in 1995 and 2000, and the M.S. degree in mathematics and sports from the University of Heidelberg, Heidelberg, Germany, in 1990.

In 2000, she joined Carnegie Mellon University, Pittsburgh, PA, as an Assistant Research Professor at the Language Technologies Institute and serves as Associate Director of the interACT Center. Her research focus on multilingual speech processing with a special emphasis on rapid language adaptation.



Alex Waibel received the B.S. degree in electrical engineering from the Massachusetts Institute of Technology, Cambridge, in 1979, and the M.S. and Ph.D. degrees in computer science from Carnegie Mellon University, Pittsburgh. PA, in 1980 and 1986, respectively.

He is a Professor of Computer Science at Carnegie Mellon University and at the University of Karlsruhe,Karlsruhe, Germany. He directs the interACT Center at both Universities with research emphasis in speech and language processing, speech translation, and multimodal and perceptual user interfaces.