Robust Far-Field Speaker Identification under Mismatched Conditions

Qin Jin and Tanja Schultz

InterACT, Language Technologies Institute, Carnegie Mellon University, USA

{qjin,tanja}@cs.cmu.edu

Abstract

While speaker identification performance has improved dramatically over the past years, the presence of interfering noise and the variety of channel conditions pose a major obstacle. Particularly the mismatch between training and test condition leads to severe performance degradations. In this paper we investigate speaker identification based on data simultaneously recorded with multiple microphones in a far-field setup under different noise and reverberation conditions. Dramatic performance degradation is observed, especially when training and test conditions mismatch. To address this mismatch we apply our robust frame-based score competition approach in which we combine and compete models trained on multiple conditions. To further improve this approach we add simulated, i.e. artificially created training data on a variety of noise conditions for additional model training. Our experimental results show that the extended approach significantly improves speaker identification performance under adverse and mismatching conditions.

Index Terms: Speaker Identification, Far-field, Frame-based Score Competition

1. Introduction

Over the years automatic Speaker IDentification (SID) has developed into a rather mature technology that is crucial to a large variety of spoken language applications. However, SID systems still lack robustness, i.e. their performance degrades dramatically when the acoustic training data mismatch with the given test conditions [1][2]. Robustness is currently the major challenge for real-world applications of speaker recognition. Traditional approaches, such as Gaussian Mixture Models (GMM) [3][4], achieve very high accuracies for speaker identification and verification tasks on high-quality data when training and test conditions are well controlled. Unfortunately, real-world applications are required to handle a large variety of speech signals which are often corrupted by adverse environmental conditions (noise, interfering speech, background, and channel). Furthermore, we cannot assume that training data are provided for all relevant conditions. Thus SID faces the situation that models for speakers are trained on one particular set of conditions but have to be applied to vastly different (mismatched) conditions. GMM-based systems are known to degrade significantly under adverse and mismatched conditions. This performance degradation becomes even more severe when speech signals are captured from the distance.

We proposed the "Frame based Score Competition" (FSC) approach in [5] to improve speaker recognition in far-field situations. In this paper we further elaborate this approach by adding simulated data on a large variety of noise conditions, i.e. we artificially create additional data by applying a filter

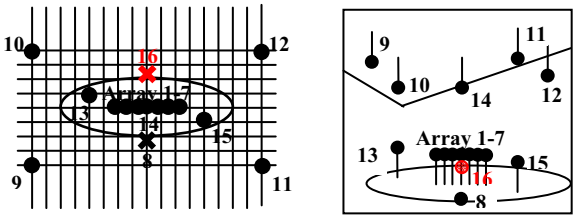

Figure 1: *Microphone setup in the FarSID database*

approach and extend the number and variety of models for the competition approach. The paper is organized as follows: In the next section we describe the far-field database as recently collected at CMU. Section 3 briefly reviews the Frame-based Score Competition approach. We describe the experimental setup and results in section 4, and discuss the outcome before we conclude in section 5.

2. Data and Setup

A Far-Field Speaker Identification (FarSID) Database has recently been collected at Carnegie Mellon University to study the performance of speaker identification algorithms in adverse conditions, including a far-field microphone setup, various interfering noise sources, and reverberant room characteristics. Similar to the database described in [5] the FarSID database consists of speech recordings from multiple far-distant microphones as depicted in Figure 1. In addition, to make the FarSID database even more challenging than its predecessor database, we additionally recorded under various noise and reverberation conditions.

The FarSID database consists of conversational speech recorded in face-to-face dialog sessions under two different reverberation conditions (small vs. medium-sized room) under six noise conditions per room. The noise conditions were applied by playing interfering noises at different Signal to Noise Ratio (SNR) levels (music, white noise, speech) while recording the respective sessions. Each condition lasted for about 13 minutes per session per speaker. The different noise conditions and their respective SNR are indicated below:

- No noise
- Music Noise -5 dB SNR and 10db SNR
- White Noise -5 dB SNR and 10db SNR
- Speaker Interfering Noise -5 dB

In total we recorded 10 native speakers of American English, where each speaker is engaged by an interviewer in a conversation about various topics. Each speaker participated in 4 recording sessions, two in a small and two in a medium-sized room, totaling to 2 hours duration per speaker. We are aware that the identification of 10 speakers does not pose a major challenge to modern SID systems typically applied to identify hundreds or thousands of different speakers. The purpose is

10.21437/Interspeech.2008-501

Copyright © 2008 ISCA

rather to study in depth the impact of our proposed algorithms to speech under various noise, SNR, and distance conditions before applying them to real-life identification problems.

