4. The Proposed Speaker Verification System

The proposed speaker verification system was described very briefly in the introduction (chapter 1). In this chapter, we present a detailed description of the system. Figure 4.1 shows a general scheme of the speaker verification system.

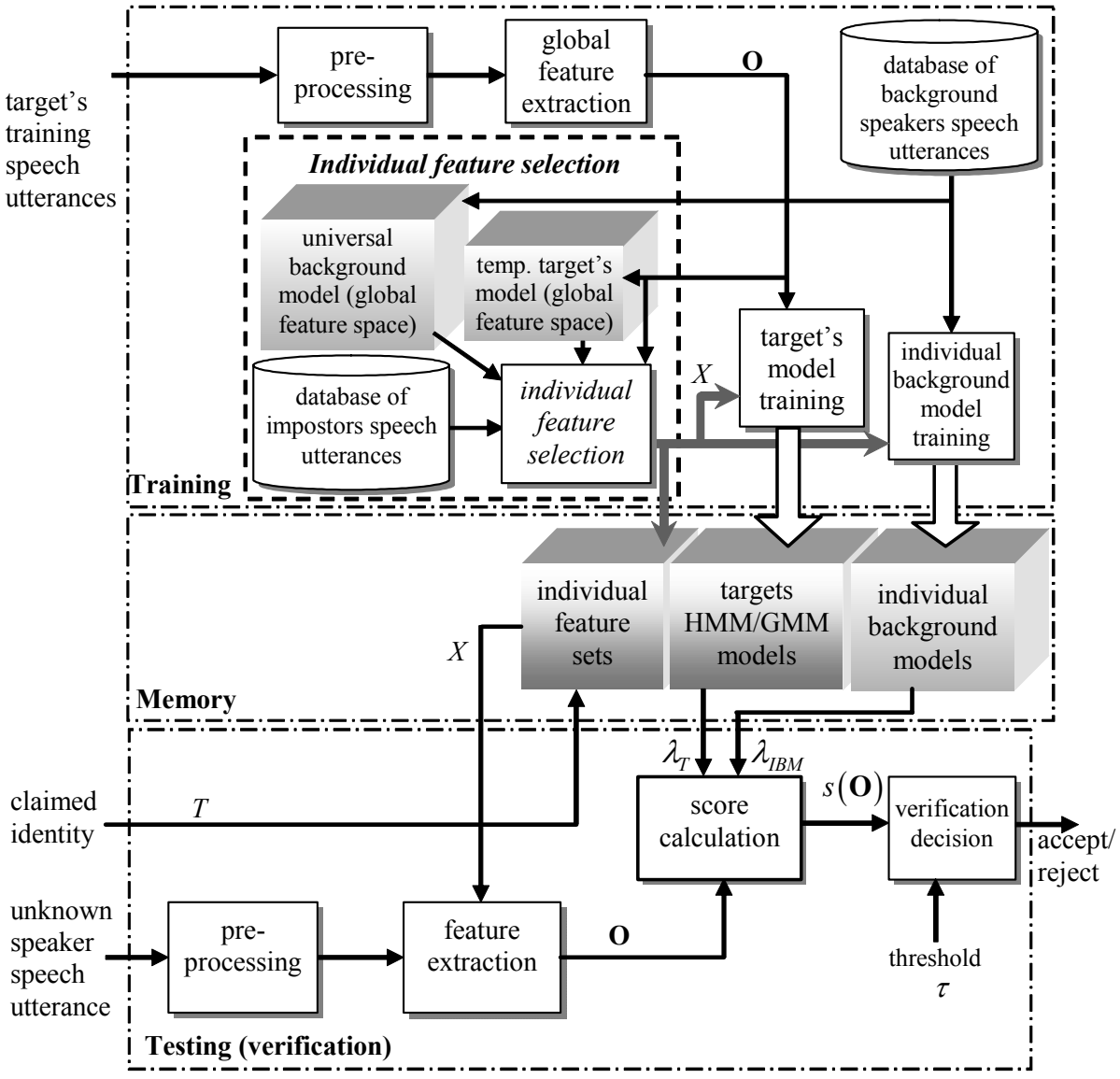

Figure 4.1: The proposed speaker verification system

The proposed speaker verification system consists of two main parts (stages): training and testing (verification).

4.1. The Training Stage

The purpose of the training stage is to train a model for each one of the target speakers. In the proposed system, the model of each target is trained in its own "individual" feature space. An Individual Background Model (IBM) is also trained in this stage. This system uses HMM for text-dependent tasks and GMM for text-independent tasks. The outputs of this training stage (for each target) are:

(1) A set of indices, which represents the selected individual feature sub-set,

(2) Target HMM/GMM model,

(3) Individual HMM/GMM background model; this model is trained for each target in its individual feature space.

These outputs are stored in the memory.

The input to the training stage is the speaker's (target's) training speech utterances (the signals). The input signals undergo pre-processing and global feature extraction (set of all pre-determined features). This pre-processing and feature extraction is sometimes referred to as front-end processing. The training of each target involves the extraction of a high dimensional feature space (termed here the "global feature space"), from which the individual's optimal feature sub-space will be extracted.

The training database consists of three different sub-bases:

(1) Target's training speech utterances;

(2) Background speakers' speech utterances;

(3) Impostors' speech utterances.

The target's training database is divided into two parts: the model training database and the optimization (validation) database. The first part is used for training two target's models:

(1) A temporary target model, using the global feature space, in the individual feature selection process (optimization process),

(2) The final target's model, using its individual feature space (after the feature selection process), which is stored in the memory.

The second part of the target's training database is used to calculate the target scores, $s(\mathbf{O}_T)$, in the individual feature selection process.

The background speakers' speech utterances database is used for training two background models[1]:

(1) A Universal Background Model (UBM), incorporating the global feature space, which is used for score normalization in the individual feature selection process,

(2) The final Individual Background Model (IBM), using the current target's individual feature space, which is stored in the memory.

The impostors' speech utterances database is used to calculate the impostors' scores, $s(\mathbf{O}_I)$, in the individual feature selection process.

4.1.1. Front-End Processing and the Global Feature Set

Conventional front-end processing is employed in the system. First, the speech signal is windowed by a 30 ms Hamming window with a 15-ms frame rate. A speech activity detector is then used to discard silence–noise frames. The speech activity detector is a self-normalizing, energy-based detector. Next, a global set of feature vectors is extracted from

[1] Background models are required for the normalization of the score. Such normalization provides a "dynamic threshold" that reduces dependency on the text, its duration, channel distortion, noise, and more.

the speech frames. The global feature set was chosen to contain K = 120 features from 10 groups of 12 order features. Table 4.1 lists the overall set of features and their assigned symbols.

Table 4.1: The features and their symbols.

#	Feature name	Order	Symbols
1	Mel Frequency Cepstral Coef. (MFCC)	12	$m_1 \div m_{12}$
2	Linear Prediction Cepstral Coef. (LPCC)	12	$c_1 \div c_{12}$
3	Log Area Ratio (LAR)	12	$a_1 \div a_{12}$
4	Linear Prediction Coef. (LPC)	12	$l_1 \div l_{12}$
5	Partial Correlation (PARCOR)	12	$p_1 \div p_{12}$
6	First diff of MFCC (Δ - MFCC)	12	$\Delta m_1 \div \Delta m_{12}$
7	First diff of LPCC(Δ - LPCC)	12	$\Delta c_1 \div \Delta c_{12}$
8	First diff of LAR (Δ - LAR)	12	$\Delta a_1 \div \Delta a_{12}$
9	First diff of LPC (Δ - LPC)	12	$\Delta l_1 \div \Delta l_{12}$
10	First diff of PARCOR (Δ - PARCOR)	12	$\Delta p_1 \div \Delta p_{12}$
	Total number of features:	120	

The MFCC features [Davis and Marmelstein, 1980] were chosen since they are commonly used in speaker verification/recognition systems. In our work, no Cepstral Mean Subtraction (CMS) is added. The other features [Deller et al., 1993] were chosen due to ease of estimation. Since the goal of the work is mainly proof of concept, only 120 features were used, in order to reduce the calculation time in the feature selection process. In future work, we plan to include other features, such as PLP's [Hermansky et al., 1992]. The features have been normalized to their standard deviation in the feature extraction process to improve results.

An algorithm for individual feature selection is executed on the global feature set in order to obtain the optimal individual feature space (X). For each speaker, an index of the selected features is stored and an HMM/GMM target model is trained in that individual feature space along with the target individual background model.

