FAR-FIELD SPEAKER RECOGNITION

Qin Jin, Yue Pan, and Tanja Schultz

Language Technologies Institute, School of Computer Science, Carnegie Mellon University
{qjin|ypan|tanja}@cs.cmu.edu

ABSTRACT

In this paper we study robust speaker recognition in far-field microphone situations such as meeting scenarios. By applying reverberation compensation and feature warping we achieved significant improvements under mismatched training-testing conditions. To capture useful information from multiple distant microphones, two approaches for multiple channel combination are investigated. This leads to 84.1% and 78.1% relative improvements on the Distant Microphone database. Furthermore, we tested the resulting system on the ICSI Meeting Corpus. The improvements are also very high on this task, which indicates that our system is robust to changing conditions in a remote microphone setting.

1. INTRODUCTION

Speaker recognition is the process of determining the identity of the person who is speaking. In a world of ubiquitous computation and communication, there are more and more applications that require the recognition of a person from his or her voice, such as verification of access permission, voice aided transactions authentication, and multimedia database management/retrieval. Speaker recognition systems are desired to perform reliably in a variety of environments, tasks and configurations. Moreover hands-free sound capture with distant microphones is required by many real applications.

Speaker recognition has achieved fairly good performance under controlled conditions as reported in the NIST annual speaker recognition evaluation [1]. However, robust speaker recognition with hands-free far-field microphones is still challenging.

Accurate far-field speaker recognition is difficult due to a number of factors. Channel mismatch as well as environmental noise and reverberation are two most prominent ones. For example, linear channel effects will shift the mean of Mel-Frequency Cepstral Coefficients (MFCC), and additive noise will tend to modify the variance [2]. During the past years, much research has been conducted towards reducing the effect of channel mismatch. A number of methods for reducing these effects were proposed. Cepstral Mean Subtraction (CMS) [3] and RASTA [4] are two of the standard feature-based approaches. However, channel mismatch and environmental noise can still cause lots of errors after CMS and RASTA. To deal with additive noise, a feature warping technique had been proposed that transforms the distribution of cepstral features to a standard distribution [2]. This technique was reported to bring more improvements compared to standard techniques. In this paper we propose a new reverberation compensation approach. It uses a different noise estimation compared to the standard spectrum subtraction approach. We applied feature warping after reverberation compensation in our system. The experimental results show that significant improvements were achieved over the baseline system. Furthermore, two multiple channel combination approaches are investigated to capture useful information from multiple distant microphones. They result in additional large improvement over the baseline system.

2. DATABASE AND EXPERIMENTAL SETUP

2.1. Distant Microphone Database

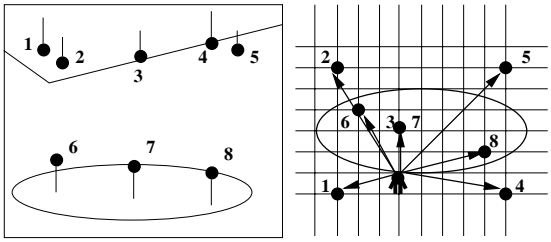

Fig. 1. Distant microphone setup in ISL meetings

In order to investigate robust speaker recognition with distant microphones, a speaker database was collected at the Interactive Systems Laboratories (ISL) in a meeting room using multiple distant microphones. The left hand-side of Figure 1 shows the distant microphone setup. Five microphones (labeled 1 to 5) are hanging from the ceiling, while three microphones (6, 7, and 8) are set up on the meeting table. We used miniature cardioid condensor microphones that are very similar to omni-directional microphones. Speakers tend to turn their head slightly during the recordings but they do not move their body too much. The right hand-side of Figure 1 illustrates the positioning of these 8 microphones in relation to the

speaker. The grid indicates the distance definition and corresponds to roughly 0.5 meters per grid unit. The vertical grid is set to 4. The distance definition of a speaker to a microphone is the Euclidean grid distance (horizontally and vertically) penalized by both the horizontal and vertical angles between the speaker (sound source) and the microphone (the receiver). For example, the distance of ceiling microphone channel 2 is computed as $D(2) = \frac{\sqrt{3^2+5^2+4^2}}{\cos(\arctan(\frac{4}{\sqrt{34}}))\cos(\arctan(\frac{3}{5}))} = 10$, 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. The distances of other channels are computed similarly, which are $D(7) = 2, D(6) = 4.3, D(8) = 10, D(3) = 10, D(5) = 11.4, D(4) = 12, D(1) = 14.5$.

