SPEAKER IDENTIFICATION WITH DISTANT MICROPHONE SPEECH

Qin Jin[1], *Runxin Li*[1], *Qian Yang*[2], *Kornel Laskowski*[1], *Tanja Schultz*[1,2]

[1]Language Technologies Institute, Carnegie Mellon University, USA
[2]Fakultät für Informatik, Universität Karlsruhe (TH), Karlsruhe, Germany

ABSTRACT

The field of speaker identification has recently seen significant advancement, but improvements have tended to be benchmarked on near-field speech, ignoring the more realistic setting of far-field-instrumented speakers. In this work we present several findings on far-field speech from the MIXER5 Corpus, in the areas of feature extraction, speaker modeling, and multichannel score combination. First, we observe that minimum-variance distortionless response (MVDR) features outperform Mel-frequency cepstral coefficient (MFCC) features, and that fundamental frequency variation (FFV) features offer complimentary information to both MFCC and MVDR features. Second, we present evidence that factor analysis significantly improves system performance, compared to the more traditional GMM/UBM strategy. Third, we find that frame-based score competition significantly improves performance under mismatched conditions with multiple channels available.

Index Terms— Speaker Identification, Distant Speech, Far-field Speech, Front-end Features, Factor Analysis

1. INTRODUCTION

Speaker identification (SID) is the process of automatically inferring speaker identity information from the speech signal. Over the years, automatic speaker identification has developed into a mature technology, crucial to a growing variety of spoken language applications [1-3].

Despite advances, SID systems still lack robustness: their performance degrades dramatically when the acoustic training data is mismatched to the test conditions [4-6].

In this work, our starting point is a state-of-the-art baseline system relying on Mel-frequency cepstral coefficient (MFCC) features and Gaussian Mixture Models/Universal Background Model (GMM/UBM) speaker modeling. We present new frontend features and speaker modeling techniques for speaker identification with distant microphone speech. The new frontend features include minimum variance distortionless response (MVDR) features and fundamental frequency variation (FFV) features. We also apply factor analysis for speaker modeling instead of the more traditional GMM/UBM technique. Finally, we explore frame-based score competition which brings significant gain under mismatched conditions when multiple acoustic/channel conditions are available.

2. MIXER5 DATA

In this paper, we conduct our experiments on the MIXER5 corpus [7], which is a new data collection with cross-channel recordings of face to face interviews used for speaker recognition evaluation undertaken by the Linguistic Data Consortium (LDC). The purpose of the MIXER5 collection was to collect conversational speech in a variety of settings. The interviews were conducted at the LDC in Philadelphia, PA and at ICSI in Berkeley, CA. All participants took part in three separate sessions, each of which involved two 30-minute interviews, separated by a 30-minute break. Further details regarding this corpus can be found in [7]. Figure 1 shows the microphone setup in the interview room. The setup includes 14 microphone channels at several distances from the speaker location. In this paper we only used the distant microphone channels labeled as from 04 to 12.

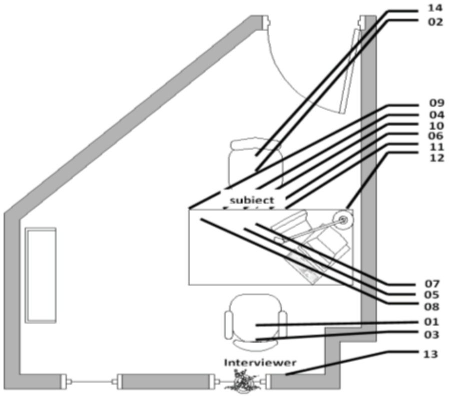

Figure 1: Microphone setup in the MIXER5 data collection

3. FRONT-END FEATURES

3.1 Baseline Features

The baseline front-end features are MFCCs, which are commonly used in speech and speaker recognition systems. We use the first 20 coefficients in this paper. Speakers are modeled with Gaussian mixture models (GMMs), whose

978-1-4244-4296-6/10/$25.00 ©2010 IEEE

training involves adaptation away from a universal background model (UBM) [15].

3.2 Minimum Variance Distortionless Response (MVDR)

As an alternative to MFCCs, we explore warped minimum variance distortionless response (MVDR) cepstral coefficients [9]. The latter have been shown to offer superior speech recognition performance in adverse acoustic conditions [8]. Although speech recognition and speaker recognition exhibit the divergent requirements of speaker-independent phoneme-discrimination and of speaker discrimination, respectively, improvements in one field have occasionally found application in the other. The flowchart in Figure 2 compares the MFCC-based front end and warped MVDR-based front end for speaker recognition. In order to compute the warped MVDR cepstral coefficients (WMVDRCC), we replace the Fourier transformation, including the Mel-scale filter bank, with warped MVDR spectral estimation [8].