Figure 1 shows the distant microphone setup in the FarSID database. The right-hand side illustrates the microphone positioning in the 3D space. Five microphones (labeled 9 to 12 and 14) are hanging from the ceiling or are mounted to high microphone stands. Seven microphones (labeled 1 to 7) are building a microphone array, with a distance of 5cm between each microphone. The microphone array and two other microphones (labeled 13 and 15) are set up on the table which is arranged between the interviewer and interviewee. In addition, we use two lapel microphones, number 8 worn by the interviewer and number 16 worn by the interviewee, i.e. the speaker whose identity is to be recognized. The left-hand side of Figure 1 illustrates the distance between the speaker (marked by a red "X") and these 16 microphones. One grid unit roughly corresponds to 0.25 meters.

3. Frame-base Score Competition (FSC)

In this section we first quickly review the decision process of our speaker identification systems based on GMM likelihood scores and then summarize our FSC approach which is described in more detail in [5]. Let S be the total number of enrolled speakers and $LL(X \mid \Theta_k)$ the log likelihood score that the test feature sequence X was generated by the GMM Θ_k of speaker k, which consists of M mixtures of Gaussian distributions. Then the recognized speaker identity S^* is given by:

$$S^* = \arg\max_k \{LL(X \mid \Theta_k)\} \qquad k = 1,2,\cdots,S \qquad (1)$$

Since the vectors of the sequence $X = (x_1, x_2, \cdots, x_N)$ are assumed to be independent and identically distributed, the likelihood score for speaker k and model Θ_k is computed as $LL(X \mid \Theta_k) = \sum_{n=1}^{N} LL(x_n \mid \Theta_k)$. Our multiple microphone setup allows us to build multiple GMM models $\Theta_k^{CH_i}$ for each speaker k and each channel CH_i, resulting in a set $\Theta_k = \{\Theta_k^{CH_1}, \cdots, \Theta_k^{CH_C}\}$ for k speakers and C channels. The key idea of the FSC approach is to use this set of multiple GMM models rather than a single GMM model for the speaker identity decision. In each frame we compare feature vector x_i provided by channel CH_h to the multiple GMMs of speaker k. The highest log likelihood score is chosen to be the frame score. In this case the likelihood score of the observed features given speaker k is computed as:

$$\begin{aligned} LL(X \mid \Theta_k) &= \sum_{n=1}^{N} LL(x_n \mid \Theta_k) \\ &= \sum_{n=1}^{N} \max \left\{ LL\left(x_n \mid \Theta_k^{CH_j}\right) \right\}_{j=1}^{C} \end{aligned} \qquad (2)$$

Note that this process does not rely on models for the test channel. Also, this competition process differs from the mono-channel scoring process in that per-frame log likelihood scores for different speakers are not necessarily derived on the same channel.

4. Experimental Setup and Results

The experiments reported in this paper are based on the FarSID database. The aim of the investigation is to demonstrate the robustness of our FSC approach for far-field scenarios and show how it addresses the challenges posed by mismatched condition, i.e. the fact that speaker models are applied to acoustic channel conditions which have not been observed during training. For this purpose we train and test our approach under four scenarios, where each scenario is designed to be more challenging than the previous one and more close to the challenges for real-world applications. The four scenarios are described next, followed by the experimental results in the respective subsections.

[Match] Matched condition: train a speaker model with data recorded under the same conditions as in the test case – this is the golden line as it reflects the best case scenario.

[Mis-MM] Mismatched condition with Multiple Microphone data: train speaker models with data simultaneously recorded by multiple microphones that cover a variety of conditions but the test condition.

[Mis-SM-k] Mismatched conditions with Single Microphone data and knowledge about test condition: train speaker models with data recorded with one microphone under one condition and tested on a different condition using some prior knowledge about the test condition.