4.1.2. The Individual Feature Selection Procedure

The proposed individual feature selection procedure is an integral part of the training stage. The procedure uses the proposed *Recognition Related Criterion* (*RRC*), which was described in section 3.4.2. The individual feature selection procedure using the *RRC* requires two sets of scores: target scores, $s(\mathbf{O}_T)$, and impostor scores, $s(\mathbf{O}_I)$. The equations of the scores are

$$s(\mathbf{O}_T) = \log p(\mathbf{O}_T \mid \lambda_T) - \log p(\mathbf{O}_T \mid \lambda_{UBM}) \tag{4.1}$$

$$s(\mathbf{O}_I) = \log p(\mathbf{O}_I \mid \lambda_T) - \log p(\mathbf{O}_I \mid \lambda_{UBM}) \tag{4.2}$$

where λ_T is the target model; λ_{UBM} is the universal background model; $\mathbf{O}_T$ is a sequence of feature vectors extracted from the target's validation database, and $\mathbf{O}_I$ is an impostor sequence of feature vectors. In the feature selection stage, universal background models are used for score normalization rather than cohorts [Tran and Wagner, 2001]. This is done due to computation speed considerations. To calculate the *RRC*, using the above scores (4.1-4.2), in each feature selection step, two appropriate models (λ_T and λ_{UBM}) are used in the current-step feature sub-space. These models are initially trained from a training database using the global feature set (global models), and for each sub-space, the models are derived from the global models, using only the tested features in the sub-space. The individual feature selection procedure requires also a set of target utterances, which are taken from the target's validation database, as well as impostor utterances. Obviously, because of computation-time considerations, one cannot use all the impostors' utterances in the database; rather a small set of impostors' utterances must be selected. This impostor set selection process is described later in section 4.1.3.

Any one of the feature selection methods discussed in section 3.1 may be used here. In our experiments (described later in chapter 5) we use several selection procedures: k-best, forward, SFFS, and DP. We found that in terms of accuracy and computation complexity the best procedure is the SFFS.

After determining the target individual feature space, an HMM/GMM model is retrained in that individual feature space, using the target's model training database. An individual background model is also retrained, using the database of background speakers, in the target's individual feature space.

4.1.3. Impostor Selection for the Feature Selection Procedure

As we previously noted, for the individual feature selection procedure using the *RRC*, a small set of impostor utterances must be selected for each target speaker. For this, we need to select several impostors (cohort). For each one of the target models, *C* "selected" impostors (cohort speakers - *c*) were determined using the Close Impostors Clustering (CIC) method [Zigel and Cohen, 2003] with the following divergence-like criterion:

$$\begin{aligned} d_D(\lambda_T,\lambda_c) = & \frac{1}{N_{\mathbf{O}_T}} \sum_{i=1}^{N_{\mathbf{O}_T}} \left[\log p\left(\mathbf{O}_T^i \mid \lambda_T\right) - \log p\left(\mathbf{O}_T^i \mid \lambda_c\right) \right] \\ & - \frac{1}{N_{\mathbf{O}_c}} \sum_{j=1}^{N_{\mathbf{O}_c}} \left[\log p\left(\mathbf{O}_c^j \mid \lambda_T\right) - \log p\left(\mathbf{O}_c^j \mid \lambda_c\right) \right] \end{aligned} \tag{4.3}$$

where $p\left(\mathbf{O}_T^i \mid \lambda_c\right)$ is the probability of the *i*th target's utterance $\mathbf{O}_T^i$ given the candidate impostor model λ_c. $\mathbf{O}_c^j$ is the *j*th impostor's utterance, and $N_{\mathbf{O}_T}$ and $N_{\mathbf{O}_c}$ are the number of target utterances and candidate impostor utterances, respectively. These models were defined for the global 120-feature space. The cohorts selected in the 120-feature space were used for all sub-spaces required by the feature selection algorithm.

4.1.3.1 The Close Impostors Clustering (CIC) Method [Zigel and Cohen, 2003]

Different cohort selection techniques have been suggested in the literature, such as closest impostors [Rosenberg et al., 1992], which choose the "closest" impostor models to the target model. The main disadvantage of such technique is that it may leave the target exposed from a certain "angle" in the feature space. The CIC [Zigel and Cohen, 2003] is an effective algorithm of choosing cohort that engulfs the target model from all (or as many as possible) space angles.

The goal of the CIC is to select the best C impostor models from the complete impostor set using clustering technique, for each target. The algorithm consists of three main steps: (1) *outliers removal* - the initial step of the algorithm is to select a subset of N impostors ($N \geq 2C$) from the complete impostor community. The subset of N impostors consists of the candidates for cohort. The impostors excluded from this set are outliers and impostors that are very un-similar to the target that may obscure the correct selection of cohort; (2) *clustering* - the subset of N impostors is clustered into C clusters. Any one of several clustering methods may be used; (3) *cohort selection procedure* – one impostor is selected from each cluster as a representative of the given cluster. Any one of several selection methods may be used, for example, selecting the "closest" (to the target) member of the given cluster.

In the version of the CIC used here, a single-link hierarchical clustering [Ripley, 1996] for finding C models has been employed:

(1) Start with initial closest set, $\mathcal{A}(T)$, which has $N = 2C$ models. $C' = 0$.

(2) Find the two closest impostors to each other (m and n) in $\mathcal{A}(T)$:

$$(m,n) = \arg \min_{\substack{m,n \in \mathcal{A}(T) \\ m \neq n}} \{ d_D(\lambda_m, \lambda_n) \} \tag{4.5}$$

(3) Between *m* and *n*, remove from $\mathcal{A}(T)$ the farthest impostor (from the target), *l*, where *l* is found by:

$$l = \arg \max_{l \in \{m,n\}} \{ d_D(\lambda_l, \lambda_T) \} \tag{4.6}$$

$C' = C' + 1$.

(4) Repeat from step (2) until $C' = N - C$.

4.2. The Testing Stage

In the testing stage, an unknown speaker's claimed identity and test utterance are introduced to the system. From the identity claim, the appropriate feature space, *X*, is drawn and feature extraction is made on the pre-processed utterance in order to yield these features, which belong to the speaker feature space. The verification algorithm provides a probabilistic score, $s(\mathbf{O})$, which is compared to a threshold (τ), to yield an accept or reject decision. The score $s(\mathbf{O})$ used here is the log likelihood ratio, $s(\mathbf{O}) = \log p(\mathbf{O} \mid \lambda_T) - \log p(\mathbf{O} \mid \lambda_{IBM})$, where $\mathbf{O}$ is the speech observations, λ_T is the target speaker model; and λ_{IBM} is the individual background model. The decision to accept or reject an identity claim is based on a comparison of the score, $s(\mathbf{O})$, with a threshold, τ :

$$s(\mathbf{O}) \begin{cases} \geq \tau & \rightarrow accept \\ < \tau & \rightarrow reject \end{cases} \tag{4.7}$$

Because each trial is tested in the claimed target feature space, one cannot use a universal threshold, rather individual thresholds must be used.

5. Experiments and Results

This chapter presents the experimental evaluation of the proposed speaker verification system, which uses individual feature space. The experiments were performed for text-dependent speaker verification as well as for text-independent speaker verification. The behavior of the system in a noisy database is examined. Comparison of performances between individual feature spaces and a common MFCC feature space is discussed. This common MFCC space was chosen for comparison because the MFCC is the most popular feature-set used in the speaker-recognition literature.

5.1. Text-Dependent Speaker Verification

5.1.1. Experimental Setup

The experiment was setup for text-dependent speaker-verification task. The model for each speaker was trained as a left-to-right Continuous Density Hidden Markov Model (CD-HMM), with 5 states and 2 Gaussians per state. Individual background models (CD-HMM with 5 states and 2 Gaussians per state) were trained using 26 speakers (one utterance from each speaker). This experiment consisted only of male speakers.

The feature selection procedure was executed for each target with the *RRC* (3.31) criterion using the evaluation database: 20 target utterances and 10 utterances from each of the six cohort impostors. The result of the selection procedure was a set of $k = 24$ features for each target speaker. This feature order of 24 was determined in order to compare the results of the feature selection algorithm with the "almost standard" MFCC feature space (12 MFCCs + 12 ΔMFCCs). Several feature selection procedures were executed: k-best, forward, DP, and SFFS.

5.1.2. The Text-Dependent Database

The algorithm was evaluated with utterances of the Hebrew word /hamesh/ (five), taken from the Hebrew Isolated Digits (HID) database [http:\\www.ee.bgu.ac.il\~spl]. The database contains high quality speech (SNR of 60dB), recorded over a six-months period and sampled at 16KHz with 12 bits resolution.

Ten male speakers from this database having the highest number of utterance repetitions were chosen to be target speakers. For each target, there were 39 male impostors. The number of utterances for each target speaker was between 70 to 400, and the number of utterances for each impostor was 45. The first 20 utterances for each target speaker were used for model training, the next 20 for the feature selection procedure, and the remaining utterances for testing.