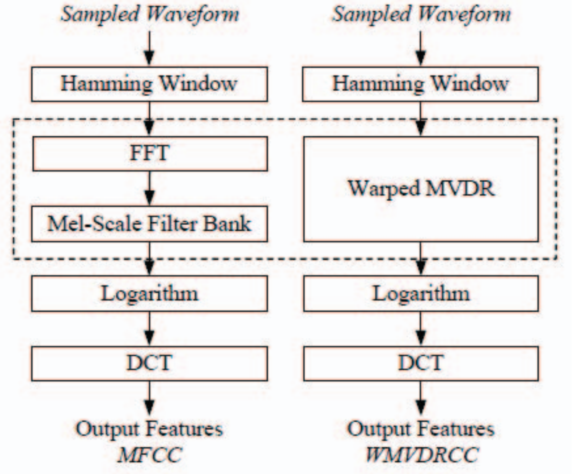

Figure 2: MFCC and WMVDRCC computation

The MVDR representation is governed by a free parameter, namely the model order, which influences spectral resolution. A higher model order shows more detail of the fine structure of the spectrum and partially captures fundamental frequency information, while a low model order reduces the influence of the excitation and the transfer function of the vocal tract. Our previous work [9] indicates that a higher model order is better for speaker recognition on distorted speech. We used order of 40 in this paper.

3.3. Fundamental Frequency Variation (FFV)

The FFV representation is a 7-element characterization of within-frame variation in fundamental frequency. Its computation, which obviates the need to first estimate the fundamental frequency itself, was described in detail in [10]; here, space limitations allow for only a brief account.

Following pre-emphasis ($1-0.97z^{-1}$), the signal is framed into 32 ms overlapping windows, with a frame step of 8 ms. Two frequency spectra, F_L and F_R, are computed for the left and right halves of each frame, respectively, using tapered and largely disjoint windows. Each of the two spectra is then dilated in frequency, over a continuum of dilation factors, while the other spectrum is kept constant. A modified dot-product yields a measure of alignment *g(r)* of their respective harmonic trains, for dilation factor *r*. The representation is then passed through a filterbank; five of the filters represent quickly falling pitch, slowly falling pitch, flat pitch, slowly rising pitch, and quickly rising pitch. The remaining 2 filters are used for normalization via a global whitening transform. In previous work [11], we observed significant performance improvement by combining MFCC and FFV features at the score level.

4. FACTOR ANALYSIS (FA)

Joint Factor Analysis (JFA) [12] can be seen as one of the model compensation methods and has been proved to be very useful in dealing with the channel variability in speaker verification tasks; therefore we applied this technique into our task of speaker identification.

The full JFA consists of three basic components: Eigenchannel, Eigenvoice and classical MAP. However, a heavy computation is required to estimate all the parameters in the complete form. To implement our FA system, we applied similar strategy as in the ALIZE/SpkDet toolkit [13]. In this paper, a simplified variant of the full JFA is applied with the form:

$$\begin{cases} M(S) = M + dz(S) \\ M_h(S) = M(S) + \mu x_h(S) \end{cases} \quad (1)$$

Where *M*, *M(s)* and $M_h(s)$ represent the supervectors for the speaker-independent model, speaker-dependent models and speaker- and utterance-dependent models; *d* is a diagonal square matrix with the same dimension as the supervectors, representing the parameters of classical MAP for speaker models; *z(s)* is a random vector with normal distribution; μ is a transformation matrix with lower dimensions representing the channel space; and $x_h(s)$ is a random vector with normal distribution representing the location of the current utterance in the channel space.

5. FRAME-BASED SCORE COMPETITION (FSC)

The goal of frame-based score competition is to combine information from multiple models. We assume that multiple mismatched models have the potential of better coverage of unknown test space. The key extension in the FSC approach is to use a set of multiple GMM models per speaker, which are trained on audio from multiple microphone channels ("CH"). For each test frame we compare feature vector x_i to the multiple GMMs $\Theta_k^{CH_i}$ for speaker *k*, and choose the highest log likelihood score $LL\left(x_n \mid \Theta_k^{CH_j}\right)_{j=1}^{C}$ to be the frame score; CH_i refers channel *i*; and C is the total number of channels. The likelihood score of the test trial against the model for speaker *k* is given by:

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

The speaker identity is then decided by selecting the k for which $LL(X \mid \Theta_k)$ is maximum. Note that this process makes no assumption about the identity of the test channel. Also, the competition process differs from mono-channel scoring in that per-frame log likelihood scores for different speakers are not necessarily derived from the same channel. Further details are available in [14].