[Mis-SM-nk] like [Mis-SM-k] but here we varied the number of microphone positions involved for model training.

The experiments were carried out on data with the first noise condition of the FarSID database (see section 2), i.e. distant speech with common background noise such as air conditioning, computer fans, and reverberation recorded by the 16 microphones (as described above) in a medium-sized room. We selected 60 seconds of these data per speaker to train the speaker model. For testing we selected 30 seconds per speaker of the same noise condition and same room. The major mismatch results from the selection of the microphone positions for training and test, as described above. In total we had 106 test trials. All described experiments are conducted as closed-set speaker identification. Performance is measured in terms of identification accuracy, i.e. the percentage of correctly identified test trials. The applied speech features X are Mel Frequency Cepstral Coefficients (MFCC); the speaker models consist of Gaussian Mixture Models (GMM) with 64 Gaussians per model.

4.1 Experiment on [Match] Scenario

To get the upper bound performance, i.e. the best case scenario we trained and tested under the matched scenario. The second column of Table 1 shows the breakdown of speaker identification performance for each microphone position. To get the performance for microphone position y we trained all speaker models on the training trials recorded by microphone y and tested on the test trials recorded by microphone y. So, on average we get 98.4% identification rate on 10 speakers for far-field recordings if we assume that we know the test condition and that we do have recordings in this condition available for each speaker.

4.2 Experiment on [Mis-MM] Scenario

The results of the second experiment are described in column 3 of Table 1. Here we assume that we do have simultaneous recordings from microphone positions 9-15. To calculate the performance on position 9 we train 6 speaker models per

speaker, one on each of the remaining positions 10, 11, 12, 13, 14, and 15. For the final number given in the table we average over the identification rates for each of these mismatching conditions. We achieve 92.1% accuracy over all positions, i.e. a drastic drop from the matched condition.

The fourth column in Table 1 compares this brute-force approach to our FSC technique. The same 6 models per speaker are now combined using competition at the scoring stage. As can been seen FSC significantly improves over the brute-force approach and even gets close to the [Match] performance. This result indicates that FSC compensates well for scenarios, in which recordings with multiple microphones but the matching one are available for a speaker. Our earlier results also showed that the SID performance further improves if the matching condition is available [5].

Table 1. *Performance for Multi-Microphone Setup*

Test Microphone	[Match]	[Mis-MM]	[Mis-MM] FSC
Position 9	98.1	85.8	97.2
Position 10	99.1	91.2	99.1
Position 11	98.1	92.1	97.2
Position 12	96.2	90.4	97.2
Position 13	100.0	94.2	100.0
Position 14	99.1	95.9	99.1
Position 15	98.1	95.0	98.1
Average	**98.4**	**92.1**	**98.2**

4.3 Experiment on [Mis-SM-k] Scenario

In this next set of experiments we investigate the performance of our FSC approach in the more realistic case where only single microphone recordings are available per speaker. We assume here without loss of generality to have training data from microphone position 4 [Mis-SM4]. As column 3 of Table 2 (labeled as [Mis-SM4]) shows, the performance drops drastically compared to the matched and the multi-microphone performance. The gap is more substantial for microphone positions 9 – 12, which is intuitively clear as these are further away from microphone 4 than microphone positions 13 – 15.

Table 2. *Performance for Single-Microphone Setup*

Test Microphone	[Match]	[Mis-SM4]	[Mis-SM4] FSC
Position 9	98.1	57.5	84.9
Position 10	99.1	66.0	93.4
Position 11	98.1	68.9	93.4
Position 12	96.2	71.7	94.3
Position 13	100.0	94.3	97.2
Position 14	99.1	95.3	98.1
Position 15	98.1	93.4	97.2
Average	**98.4**	**78.2**	**94.1**

The key idea to make use of our FSC approach in the single microphone case is to simulate multiple microphone recordings from the single microphone data. In order to apply FSC, we simulated different channels from the microphone 4 speech. This simulation targets microphone positions 9-15 by convolving the source speech with the simulated Room Impulse Response (RIR) generated using [6]. While the RIR program incorporates the dependence of room impulse response on many different physical characteristics of the room and environment, we focused here on three major attributes, the room size, reverberation time, and distance between the speaker and the microphone.