5.1.3. Results and Discussion

Figure 5.1 shows the maximum value of the *RRC* criterion (3.31) as a function of the dimension of the selected feature space, *k,* as evaluated by the different feature selection procedures: k-best, forward, DP and SFFS. The data is from the training database. These curves indicate that the worst selection procedure is, as expected, the k-best, followed by the forward selection procedure. The two best selection procedures are the DP and the SFFS. The SFFS yields similar results to the DP, however, it is more efficient than the DP in terms of the calculation load. Therefore, we used the SFFS as the selection procedure for our individual feature selection system.

Figure 5.2 shows EER test results of the various selection methods as a function of the feature space dimension, for speaker #3, using the testing database. One can see that the dimension of $k = 33$ yields the best results (for the SFFS). For dimension sizes above 35, the

EER increases, most probably due to overfitting in the training (sometimes referred to as: "the curse of dimensionality"). In practical situations, one would like to determine the order of the feature space from the training/evaluation data, during the training process.

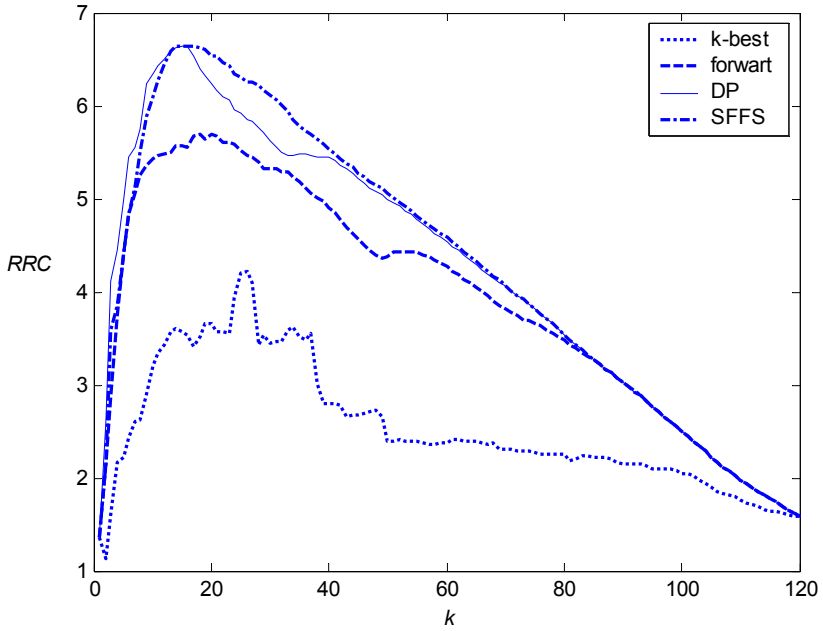

Figure 5.1: Maximum RRC criterion as a function of the feature space dimension, k, for several feature selection procedures (for speaker #3).

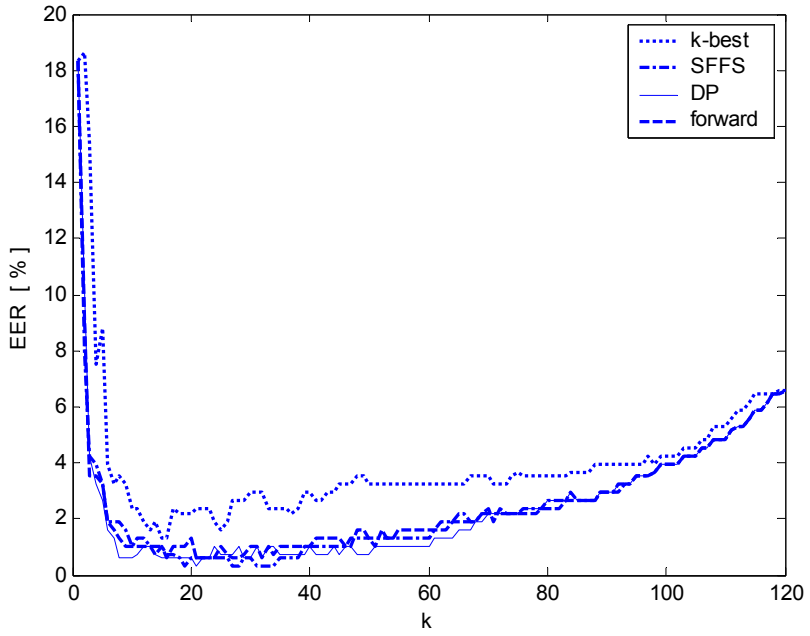

Figure 5.2: Real EER test results of the different feature selection procedures in different feature space dimension (for speaker #3).

Table 5.1 shows the 24 selected feature subsets for each of the first five target speakers, using the SFFS feature selection procedure. The SFFS employed the *RRC* criterion (3.31). The table shows that different feature spaces were selected for the different target speakers, and that the dominant features in the optimal sets belong to the MFCC family.

Table 5.1: Selected features for the first 5 target speakers.

Speaker #	**Selected features**
1	$m_4\ m_5\ m_{10}\ c_8\ a_2\ l_{11}$ $\Delta m_2\ \Delta m_4\ \Delta m_5\ \Delta m_6\ \Delta m_7\ \Delta m_8\ \Delta m_9\ \Delta m_{11}\ \Delta m_{12}$ $\Delta c_3\ \Delta a_2\ \Delta a_{12}\ \Delta l_4\ \Delta p_2\ \Delta p_4\ \Delta p_5\ \Delta p_8\ \Delta p_{10}$
2	$m_2\ m_4\ m_5\ m_8\ m_9\ a_{12}\ l_8\ l_{10}\ l_{12}\ p_{11}$ $\Delta m_1\ \Delta m_6\ \Delta m_7\ \Delta m_9\ \Delta m_{10}\ \Delta m_{11}\ \Delta m_{12}$ $\Delta a_1\ \Delta a_4\ \Delta a_5\ \Delta a_{12}\ \Delta p_1\ \Delta p_4\ \Delta p_5$
3	$m_5\ m_8\ m_9$ $\Delta m_3\ \Delta m_5\ \Delta m_6\ \Delta m_7\ \Delta m_9\ \Delta m_{10}\ \Delta m_{11}\ \Delta m_{12}$ $\Delta a_2\ \Delta a_3\ \Delta a_5\ \Delta a_6\ \Delta a_9\ \Delta a_{10}\ \Delta a_{11}\ \Delta l_{12}\ \Delta p_1\ \Delta p_2\ \Delta p_9\ \Delta p_{10}$
4	$m_3\ m_7\ m_8\ m_9\ m_{10}\ a_4\ a_6\ a_{11}\ l_6\ l_{11}\ p_6\ p_8\ p_{11}$ $\Delta m_4\ \Delta m_5\ \Delta m_8\ \Delta m_{10}\ \Delta m_{12}$ $\Delta a_2\ \Delta a_8\ \Delta l_8\ \Delta l_9\ \Delta l_{10}\ \Delta p_2$
5	$m_4\ m_7\ m_{12}\ a_7\ a_8\ a_9\ a_{10}\ a_{11}\ p_7\ p_8\ p_9\ p_{10}$ $\Delta m_4\ \Delta m_5\ \Delta m_7\ \Delta m_9\ \Delta m_{11}\ \Delta m_{12}$ $\Delta a_1\ \Delta a_7\ \Delta a_{10}\ \Delta a_{11}\ \Delta p_2\ \Delta p_{10}$

Figure 5.3 is a histogram of feature occurrences (only the top probable features) in the individual selected feature subsets (from the ten targets). As the figure shows, most of the selected features belong to the ΔMFCCs, especially the highest order coefficients $\Delta m_4 \div \Delta m_{12}$.