There are 24 speakers in total in the Distant Microphone database. Each speaker was recorded in one session. Each speaker was required to talk about a selection of 10 given topics of personal interest. So the speaking style is spontaneous free speech. The speech duration varies from 8 minutes to 20 minutes. 2 minutes of speech was randomly chosen from the first 80% of a speaker's entire recording as training data for that speaker. The remaining 20% of speech is split into 20 seconds segments and each of them is used as one test trial. There are in total 183 test trials.

2.2. ICSI Meeting Corpus

The ICSI Meeting Corpus [5] is a collection of 75 meetings with simultaneous multi-channel audio recordings collected at the International Computer Science Institute (ICSI) in Berkeley. There are a total of 53 unique speakers in this corpus. We selected 24 speakers based on their positions and whether they have enough speech in meetings. Figure 2 is a simple diagram of the distant table microphone arrangement in the ICSI meeting room and the speaker position we selected. The table microphones are desktop omni-directional Crown Pressure Zone Microphone (PZM) microphones. They were arranged in a staggered line along the center of the conference table. 90 seconds of speech was randomly selected from meetings for each speaker as training data. The remainder speech was used for testing. We use the manual transcription to keep the test segments as they are if they were not longer than 20 seconds. Otherwise the segment is split into several 20 seconds chunks. There are 397 test trials in total.

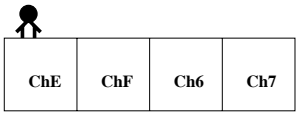

Fig. 2. Distant table microphone setup in ICSI meetings

2.3. Speaker Modeling and Performance Measure

Over the past decades, GMM has become the dominant approach for speaker modeling in speaker recognition systems which use untranscribed training data [6]. In our system a GMM with 128 mixtures was trained for each speaker via the EM algorithm. We assume that the testing speaker is one of the trained speakers, which means closed-set speaker recognition is evaluated in this paper. The system performance is measured using recognition accuracy, which is the percentage of correctly recognized test trials over all test trials.

3. ROBUSTNESS TO FAR-FIELD

3.1. Methods

3.1.1. Reverberation Compensation

A distant-talking speech signal is 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,

$$y[t] = x[t] * h[t] + n[t] \quad (1)$$

where the source signal $x[t]$ is the clean speech, $h[t]$ is the impulse response of room reverberation, and $n[t]$ is recording noise. Cepstrum Mean Subtraction (CMS) has been used successfully 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 16ms-32ms. Unfortunately, the duration of impulse response of reverberation usually has a much longer tail, as long as more than 50ms. Therefore, traditional CMS will not be as effective under these conditions.

We separate the impulse response $h[t]$ into two parts $h_1[t]$ and $h_2[t]$, where, $h[t] = h_1[t] + \delta(t-T)h_2[t]$

$$h_1[t] = \begin{cases} h[t] & t < T \\ 0 & otherwise \end{cases} \quad h_2[t] = \begin{cases} h[t+T] & t \geq 0 \\ 0 & otherwise \end{cases}$$

and rewrite formula (1) 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 whose length is smaller than the DFT analysis window, thus it can be compensated by the conventional CMS. For $x[t-T] * h_2[t]$, we treat it the same as 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 $y[t-T]$, then 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, 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, performance degrades significantly. For the recording setup in this paper, we found

$a = 1.0$, $b = 0.1$ and $g(\omega) = |1 - 0.9e^{j\omega}|$ optimal in most changing conditions based on development data. Standard CMS is applied after spectrum subtraction to eliminate the effect of $h_1[t]$.

3.1.2. Feature Warping

The feature warping method applied here was proposed in [2]. It warps the distribution of a cepstral feature stream to a standardized distribution over a specified time interval. The warping is implemented via Cumulative Distribution Functions (CDF) matching as described in [7]. In our experiments, the window size is 300 frames and the window shifts one frame. Zeros are padded at the beginning and at the end of the raw feature stream.