6. EXPERIMENTAL RESULTS

6.1. Experimental Setup

We conducted our experiments results on the MIXER5 data collected at LDC. We used only the distant microphone channels, which are the nine channels labeled 04 to 12 in Figure 1. There are in total 66 speakers (39 female and 27 male). The speaker models are trained on speech data from session 2 and tested on speech data from session 3. We define two train-test conditions:

- Long-Long: 90-sec of training and 30-sec of test, 983 test trials in total
- Short-Short: 30-sec of training and 10-sec of test, 2949 test trials in total

For FA training (estimation of the transformation matrix), two data sets are used:

- SRE08 development data, including 6 speakers and 288 audio files (8 channels each speaker). This is labeled as SRE08 in the following section.
- MIXER5 ICSI data, including 20 speakers and 280 audio files (14 channels each speaker). This is labeled as ICSI in the following section.

50 channel factors are used in all of our experiments.

6.2. Experimental Results

The results under the Long-Long train-test condition are shown in Figure 3 and 4. From Figure 3 we can see that MVDR features achieve better performance than MFCC features under mismatched condition. The performance of the FFV system alone is not comparable to the performance of the other two systems; however, combination with FFV features is always beneficial, as shown in Figure 4, indicating that FFV information is complementary to that in MFCC or MVDR features. A relative improvement of 19% over the best single system under mismatched condition and a relative improvement of 12% over the best single system under matched conditions are achieved.

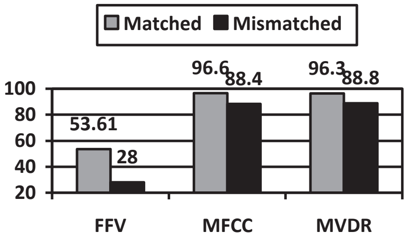

Figure 3: SID accuracy (Long-Long, single features)

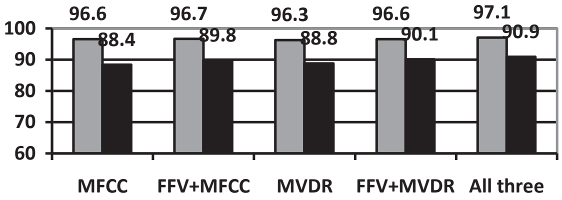

Figure 4: SID accuracy (Long-Long, combined features)

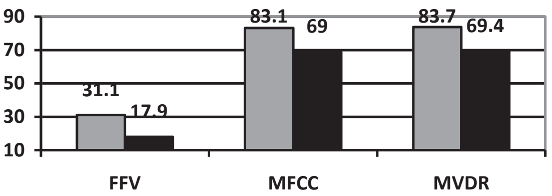

Figure 5: SID accuracy (Short-Short, single features)

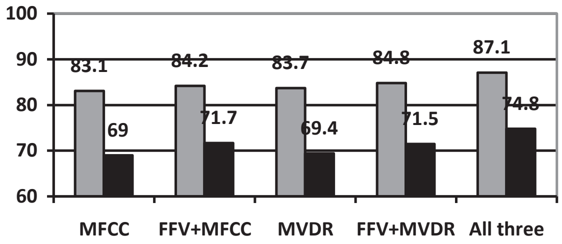

Figure 6: SID accuracy (Short-Short, combined features)

Figures 5 and 6 show the results under the Short-Short train-test condition, in which less audio was used for both training and testing. The trends are similar to those observed for the Long-Long condition; an relative improvement of 18% over the best single system under the mismatched condition, and a relative improvement of 21% over the best single system under the matched condition were achieved. FFV features appear complementary to both MFCC and MVDR features in the Short-Short condition as well.

We applied FSC only in the MVDR system under the mismatched condition, which means that for example when the test channel is 04, then the multiple competing models are models trained on channel 05 to 12 respectively, no model trained on channel 04. We can see from Figure 7 that FSC significantly improves system performance for the mismatched condition. A relative improvement of 42% was observed when using FSC for the best single-feature-type system in the Long-Long condition; for the combined system with three front-end features, FSC improves performance by 29% relatively. Similarly, in the short-short condition, relative improvements of 34% and 20% were observed for the best single-feature-type system and for the combined system, respectively.