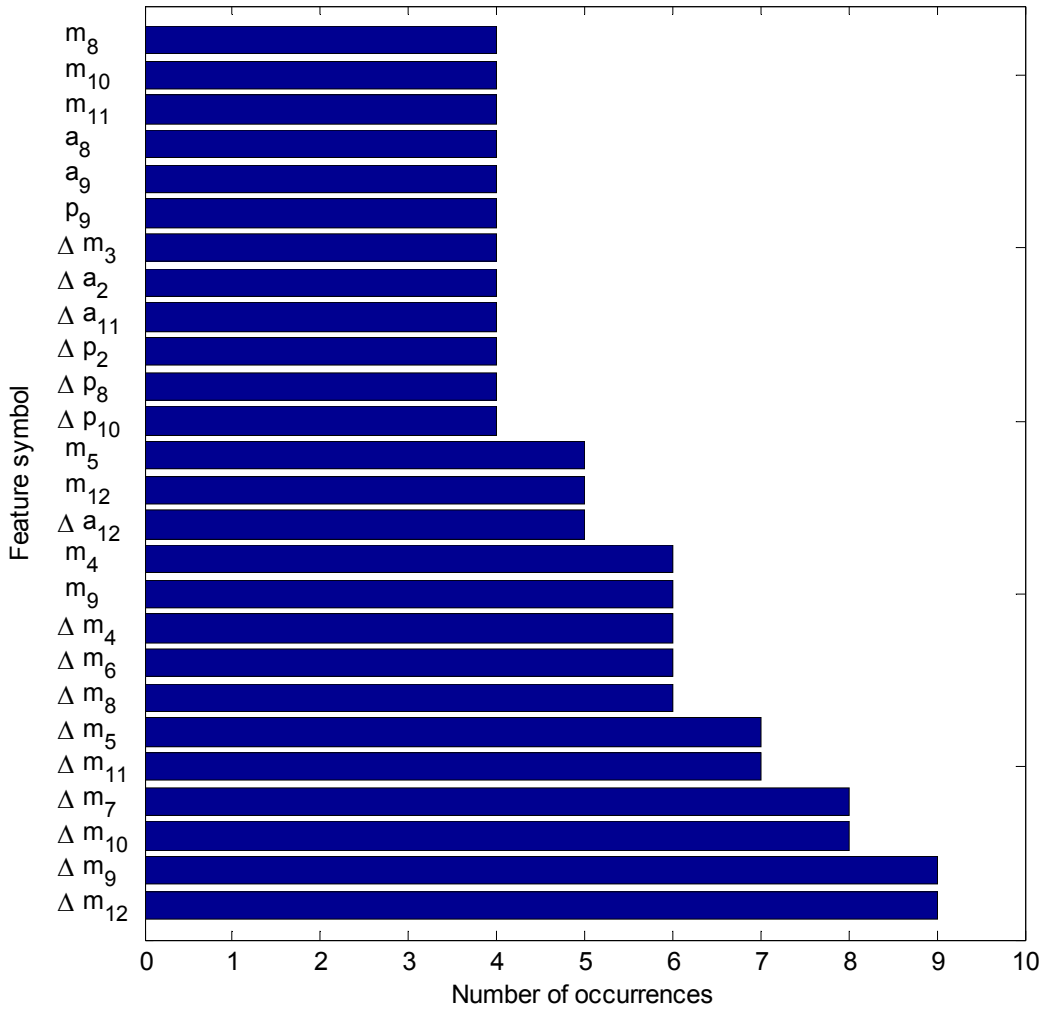

Figure 5.3: Number of feature occurrences in the individual selected feature subsets (ten targets).

Figure 5.4 shows the results of verification experiments obtained with different feature spaces. Results are presented by Detection Error Tradeoff (DET) plots. DET plots show the system tradeoff of misses versus false acceptances. The figure shows the average[2] DET curves of the full set of 120 features and two different (24 dimensional) spaces: the MFCC (12 MFCC + 12 ΔMFCC) feature space and the individual selected feature space. Each curve is an average of ten DET curves of the ten target speakers. The number of target trials is 1734, and the number of impostor trials is 6600.

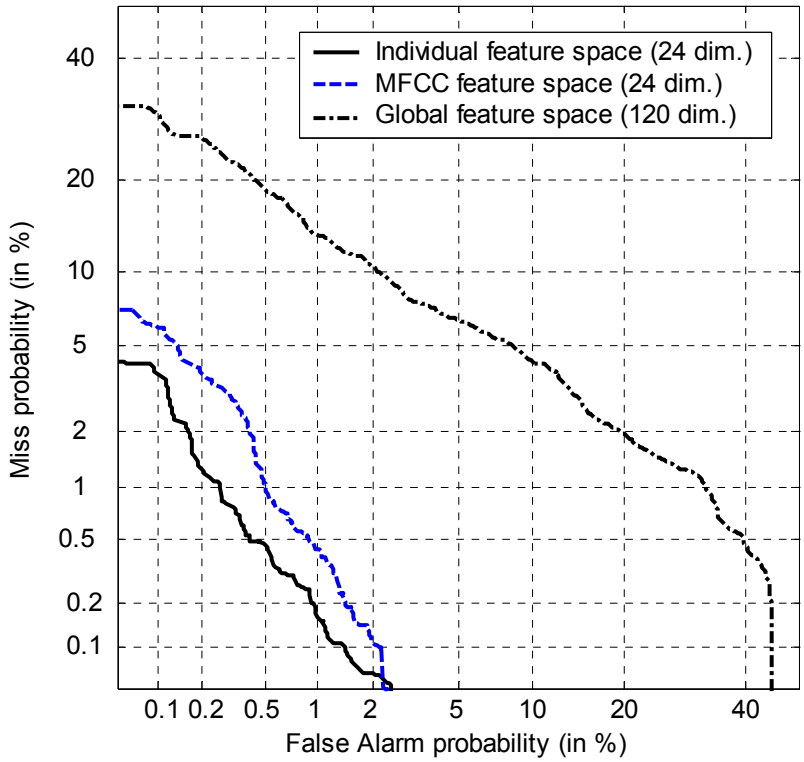

Figure 5.4: Average DET curves of speaker verification results (feature spaces: global 120 features, 24 MFCC and Del MFCC space, and 24 individual optimal space)

[2] Note that the DET curves here are not performed conventionally (universal threshold), since scores for each target are given in a different feature space. Therefore, individual DET curves are calculated individually for each of the targets and average DET curve is performed by averaging all individual DET curves.

According to figure 5.4 the individual selected feature system yields the best results. The worst results were obtained with the overall 120-feature space, probably due to the "curse of dimensionality."

Table 5.2 shows the mean EER values for each tested feature space. An average EER of 0.48% was achieved with the individual selected feature space. This is an improvement of 31% comparing to the 'almost standard' MFCC feature space (average EER = 0.7%).

Table 5.2: Mean equal error rate of the verification results

Feature Space	**Mean Equal Error Rate (EER)** in %
120 features	6
MFCC	0.7
FS	0.48

5.2. Text-Dependent Speaker Verification in Noisy Speech

This section presents two experiments that examine speaker verification performance of the HMM text-dependent speaker verification system with respect to different Signal to Noise Ratios (SNR). The first experiment was performed with a noisy training and a noisy testing database, and the second with a clean training database and a noisy testing database.

5.2.1. Training on a Noisy Database

The database used in this experiment was the same as that used in the previous text-dependent experiment (section 5.1), except that to each speech file (including training, evaluating, and testing speech files), was added white noise in different SNRs of 20dB and 5dB. The background models, as well as the target models, were retrained using the noisy database. The individual feature selection process used the noisy training database and the testing was performed using the noisy testing database.

Table 5.3 shows the 24 feature subsets selected from the **20dB** noisy database, for each one of the first five target speakers (the same speakers as in table 5.1). The individual feature selection process again employed the SFFS selection procedure and the *RRC* criterion (3.31). Again, different feature spaces were selected for the different target speakers. Moreover, the selected features are somewhat different from the "clean" database case (table 5.1). One can see also that the dominant features in the optimal sets belong to the MFCC family.

Table 5.3: Selected features for the first 5 target speakers on 20dB noisy database.

Speaker #	Selected features
1	$m_5\ m_7\ m_9\ m_{10}\ a_3\ a_4\ a_5\ a_9\ l_9\ p_5\ p_9\ p_{10}$ $\Delta m_5\ \Delta m_7\ \Delta m_9\ \Delta m_{10}\ \Delta m_{11}\ \Delta m_{12}$ $\Delta a_3\ \Delta a_{12}\ \Delta l_{12}\ \Delta p_9$
2	$m_2\ m_3\ m_6\ m_8\ m_{11}\ m_{12}\ a_3\ l_3\ l_6\ l_7\ l_{10}\ p_3\ p_6$ $\Delta m_4\ \Delta m_5\ \Delta m_7\ \Delta m_8\ \Delta m_9\ \Delta m_{12}$ $\Delta a_{10}\ \Delta a_{11}\ \Delta l_9\ \Delta l_{10}\ \Delta p_{10}$
3	$m_5\ m_6\ m_7\ m_9\ a_5\ a_9$ $\Delta m_5\ \Delta m_8\ \Delta m_9\ \Delta m_{11}\ \Delta m_{12}$ $\Delta a_1\ \Delta a_2\ \Delta a_4\ \Delta a_8\ \Delta a_{10}\ \Delta a_{11}\ \Delta l_6\ \Delta l_{10}\ \Delta p_2\ \Delta p_4\ \Delta p_8\ \Delta p_{11}$
4	$m_7\ m_8\ m_9\ m_{11}\ m_{12}\ c_2\ a_6\ a_8\ a_9\ l_2\ l_4\ l_9\ l_{12}\ p_8\ p_{10}\ p_{12}$ $\Delta m_5\ \Delta m_8\ \Delta m_{10}\ \Delta m_{11}\ \Delta m_{12}$ $\Delta a_7\ \Delta l_8\ \Delta p_7$
5	$m_5\ m_{10}\ m_{11}\ m_{12}\ a_3\ a_{10}\ a_{12}\ l_3\ p_1\ p_3\ p_{10}$ $\Delta m_2\ \Delta m_3\ \Delta m_5\ \Delta m_9\ \Delta m_{10}\ \Delta m_{11}$ $\Delta c_7\ \Delta a_5\ \Delta a_9\ \Delta l_{10}\ \Delta p_5\ \Delta p_6\ \Delta p_7$

Figure 5.5 is a histogram of the top probable feature occurrences in the individual selected feature subsets (from the ten targets), using the 20dB noisy database. From this figure one can see that, like the "clean" database (figure 5.3), most of the selected features belong to the ΔMFCCs. However, the MFCCs are more prominent now than in the "clean" database case.