Table 1. *Detailed baseline system performance (in %)*

Train–Test	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**

3.2. Experimental Results

The front-end processing of the baseline system relies on MFCC analysis. The signal is characterized by 13-dimensional MFCC every 16ms. A speech detection process based on normalized energy is used in order to remove non-informative frames. The mean feature vector is computed on the informative frames only. The non-informative frames are discarded during training speaker models. The improved system adds reverberation compensation and feature warping (RC+Warp) in the front-end processing while keeping other system components the same as the baseline system.

Table 1 presents the speaker recognition accuracy of the baseline system under all possible training-testing conditions. It shows that accuracies under matched conditions (numbers in bold) are much better than under the mismatched conditions (off the diagonal). The average accuracies under matched and mismatched conditions are 94.0% and 74.2% respectively.

Figure 3 shows the reverberation compensation plus feature warping effect on system performances on both data sets. We can see that significant improvements were achieved under matched and mismatched conditions on both data sets. On average, 45.5% and 41.6% relative improvements are achieved under matched and mismatched conditions respectively on the Distant Microphone database, and 31.9% and 34.1% on the ICSI Meeting Corpus, indicating that the applied methods are robust under different conditions.

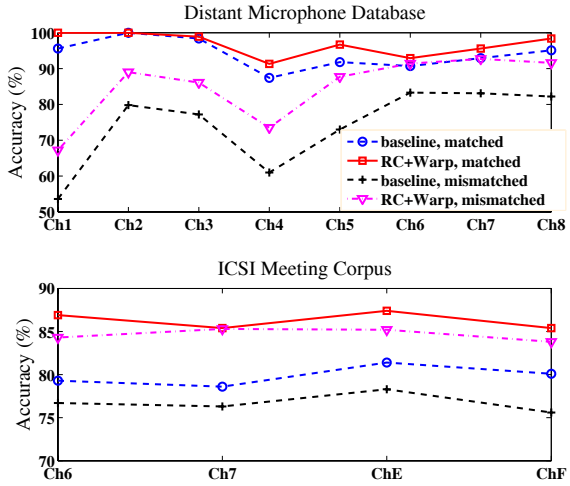

Fig. 3. Performance improvement by RC+Warp

Table 2 shows the performance improvement by reverberation compensation alone, feature warping alone and reverberation compensation plus feature warping on the Distant Microphone database. Each of the two approaches improves performance under both matched and mismatched conditions. Combining both approaches provide more improvement.

Table 2. *RC and Warp effect on system performance (in %)*

Condition	baseline	RC	Warp	RC+Warp
Matched	94.0	94.8	96.4	96.7
Mismatched	74.2	78.1	79.1	84.9

4. MULTI-CHANNEL COMBINATION

4.1. Methods

Hands-free multiple distant microphones are easy to set up and quite common in applications such as meetings and lectures. In order to benefit from the multiple channel setup, two multi-channel combination approaches are investigated: one is applied at the data source level, the other at the decision level. "Data Combination" means the speaker models are trained using data from multiple mismatched channels. For example, for test on channel 1, the speaker models are trained using all channels (Ch2 to Ch8) but channel 1. Consequently, the training data does not cover the test channel, so that the tests are still performed under mismatched condition. "Decision Combination" means combination of the decision scores from the 7 GMM classifiers, each of which is trained on one of the 7 mismatched channels. For example, test trials from channel 1 are evaluated with 7 mismatched classifiers which are trained on channel 2 to channel 8 and the 7 decision scores are linearly combined with equal weights.

4.2. Experimental Results

Figure 4 presents the system improvement achieved by adding the two multi-channel combination approaches to the im-

proved system under mismatched condition. Significant improvements were achieved by both combination approaches on both data sets. On average, 84.1% relative improvement over the baseline was gained by data combination on the Distant Microphone database. We want to point out that in this approach, we control the amount of training data to be the same as in the baseline system by randomly choosing $\frac{1}{7}$ data from each of the original mismatched channel. So the improvement proves that seeing more variability in training improves the recognition robustness. 78.1% relative improvement over the baseline on average was achieved by decision combination on the Distant Microphone database. This indicates that it is beneficial to use information from multiple sources even though each of them is not very powerful. 40.5% relative improvement was achieved by data combination and 37.8% was achieved by decision combination on the ICSI Meeting Corpus. We also observe additional gain when two combinations are used together. For example, 50.8% relative improvment over the baseline was achieved on the ICSI Meeting Corpus compared to 40.5% and 37.8%.