Column 4 in Table 2 (labeled [Mis-SM4]-FSC) shows the performance when simulating the 7 microphone positions. FSC on simulated channels significantly outperforms the baseline under mismatched conditions although it cannot beat the performance of FSC on real multi-microphone data "[Mis-MM] FSC" from Table 1. Please note that for both, the FSC on real multi-microphone data and on simulated multi-microphone data, we purposely exclude the data from the matching channel, i.e. we assume to not know the microphone position of the test condition. However, we do assume in the simulation to have some knowledge about possible microphone positions, i.e. the RIR filter do use the actual room size and reverberation time, and create realistic microphone distances.

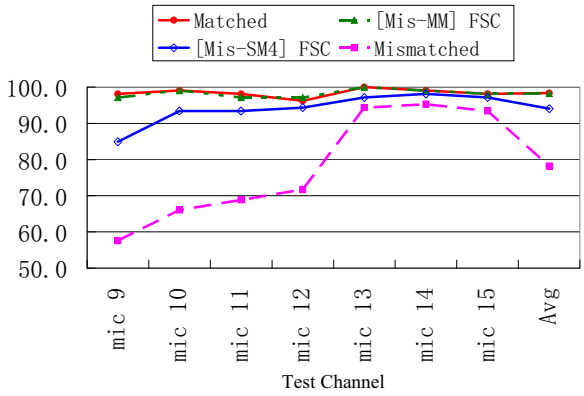

Figure 2: *SID Performance comparison under all setups*

Figure 2 summarizes our findings on the four cases, i.e. trained and tested on microphone 4 [Match], real multi-microphone conditions with FSC [Mis-MM]-FSC, single microphone conditions with simulation [Mis-SM4]-FSC, and the mismatched case [Mismatched], in which the models are trained on single microphone data at position 4 and applied to position 9-15 microphones.

4.4 Experiment on [Mis-SM-nk] Scenario

In this final set of experiments, we compared the impact of the number of microphone positions on performance. We tested this by repeating the [Mis-SM4]-FSC experiments but this time applying FSC on different numbers of simulated multi-microphone data streams. FSC6 refers to the case, in which we used all 6 mismatched microphone data (position 9 -15 except the position matched with test condition) to train 6 models. This corresponds to experiment [Mis-SM4]-FSC. FSC5 refers to the case where we used only 5 out of 6 simulated multi-microphone data. Since we can have 6 over 5 = 6 different choices, we averaged the performance over all different choices. In addition, we calculated the best and worst performance depending on the choice. FSC4, FSC3, FSC2 repeated the same experiments with fewer microphones giving us 6 over 4 = 15, 6 over 3 = 20, and 6 over 2 = 15 choices, respectively. Figures 3 and 4 show the best and worst performance for all selections.

As can been seen the worst selection of microphone positions for the data simulation does not have a significant impact on the

system performance compared to the best choice. In other words, the success of the FSC approach does not depend on a proper selection of microphone positions for the simulation. Figure 5 compares the worst, average, and best case and with the mismatched condition. Even in the worst case, FSC still significantly improves performance compared to the baseline performance under mismatched condition (noFSC in Figure 5).

