

AMADEUS

USER VOICE IDENTIFICATION  
(URVIN)

Aladdin Ariyaeinia and Amit Malegaonkar  
University of Hertfordshire

Speech Spectral Features

The classical approach to speaker recognition is through the use of short-term spectral templates. The approach involves applying an appropriate analysis to a spoken utterance to generate a sequence of short-term spectral feature vectors. Such templates have been found to contain significant voice characteristics, and may therefore be used to effectively discriminate amongst speakers. Amongst various types of speech features, LPCC (linear prediction-based cepstral coefficients) and MFCC (mel-frequency cepstral coefficients) have been found to be superior for speaker recognition. This section presents the experimental investigations conducted into the relative effectiveness for speaker recognition of the above feature types.

Relative Effectiveness of LPCC and MFCC

In this study, the effectiveness of LPCC and MFCC features are compared for the purpose of open-set, text-independent speaker identification (OSTI-SI). All the experiments are performed using the TIMIT database. The study also includes investigations into the usefulness of score adjustment based on the unconstrained cohort normalisation (UCN) method.

**Database Division**

TIMIT is a clean speech database with 630 speakers. There are 438 male and 192 female speakers, each with 10 utterances. For this experimental study, speakers are divided into 2 equal groups. Each group has 219 males and 96 females. These are called known and unknown speaker pools respectively. For each speaker in the known speaker pool, 9 out of 10 available utterances are used for training a model. The last utterance is used as a test utterance. In the case of the unknown pool, all 10 utterances from each speaker are used as test utterances. Therefore, there are 315 known and 3150 unknown test utterances.

**Experimental set up**

<b>Parameter</b>	<b>Type</b>
Database for experiments	TIMIT
Nature of speech data	Clean speech
Number of Known (Registered) Speakers	315 (219 Males, 96 Females)
Number of Unknown Speakers	315 (219 Males, 96 Females)
Number of Known Speaker Utterances (Known Tests Utterances)	315 ( 1 per speaker)
Number of Unknown Speaker Utterances (Unknown Test Utterances)	3150 ( 10 per speaker)
Features under consideration	LPCC, MFCC
Speaker Modelling	GMM with 32 components
Training data duration	35-40 seconds
Test data segment duration	2-4 seconds
Performance Measure - system	Identification Error, Open-Set Equal Error Rate

## Features

The procedures for feature extraction are as follows.

- LPCC  
Linear prediction coefficients are obtained for each frame using Durbin Recursive method. These coefficients are then converted to cepstral coefficients.
- MFCC  
Fast Fourier Transform is computed for each frame and then weighted by a Mel-scaled filter bank. The filter bank outputs are then converted to cepstral parameters by applying the discrete cosine transformation.

## Unconstrained Cohort Normalisation

This score normalisation method aims to reduce the effect of utterance degradation. This is achieved at the score level by parting known and unknown score distributions further apart. Each score is normalised by choosing mean of the best N scores obtained from rest of the models, where, N is the cohort size. The attraction of this method is that, no offline work is necessary for normalisation as it is based on test scores only.

## Feature Performance in Open-Set, Text-Independent Speaker Identification

Given a set of registered speakers and a sample utterance, open-set speaker identification is defined as a twofold problem. Firstly, it is required to identify the speaker model in the set, which best matches the test utterance. Secondly, it must be determined whether the test utterance has actually been produced by the speaker associated with the best-matched model, or by some unknown speaker outside the registered set.

As shown below, stage 1 is a closed-set identification process where a possible speaker is selected based on the maximum score among the set of N scores. Then this possible score is sent to stage 2 where it is compared against a threshold to decide if it really belongs to the nominated registered speaker. There are two types of errors associated with the process. Error in stage 1 is an identification error. This arises when a registered speaker scores badly against his/her own model.

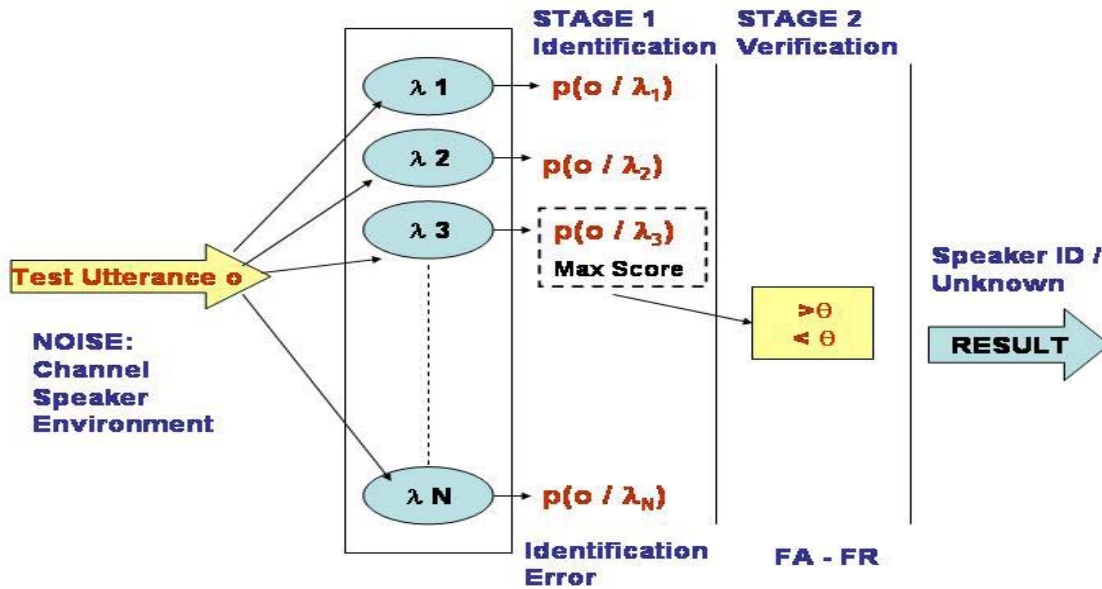
$$IDE = \{1 - [(correct\ identifications) / (total\ tests)]\} \times 100\%$$

