SPEAKER IDENTIFICATION WITH DISTANT MICROPHONE SPEECH

Qin Jin, Runxin Li, Qian Yang, Kornel Laskowski, and Tanja Schultz

Motivation

Speaker identification (SID) technologies have been substantially advanced in the past decades. Despite advances, SID systems still lack robustness: their performance degrades dramatically when the acoustic training data is mismatched to the test conditions. In this paper, We present new frontend features and speaker modeling techniques for speaker identification with distant microphone speech. Including:

❑ minimum variance distortionless response (**MVDR**) features

❑ fundamental frequency variation (**FFV**) features

❑ factor analysis for speaker modeling on top of GMM/UBM

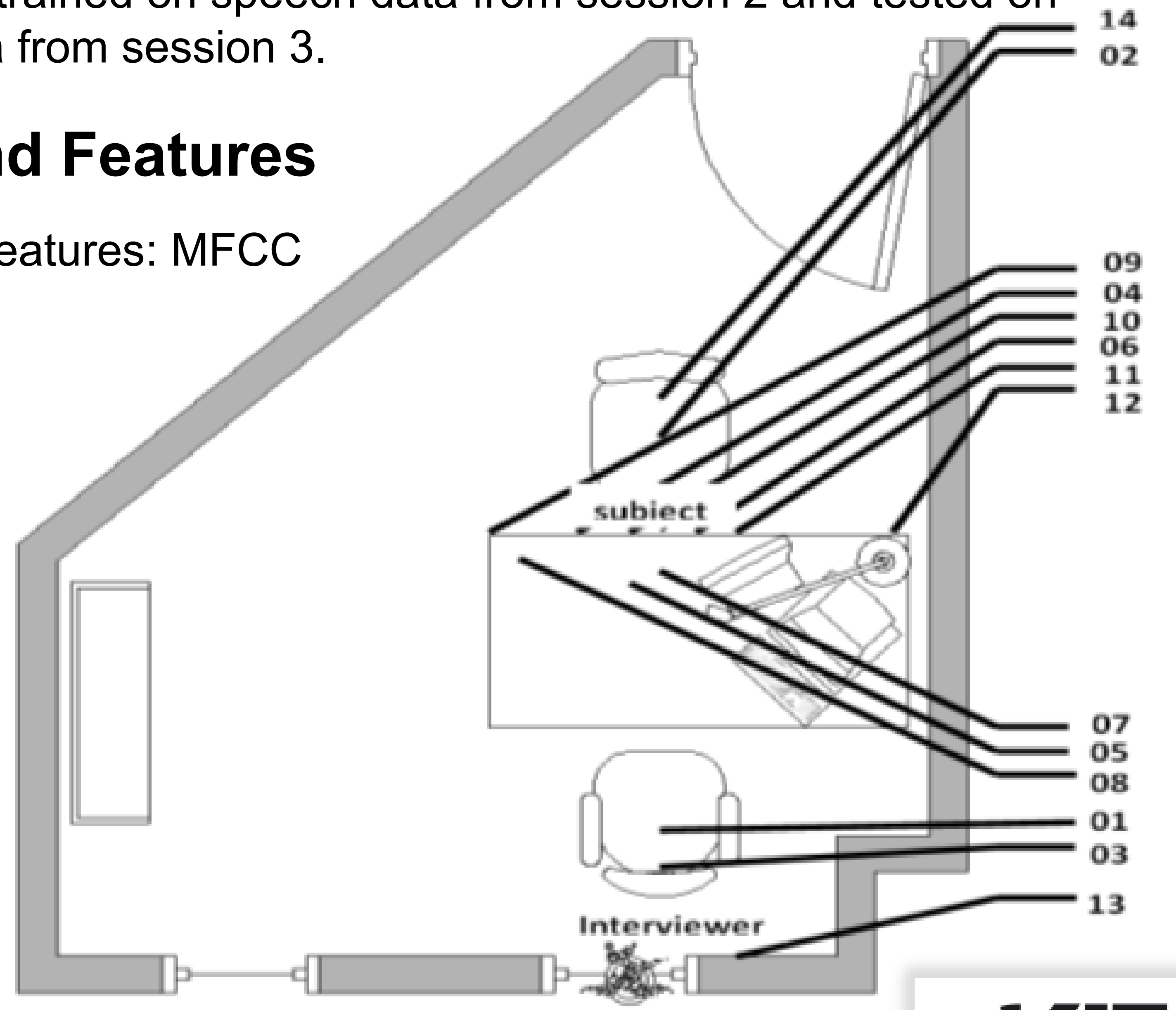
❑ frame-based score competition

MIXER5 Data

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 interviews were conducted at the LDC in Philadelphia, PA and at ICSI in Berkeley, CA.