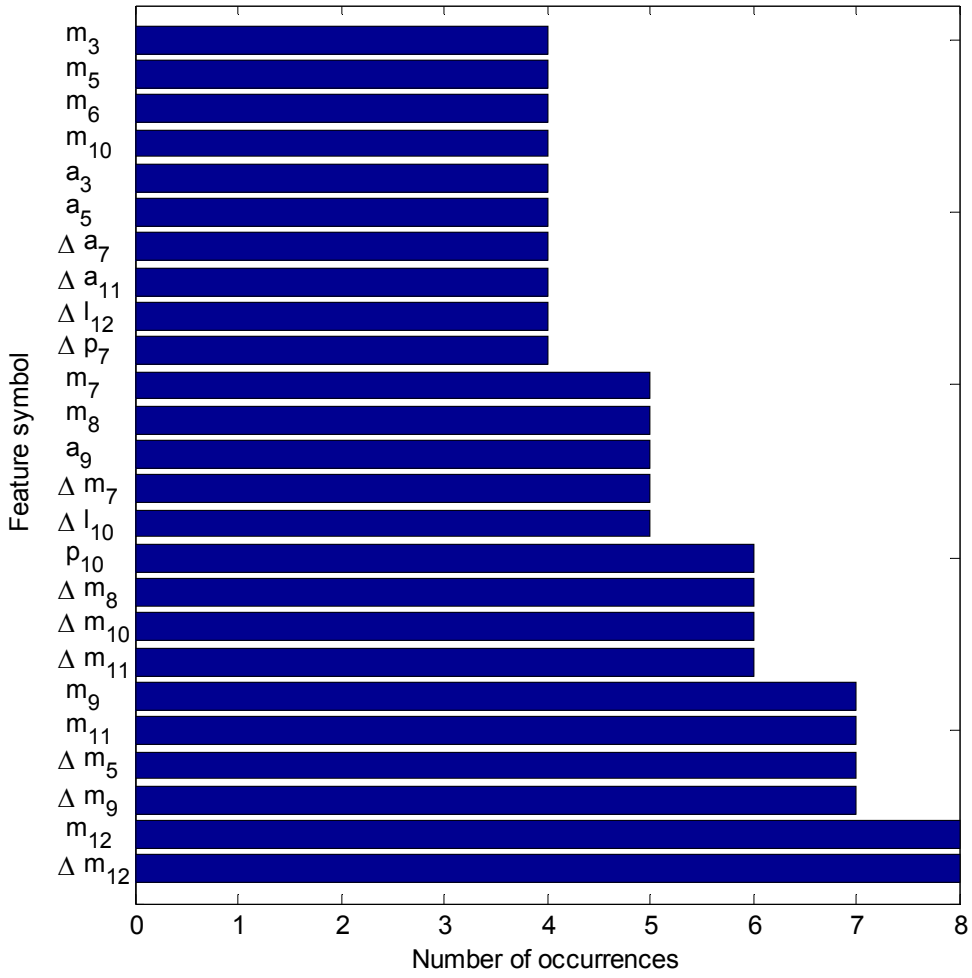

Figure 5.5: Number of feature occurrences in the individual selected feature subsets (ten targets) – using the 20dB noisy database.

Figure 5.6 shows the results of the verification experiments using the 20dB SNR database, obtained with two different 24-dimensions feature spaces: 1) the individual feature

spaces, and 2) MFCC feature space. The figure shows that the best results, like the "clean" database, are achieved with the individual feature space (EER of 2.93%). The MFCC feature space yielded an EER of 3.43%.

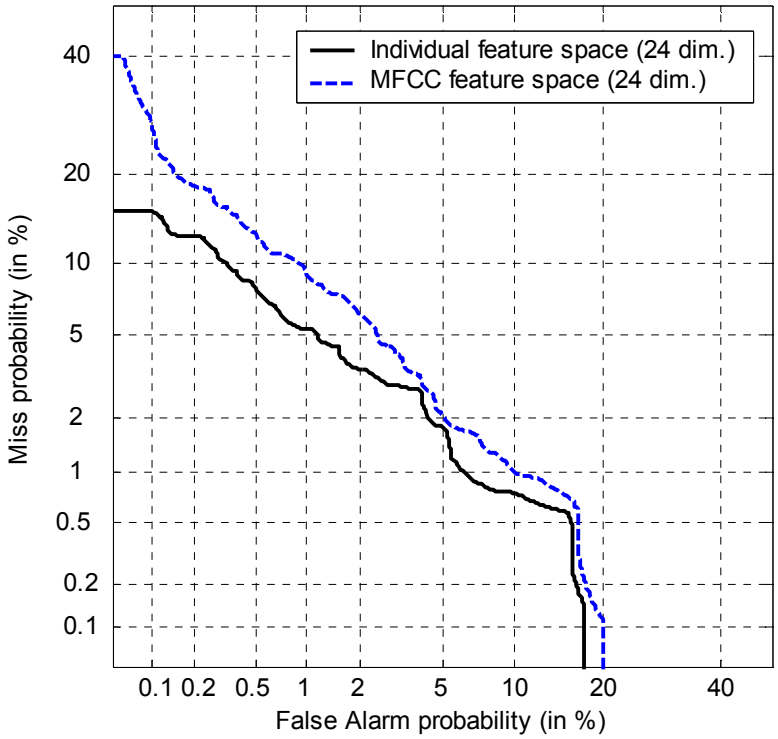

Figure 5.6: Average DET curves of speaker verification results in 20dB noisy database.

A similar experiment was performed using **the 5dB** noisy database. Figure 5.7 is a histogram of feature occurrences in the individual selected feature subsets, using the 5dB noisy database. As in the "clean" and 20dB noisy database cases (figure 5.3 and figure 5.5), most of the selected features belong to the MFCC family. However, in this case the MFCCs are more prominent than the ΔMFCCs. Moreover, the feature distribution is more uniform than for the two other SNRs. The best one feature, which is most common in all the three SNR cases and almost in all the targets' individual feature spaces is Δm_{12}.

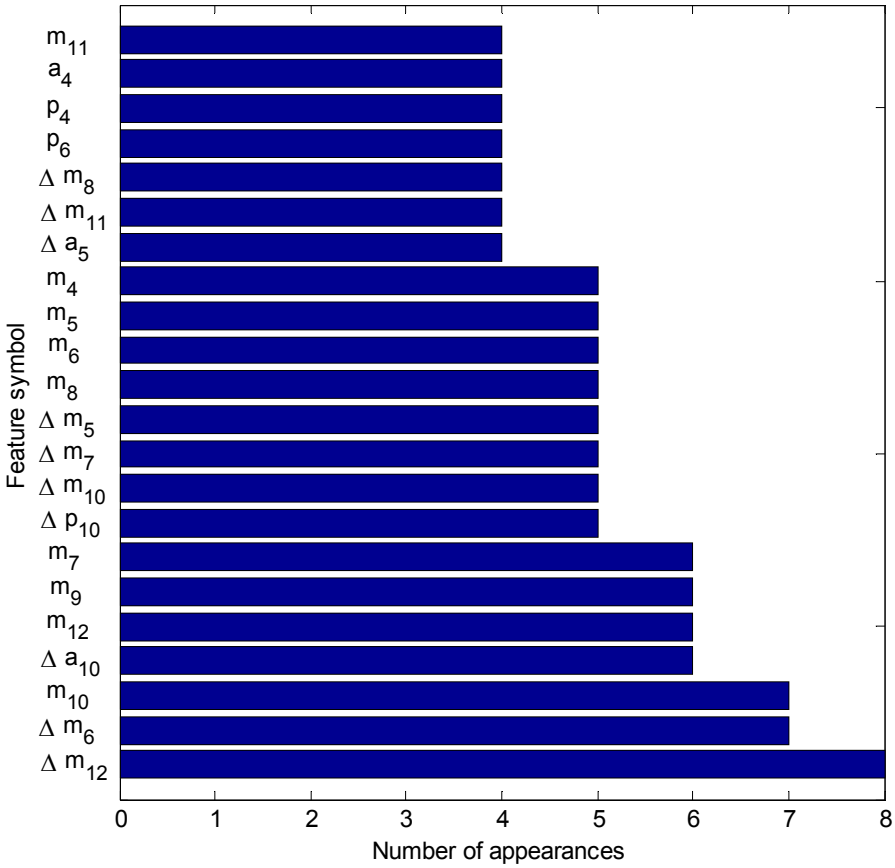

Figure 5.7: Number of feature occurrences in the individual selected feature subsets (ten targets) – using the 5dB noisy database.