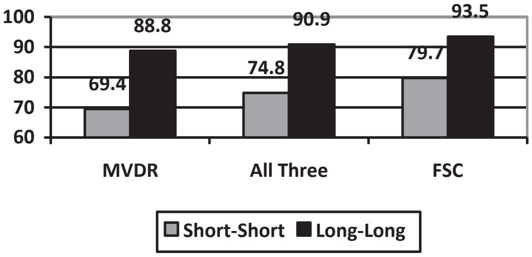

Figure 7: Improvement due to FSC for the best single-feature-type system (MVDR) and the combined system with all three front-end features under mismatched condition

Table 1: SID accuracy (Long-Long and matched)

	baseline	FA-SRE08	impv.	FA-ICSI	impv.
MFCC	96.03%	96.54%	12.8%	97.86%	46.1%
MVDR	95.73%	96.75%	23.9%	98.07%	51.4%

Table 2: SID accuracy (Long-Long and mismatched)

	baseline	FA-SRE08	impv.	FA-ICSI	impv.
MFCC	85.81%	89.94%	29.1%	94.18%	59.0%
MVDR	87.72%	90.84%	25.4%	95.16%	60.6%

Tables 1 and 2 compare the speaker identification performance under the Long-Long train-test condition of the baseline system with GMM/UBM speaker modeling and the system with factor analysis applied. We can see that factor analysis provides significant improvements over the baseline UBM/GMM speaker modeling approach. We all compared the performance of factor analysis based on different FA training data, SRE08 vs. ICSI. FA trained on ICSI data achieved better performance that trained on SRE08 data. We think the reason is that ICSI data as part of the MIXER5 data collection is more similar to the test data. We observed the same trend under the Short-Short train-test condition.

7. CONCLUSIONS

In this paper we conducted speaker identification experiments on the MIXER5 corpus with distant microphone speech. We applied two new sets of features (MVDR and FFV) for speaker feature extraction, factor analysis for speaker modeling, and frame-based score competition for likelihood score computation when multiple distant channels are available. Our results show: (1a) that MVDR features outperform traditional MFCC features under mismatched conditions; (1b) that FFV features are complementary to MFCC and MVDR features, leading to 20% relative improvements; that (2) factor analysis can significantly improve performance compared to the GMM/UBM strategy; and that (3) frame-based score competition can significantly improve performance under mismatched conditions when multiple acoustic/channel conditions are available.

8. REFERENCES

[1] G. Doddington, “Speaker recognition—identifying people by their voices”, in Proceedings of the IEEE, Vol. 73, No. 11, pp. 1651- 1664, 1985.

[2] A. Kanak, E. Erzin, Y. Yemez, A. Tekalp, “Joint audio-video processing for biometric speaker identification”, in Proceedings of Multimedia and Expo. (ICME), pp. 561-4, 2003.

[3] S.E. Tranter, D.A. Reynolds, “An overview of automatic speaker diarization systems”, in IEEE Trans. on Audio, Speech, and Language Processing, Vol. 14. No. 5, pp. 1557-1565, 2006.

[4] C.H. Lee, F.K. Soong, K.K. Paliwal, “Automatic Speech and Speaker Recognition: Advanced Topics”, Springer, 1996, ISBN:0792397061.

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

[6] 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.

[7] L. Brandschain, C. Cieri, D. Graff, A. Neely, K. Walker, “Speaker Recognition: Building the Mixer 4 and 5 Corpora”, in Proceedings of the LREC, 2008.

[8] M. Woefel and J.W. McDonough, “Minimum variance distortionless response spectral estimation, review and refinements,” IEEE Signal Processing Magazine, vol. 22, no. 5, pp. 117–126, Sept. 2005.

[9] M. Wolfel, Q. Yang, Q. Jin, T. Schultz, “Speaker Identification usingWarped MVDR Cepstral Features”, in Interspeech 2009.

[10] K. Laskowski, J. Edlund, and M. Heldner, “An instantaneous vector representation of delta pitch for speaker-change prediction in conversational dialogue systems, ” in Proc. ICASSP, 2008.

[11] K. Laskowski and Q. Jin, “Modeling Instantaneous Intonation for Speaker Identification Using the Fundamental Frequency Variation Spectrum”, in proceedings of ICASSP, 2009.

[12] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, “Factor analysis simplified,” in *ICASSP*, 2005.

[13] J.-F. Bonastre, N. Scheffer, D. Matrouf, C. Fredouille, A. Larcher, A. Preti, G. Pouchoulin, N. Evans, B. Fauve, J. Mason "ALIZE/SpkDet: a state-of-the-art open source software for speaker recognition" in Proc. Odyssey, 2008.

[14] Q. Jin, T. Schultz, and A. Waibel. “Far-field Speaker Recognition”, in IEEE transactions on Audio, Speech, and Language Processing (TASL), Vol. 15, No. 7, September, 2007.

[15] D. Reynolds, T. Quatieri, and R. Dunn, “Speaker verification using adapted Gaussian mixture models,” in *Digital Signal Processing*, 2000, vol. 10, pp. 19–41.