In this paper we only used the data collected at LDC and distant microphone channels labeled as from 04 to 12.

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.



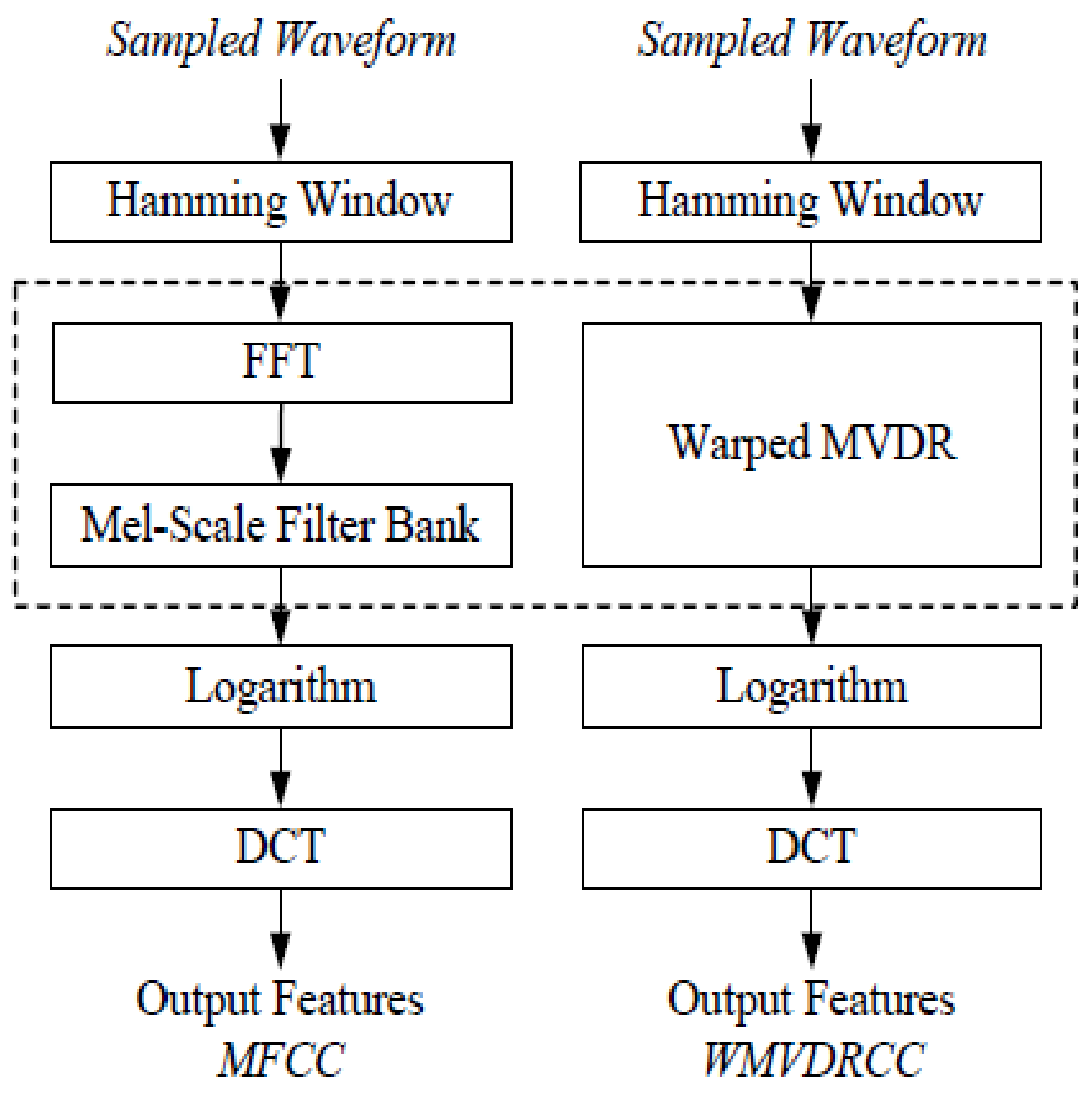
Front-end Features

❑Baseline features: MFCC

MVDR

❑ MVDR features have been shown to offer superior speech recognition performance in adverse acoustic conditions .