Figure 5.8 shows the results of the verification experiments using the 5dB SNR database, obtained with two different 24-dimension feature spaces: 1) the individual feature spaces, and 2) the MFCC feature space. From this figure one can see that the results are similar. However, in the EER point the MFCC feature space (EER of 10.45%) has an advantage over the individual feature space system (EER of 11.91%).

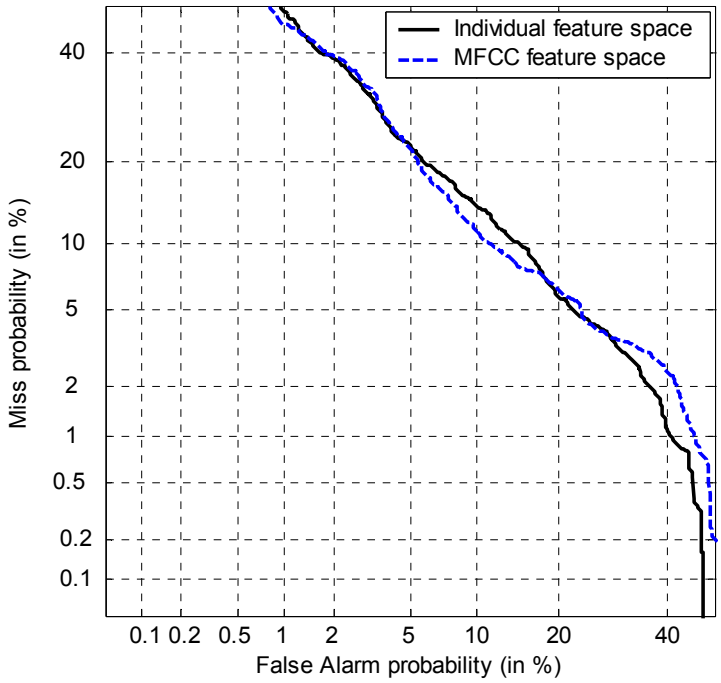

Figure 5.8: Average DET curves of speaker verification results in 5dB noisy database).

Figure 5.9 summarizes the EER values of the speaker verification systems (conventional MFCC space and the proposed system using individual feature space) vs. the SNR. The "clean" database case is indicated by 60dB.

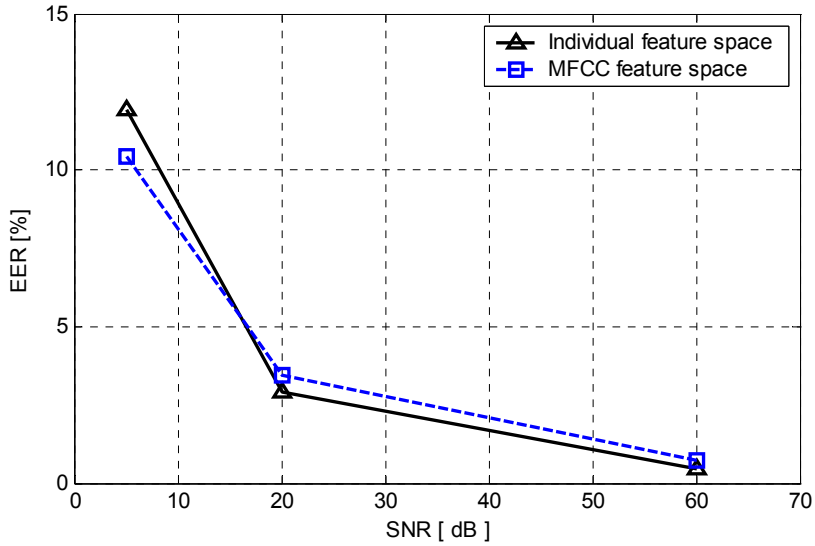

Figure 5.9: Average EER vs. SNR of the two verification systems.

5.2.2. Training on the Clean Database (60 dB)

In this experiment, the database is the same database as that used in the previous "clean" text-dependent experiment (section 5.1), except that to each **testing** speech file was added a white noise of two different SNRs: 20dB, and 5dB. The individual feature subsets are the same as in the "clean" database case, as well as are the background models and the target models.

Figure 5.10 shows the results of the verification experiments using the 20dB SNR testing database, obtained with two different 24-dimension feature spaces: 1) the individual feature spaces, and 2) the MFCC feature space. From this figure, one can see that the best results are with the MFCC feature space (EER of 5.8%). The individual feature space yields an EER of 11.37%.

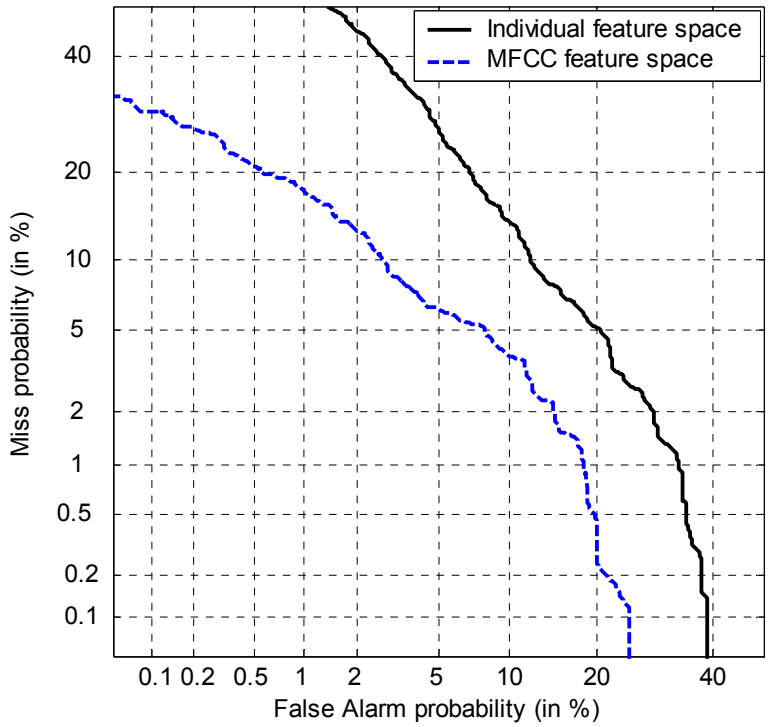

Figure 5.10: Average DET curves of speaker verification results in 20dB noisy testing database (using a clean training database).

Figure 5.11 shows the results of the verification experiments with the 5dB SNR testing database, obtained with different 24-dimension feature spaces (individual feature spaces and the MFCC feature space). From this figure, one can see that the best results are, again, obtained with the MFCC feature space (EER of 28.16%). The individual feature space yields an EER of 41.07%.

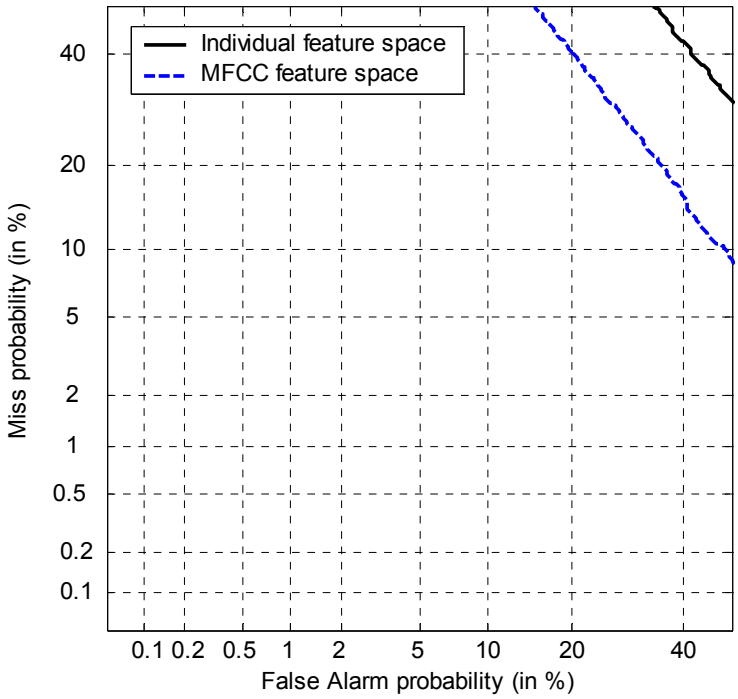

Figure 5.11: Average DET curves of speaker verification results in 5dB noisy testing database (using clean training database).