Error in stage 2 is a general verification error with a possibility of false acceptance or false rejection.

## Testing Procedure

- Speech is passed through a voice activity detector (VAD) to remove silences and is pre-emphasised.
- Required features are extracted from training as well as testing utterances.
- Gaussian Mixture Models (GMMs) are generated using the training utterances for each known speaker.
- Test utterances of all speakers are compared against the GMMs of the registered speakers to generate scores.
- Identification error and open-set EER are calculated from scores.
- EERs are recalculated using UCN.

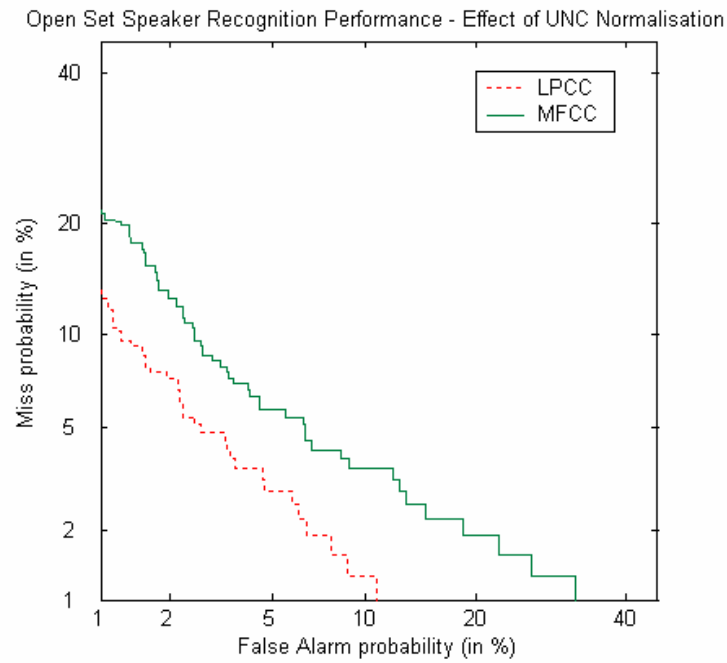
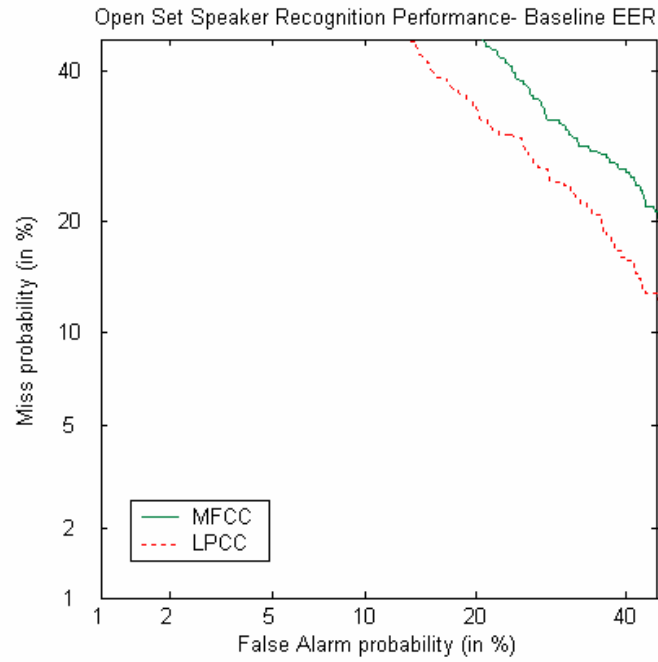
## OPEN SET IDENTIFICATION



## Results

Feature Type	IDE (%)	Without UCN	With UCN
		OS-EER (%)	OS-EER (%)
LPCC	0.31	26.98	3.65
MFCC	1.58	31.46	5.49

It is seen that, in general, LPCC performs better than MFCC at both Stage 1 and Stage 2 of the OSTI-SI process. The DET Plots for OS-EER with and without the use of UCN are given below.



### Summary of the Results

- LPCC are more effective than MFCC for speaker recognition.
- Unconstrained cohort normalisation improves the performance considerably.