❑ In order to compute the warped MVDR cepstral coefficients , we replace the Fourier transformation, including the Mel-scale filter bank, with warped MVDR spectral estimation



FFV

The FFV representation is a 7-element characterization of within-frame variation in fundamental frequency. We observed significant performance improvement by combining MFCC and FFV features at the score level.

Factor Analysis

Joint Factor Analysis (JFA) 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.

Frame-based Score Competition

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.

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

Experimental Setup

We define two train-test conditions:

Long-Long: 90sec training and 30sec test, 983 test trials in total

Short-Short: 30sec training and 10sec test, 2949 test trials in total

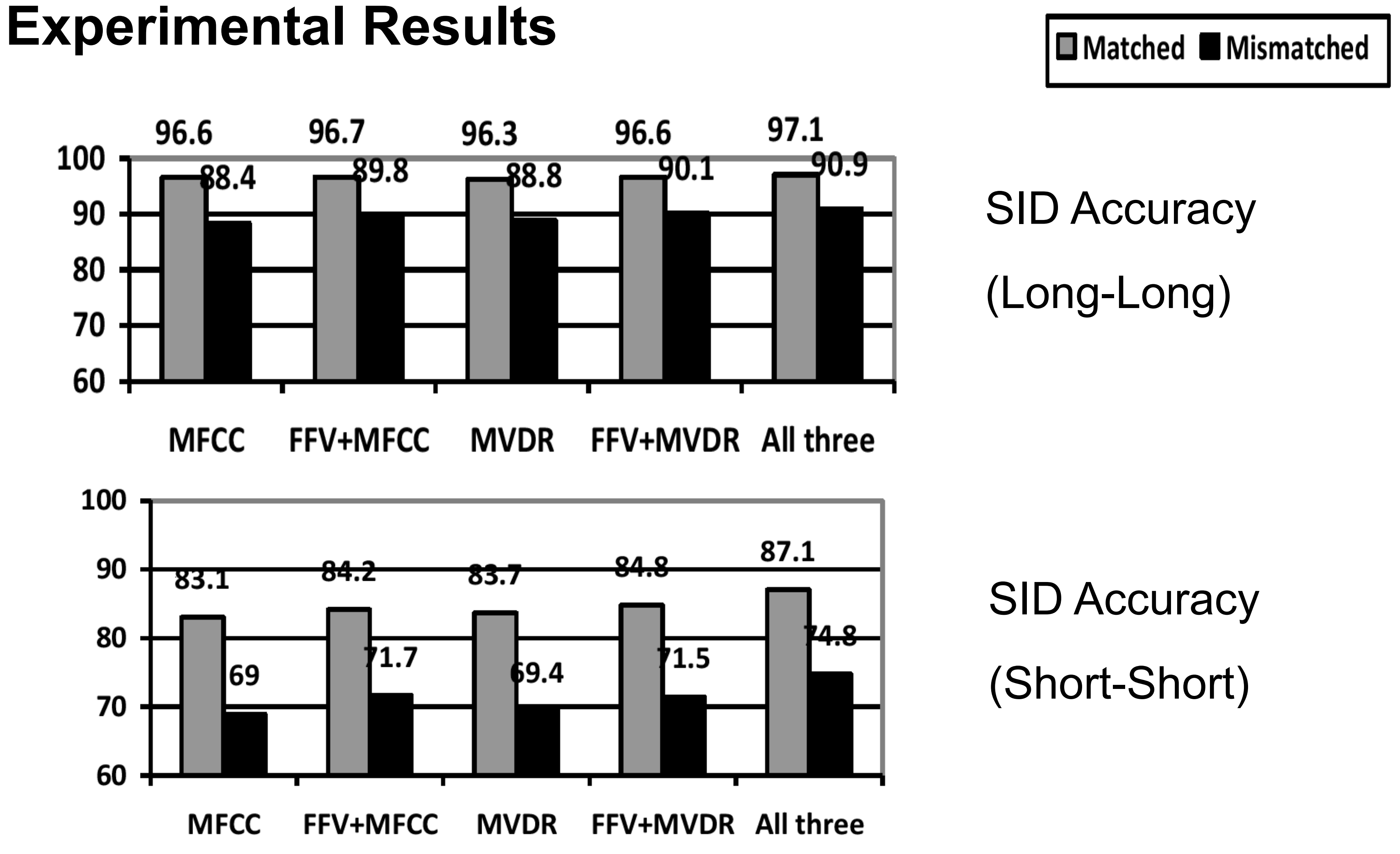
For FA training, two data sets are used:

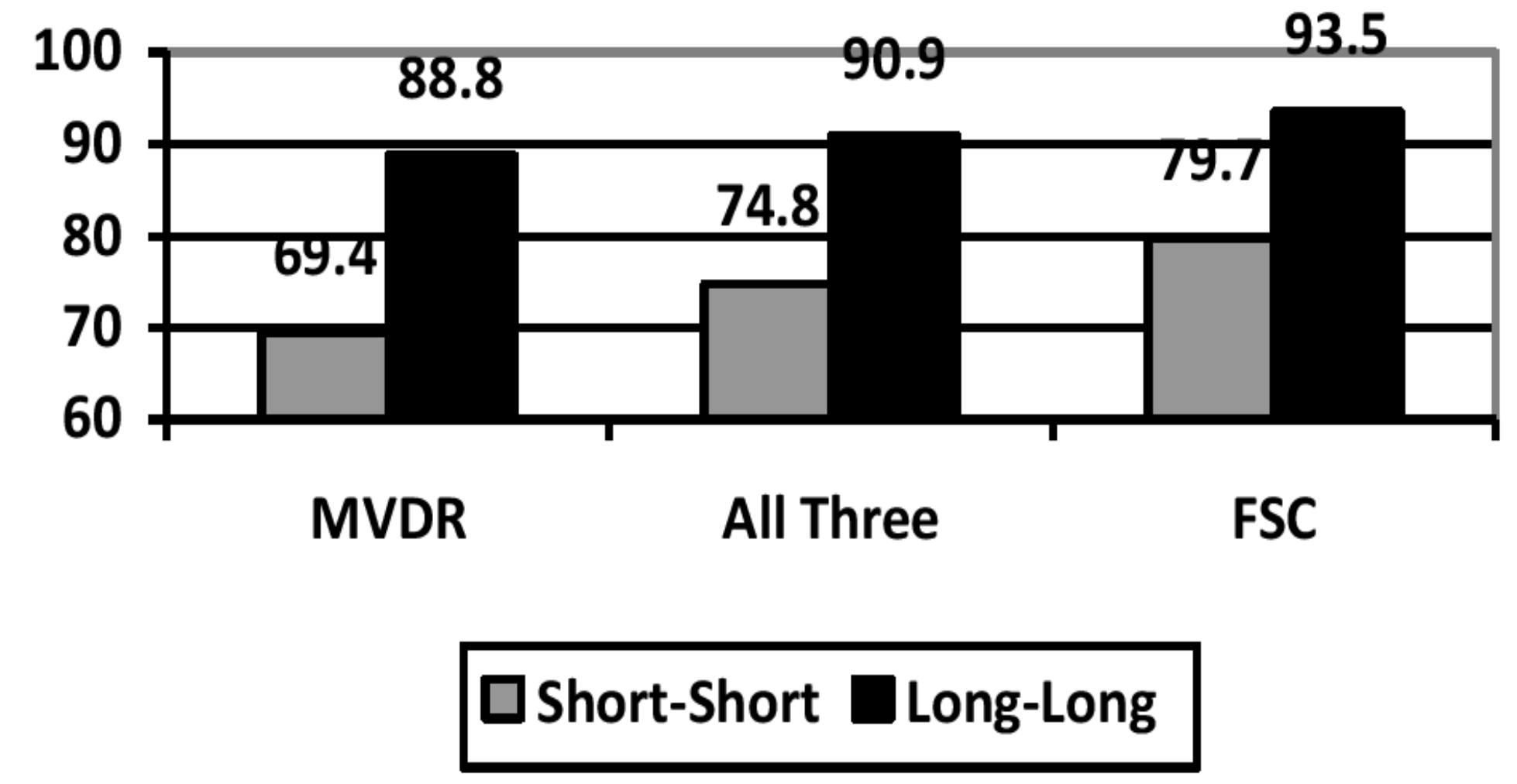
○ SRE08 development data, including 6 speakers and 8 channels each speaker.

○ MIXER5 ICSI data, including 20 speakers and 14 channels each speaker. This is labeled as ICSI in the following section.

50 channel factors are used in our experiments.

Experimental Results





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%

Conclusions

❑ Two new sets of features for speaker feature extraction on distant speech

- ○ MVDR outperform MFCC features under mismatched conditions
- ○ FFV features are complementary to MFCC and MVDR features

❑ FA can significantly improve performance over the GMM/UBM strategy

❑ FSC can significantly improve performance under mismatched conditions