Figure 5.12 summarizes the EER values of these two speaker verification systems vs. SNR. The figure shows that the verification system using the individual feature space exhibits much worse results as the SNR decreases (using clean training database). As opposed to the performances using the noisy training and testing database (section 5.2.1) or the clean database (section 5.1), under the present conditions the MFCC feature space system gives better verification results. One can conclude from these results that the proposed

system using the individual feature space is much more sensitive to changing environmental conditions than the system using the MFCC feature space. For some applications, this could be a disadvantage. On the other hand, it is more robust to impostors using recording devices for purpose of fraud.

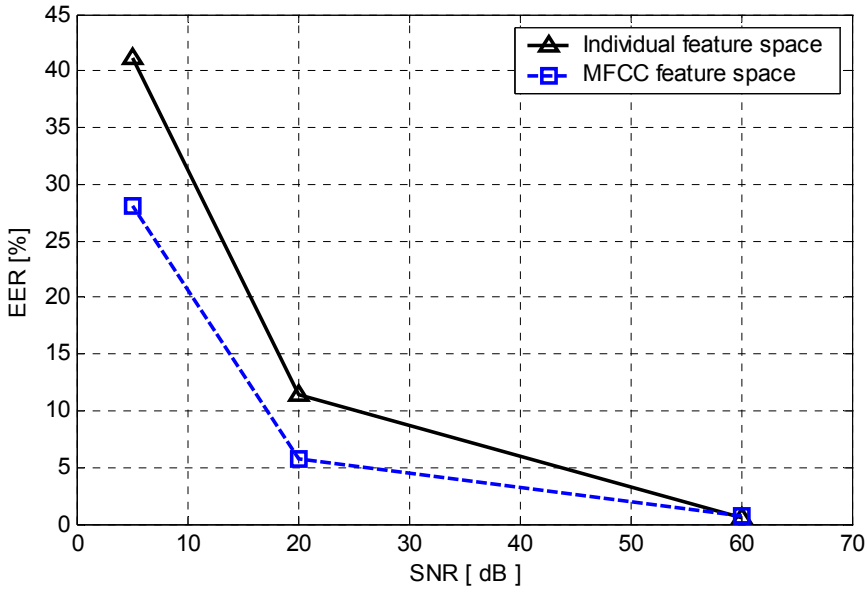

Figure 5.12: Average EER vs. SNR of the two verification systems, using a noisy testing database.

5.3. Text-Independent Speaker Verification

5.3.1. Experimental Setup

The experiment was conducted for a text-independent speaker-verification task. This experiment consists of only male speakers taken from the NIST99 and NIST98 databases. Front-end processing was employed, as discussed in section 4.1.1, and individual feature selection was employed, as discussed in section 4.1.2. The feature selection procedure was executed for each target with the *RRC* (3.31) criterion and with the forward selection procedure. The SFFS selection procedure was not employed due to computation speed considerations. Impostor selection for the individual feature selection process was not employed as well; instead a fixed group of impostors was used for all the targets' feature selection procedures. The selection procedure resulted in a set of $k = 24$ features for each target speaker. As in the test-dependent experiments, a feature order of 24 was determined in order to compare the results of the feature selection algorithm with the "almost standard" MFCC feature space (12 MFCCs + 12 ΔMFCCs). The model for each target was trained as a Gaussian Mixture Model (GMM), with 16 (diagonal covariance) Gaussians. Individual background models (GMM with 32 Gaussians) were trained using 30 speakers. These numbers of Gaussians were chosen using computation-time, memory, and overfitting considerations.

5.3.2. The Text-Independent Databases

As mentioned in the previous section, we used the NIST98 and NIST99 (1SPK) databases [Martin and Przybocki, 2000]. These databases are derived from the Switchboard-II corpus and consist of variable length utterances (0.5 – 60 seconds) extracted from one-sided

conversational telephone speech. For the experiments, we used only part of these databases. All the speakers (targets, background speakers, and impostors) were males, used electret handsets.

The target speakers were taken from the NIST99 database. For each target, the training data consisted of two minutes of total speech, one minute from each of two conversations. From this training database, one minute was taken to train the target model and one minute for the individual feature selection process. The file used for the feature selection process was segmented into 30 8-sec segments (after silence removal) with an 0.7-sec segment rate, to yield 30 target scores. 30 speakers from the NIST98 database were arbitrarily selected to be the background speakers; one 10-second speech file for each speaker. The impostor speech utterances for the feature selection procedure were taken also from the NIST98 database and consisted of 50 speech files, each one from a different speaker, arbitrarily selected (~10-second length).

For the testing database, 50 male impostors were taken from the NIST98 database, which are present neither in the background speaker database nor in the impostor feature selection database. 50 one-minute impostor files were arbitrarily chosen. The impostor trials were segmented from these files. Each impostor trial segment was 10-sec long (after silence removal) with 0.5-sec segment rate. The target trials (10-sec, 0.5-sec segment rate) were segmented from the other target files (NIST99) which were not used in the training process. The verification test includes 941 target trials (from ten target speakers) and 23670 impostor trials (males only).

5.3.3. Results and Discussion

Table 5.4 lists the 24 selected feature subsets for each of the first five target speakers, which were selected using the forward feature selection procedure along with the *RRC* criterion (3.31). From this table one can see that like for the text-dependent task, different feature spaces were selected for the different target speakers.

Table 5.4: Selected features for the (first 5) target speakers (text-independent).

Speaker #	Selected features
1	$m_7\ a_6\ a_{11}\ l_3\ l_8\ p_2\ p_4$ $\Delta m_3\ \Delta m_4\ \Delta m_8\ \Delta m_9\ \Delta m_{11}\ \Delta m_{12}$ $\Delta a_2\ \Delta a_3\ \Delta a_7\ \Delta a_{12}\ \Delta l_3\ \Delta l_8\ \Delta l_{12}\ \Delta p_2\ \Delta p_4\ \Delta p_7\ \Delta p_{12}$
2	$m_{10}\ a_{12}\ \Delta m_3\ \Delta m_5\ \Delta m_6\ \Delta m_7\ \Delta m_9\ \Delta m_{11}$ $\Delta a_2\ \Delta a_6\ \Delta a_7\ \Delta a_8\ \Delta a_9\ \Delta a_{10}\ \Delta a_{11}$ $\Delta l_7\ \Delta l_{11}\ \Delta p_2\ \Delta p_6\ \Delta p_7\ \Delta p_8\ \Delta p_9\ \Delta p_{10}\ \Delta p_{11}$
3	$m_4\ m_6\ m_7\ a_4\ a_5\ a_8\ l_5\ p_{10}$ $\Delta m_2\ \Delta m_3\ \Delta m_4\ \Delta m_8\ \Delta m_9$ $\Delta a_9\ \Delta a_{10}\ \Delta a_{12}\ \Delta l_3\ \Delta l_4\ \Delta l_7\ \Delta l_{12}\ \Delta p_5\ \Delta p_8\ \Delta p_9\ \Delta p_{10}$
4	$m_3\ m_5\ m_6\ m_7\ m_8\ m_9\ m_{10}\ m_{11}\ m_{12}$ $c_5\ c_6\ a_5\ a_8\ a_{11}\ l_8\ p_2\ p_5\ p_{11}$ $\Delta m_6\ \Delta m_{12}\ \Delta a_{10}\ \Delta a_{11}\ \Delta p_{10}\ \Delta p_{11}$
5	$m_{12}\ a_8\ a_{12}$ $\Delta m_1\ \Delta m_5\ \Delta m_6\ \Delta m_7\ \Delta m_8\ \Delta m_9\ \Delta m_{10}\ \Delta m_{11}\ \Delta m_{12}$ $\Delta c_1\ \Delta a_8\ \Delta a_9\ \Delta a_{10}\ \Delta a_{11}\ \Delta a_{12}\ \Delta l_{12}\ \Delta p_6\ \Delta p_8\ \Delta p_9\ \Delta p_{11}\ \Delta p_{12}$

Figure 5.13 is a histogram of feature occurrences (the top probable features, not all the 120 features) in the individual selected feature subsets. From this figure, one can see that most of the selected features are dynamic (delta) features. Moreover, most of them consist of higher order coefficients from MFCC ($\Delta m_3 \div \Delta m_{12}$), PARCOR ($\Delta p_7 \div \Delta p_{12}$), and LAR ($\Delta a_7 \div \Delta a_{12}$). As in the case of the text-dependent task, the MFCC features are the most prominent, however, here, the PARCOR and LAR are much more noticeable than in the text-dependent case.