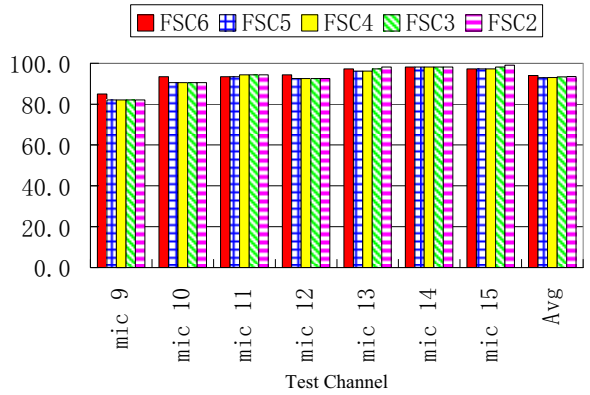

Figure 3: *Best performance of [Mis-SM4]-FSC*

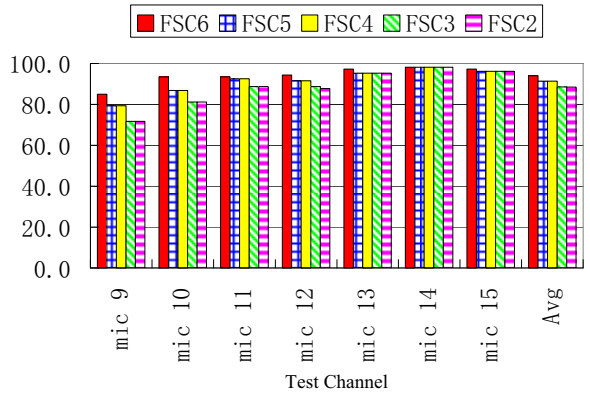

Figure 4: *Worst performance of [Mis-SM4]-FSC*

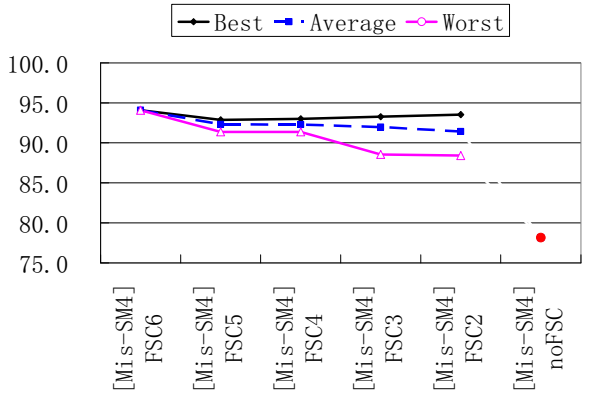

Figure 5: *Performance summary of FSC with different number of channels*

5. Discussion and Conclusions

In this paper we reported far-field speaker recognition performances under mismatched conditions. The aim of the investigation is to demonstrate the robustness of our frame-based score competition approach (FSC) for far-field scenarios and show how it addresses the challenges posed by mismatched condition, i.e. the fact that speaker models are usually applied to acoustic channel conditions which have not been observed during training. For this purpose we trained and tested our approach under four scenarios, where each scenario is designed to be more challenging than the previous one and more close to the challenges for real-world applications. The first scenario assumes to know the test condition, i.e. the best but most unlikely case. The second case assumes to not know the test condition but to have training samples recorded from multiple microphone positions for the speaker in question. Here, FSC significantly improves over the mismatched case, i.e. applying multi-microphone data gives more robustness since they cover multiple microphone positions and thus better prepare for the unknown. In the third scenario we assume to have only single-microphone data available and compensate this lack by simulating multi-microphone data using room impulse response filters. FSC manages to still significantly outperform the mismatched scenario. In other words, even when only single microphone recordings from a speaker are available, the simulation of multiple microphone recordings combined with our FSC approach improves the overall performance significantly. In the last scenario we vary the selection of microphone positions for the simulated data and show that even if we make the worst choice of microphone positions, we still see significant improvements over the mismatched case.

6. Acknowledgements

The work was funded in part by the Department of Defense. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Department of Defense, the United States Government, or Rosettex. We would like to thank Kshitiz Kumar for organizing and carrying out the database collection and providing the simulated data. Furthermore, we thank Rich Stern for his very valuable contributions and Fred Goodman for useful discussions and comments.

7. References

[1] Chin-Hui Lee, Frank K. Soong, Kuldip K. Paliwal, “Automatic Speech and Speaker Recognition: Advanced Topics”, Springer, 1996, ISBN:0792397061.

[2] S. Furui, “Towards Robust Speech Recognition Under Adverse Conditions”, ESCA Workshop on Speech Processing in Adverse Conditions, p. 31-42, 1992.

[3] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska-Delacretaz and D. A. Reynolds, “A Tutorial on Text-Independent Speaker Verification,” Journal on Applied Signal Processing 4, pp. 430-451, 2004.

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

[5] Q. Jin, T. Schultz, and A. Waibel, “Far-field Speaker Recognition,” IEEE Transactions on Audio, Speech, and Language Processing, Vol. 15, No.7, p. 2023-2032, 2007.

[6] D. Campbell, K. Palomäki, and G. Brown, “A MATLAB Simulation of “Shoebox” Room Acoustics for use in Research and Teaching. http://cis.paisley.ac.uk/research/journal /V9/V9N3/campbell.doc