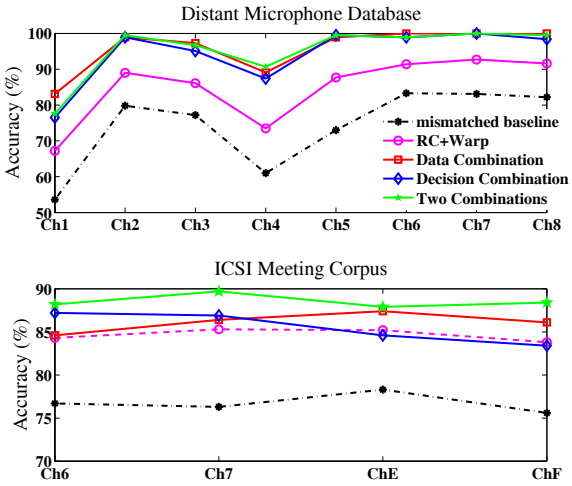

Fig. 4. Performance improvement by combination

5. CONCLUSION

In this paper we presented our robust speaker recognition system in meeting scenario with multiple distant microphones. A new reverberation compensation approach plus feature warping significantly improve the system robustness under mismatched training-testing conditions. 41.6% relative improvement is achieved on the Distant Microphone Database and 34.1% relative improvement is achieved on the ICSI Meeting Corpus. Furthermore, two multi-channel combination approaches are investigated in order to capture useful information from multiple channel sources. 84.1% and 78.1% relative improvements are achieved with these two approaches on the Distant Microphone database, which shows that seeing more variability in training and combining supplementary information from multiple sources improves the system robustness. The improvement carries over to the ICSI Meeting Corpus (40.5% and 37.8% relative improvement), which indicates that our system is robust across datasets with different multiple distant microphone settings.

Figure 5 shows the relationship between recognition accuracy and channel distance on the Distant Microphone database. Apparently the performance is a function of the distance value: after surpassing a critical distance between speaker and microphone (mic 5,4,1) the performance decreases significantly. The distance value can act as heuristic information to better combine multiple channels at the decision level. For example, we can give higher weights to the decisions from the classifiers belonging to the closer microphones. Our preliminary experimental results indicate that such heuristic weights combination outperforms equal weights combination.

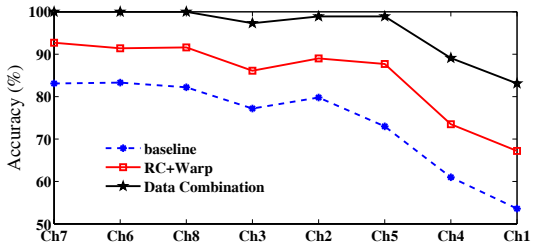

Fig. 5. Performance over microphone distance

6. REFERENCES

[1] NIST Annual Speaker Recognition Evaluation http://www.nist.gov/speech/tests/spk/index.htm.

[2] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," Proc. Speaker Odyssey 2001 conference, June 2001.

[3] S. Furui, "Cepstral analysis technique for automatic speaker verification," IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 19, p. 254-272, 1981.

[4] H. Hermansky and N. Morgan, "RASTA processing of speech," IEEE Trans. Speech and Audio Processing, vol.2, no.4, p.578-589, 1994.

[5] 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

[6] D. Reynolds, "Speaker Identification and Verification Using Gaussian Mixture Speaker Models," Speech Communication, Vol. 17, No. 1-2, p. 91-108, August 1995

[7] B. Xiang, U. Chaudhari, J. Navratil, G. Ramaswamy and R. Gopinath, "Short-time Gaussianization for Robust Speaker Verification," in Proc. ICASSP, 2002.