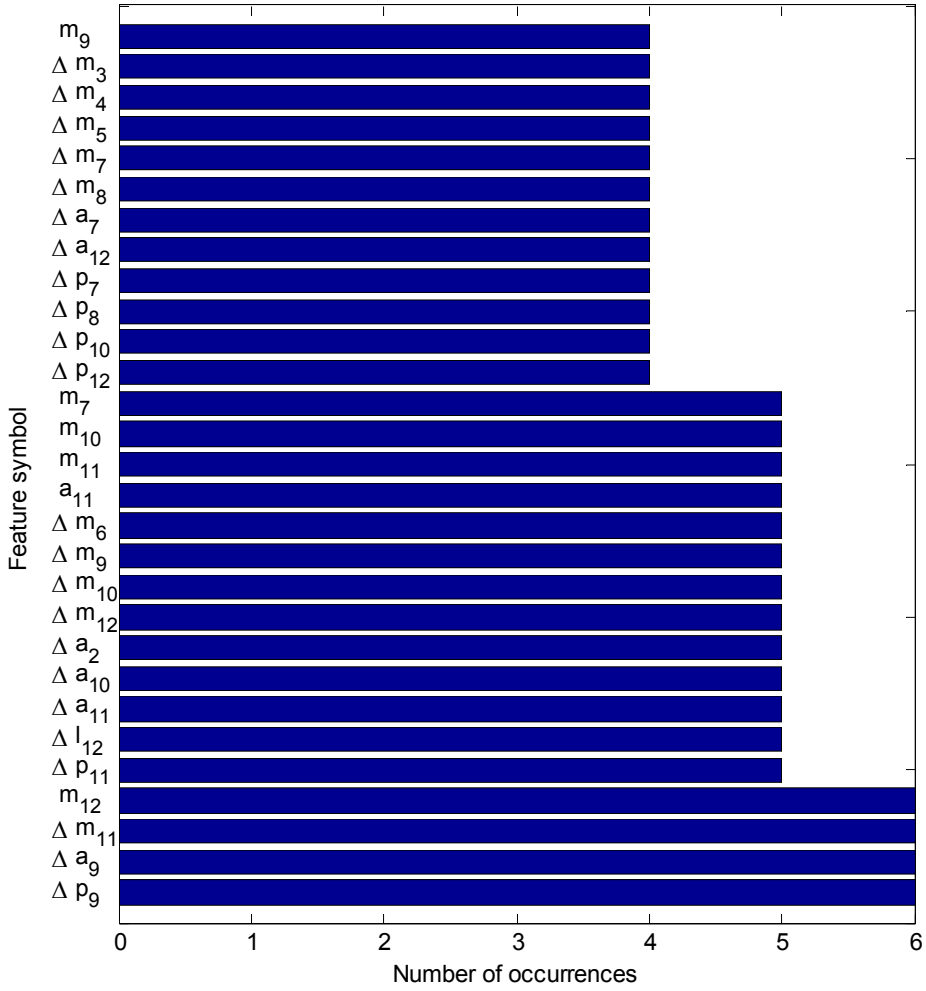

Figure 5.13: Number of feature occurrences in the individual selected feature subsets (ten targets).

Figure 5.14 shows the results of the verification experiments obtained with different feature spaces. The results are presented by average DET curves of the two different (24 dimensional) spaces: 1) the MFCC (12 MFCC + 12 ΔMFCC) feature space, and 2) the individual selected feature space. Each curve is an average of ten DET curves of the ten target speakers. The number of target trials is 941, and the number of impostor trials is 23670. From the figure we can see that the verification results in the individual feature space (EER of 4.15%) is much better than those in the MFCC feature space (EER of 6.14%).

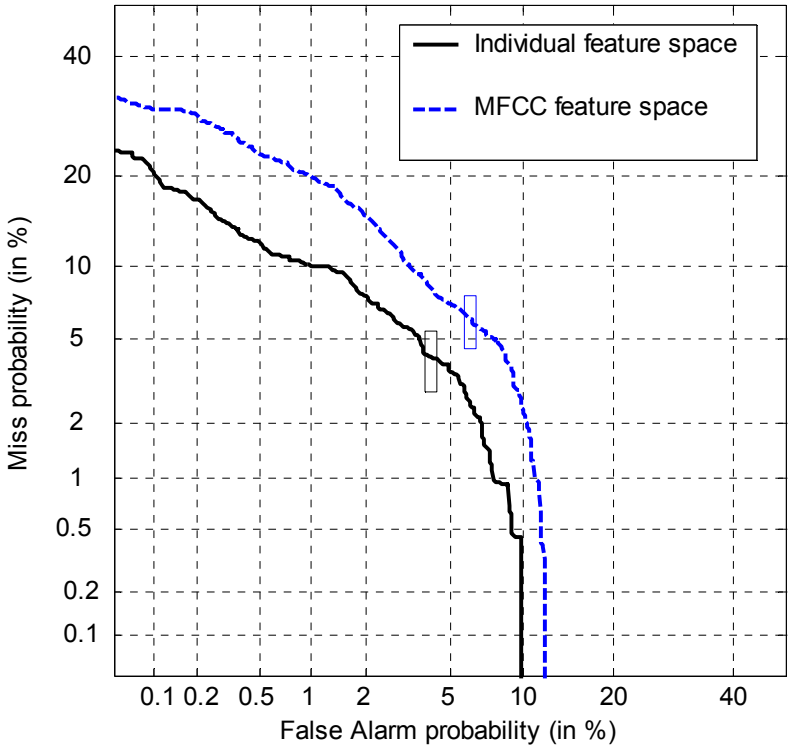

Figure 5.14: Average DET curves of text-independent speaker verification results using individual thresholds (feature spaces: 24 MFCC and Del MFCC space, and 24 individual optimal space)

As mentioned in section 4.2, because of the use of different feature spaces for each target, one cannot use a universal threshold; rather individual thresholds must be used.

5.3.3.1 Statistical Significance

The statistical significance [Mood et al., 1974] of the results in figure 5.14 can be visualized by shown by plotting a rectangle around the EER point of each curve which indicates a 95% confidence interval. This rectangle is computed under the assumption that each verification test is an independent trial and that misses and false alarms are decorrelated errors [Dunn et al., 2000]. The 95% confidence rectangle at the operating point (P_{miss}, P_{fa}) is bounded by the values

$$P_{\text{miss}} \pm z\sqrt{\frac{P_{\text{miss}}\left(1-P_{\text{miss}}\right)}{N_{\text{tgt}}}} \tag{5.1}$$

$$P_{\text{fa}} \pm z\sqrt{\frac{P_{\text{fa}}\left(1-P_{\text{fa}}\right)}{N_{\text{imp}}}} \tag{5.2}$$

where P_{miss} is the probability of miss, P_{fa} is the probability of false alarm, N_{tgt} is the number of target trials, and N_{imp} is the number of impostor trials. z is defined by

$$\Phi\left(z\right)=\frac{1+\gamma}{2} \tag{5.3}$$

where γ is the confidence coefficient, $\Phi\left(z\right)$ is the cumulative normal distribution, which is defined by

$$\Phi\left(z\right)=\int_{-\infty}^{z}\frac{1}{\sqrt{2\pi}}e^{-t^2/2}dt \tag{5.4}$$

In this statistical significance test, $\gamma=0.95\rightarrow\Phi\left(z\right)=0.975\rightarrow z=1.96$, $N_{\text{tgt}}=941$, and $N_{\text{imp}}=23670$. The non-overlapping error rectangles in figure 5.14 indicate that the performance improvement of the FS system is statistically significant.

6. Conclusions and Future Work

This work has proposed an individual feature selection algorithm for text-dependent and text-independent speaker verification systems based on HMM/GMM classifiers. The individual feature selection algorithm employed the proposed Recognition Related Criterion (RRC), which was proved in this work to correlate with the verification results.

It was shown that the use of an individual feature space can significantly improve speaker verification accuracy in text-dependent as well as in text-independent tasks. It was also shown that the individual feature space is more sensitive to SNR mismatch than the "conventional" MFCC space. Using an individual feature space requires the use of individual verification thresholds. Obviously, an individual feature selection process causes a greater computational load during the training process, and therefore consumes more time.

The proposed HMM-based verification system was evaluated on a local text-dependent database. A significant improvement over the "standard" MFCC space (12 MFCC + 12 ΔMFCC) in verification results was demonstrated with the selected individual feature space. An EER of 0.7% was achieved when the feature set was the MFCC space. Under the same conditions, the system based on the selected individual feature space (order of 24) yielded an EER of only 0.48%.

The proposed GMM-based verification system was evaluated on a text-independent database (NIST98 & NIST99). A significant improvement in verification results was demonstrated with the selected individual feature space. An EER of 6.14% was achieved when the feature set was the MFCC space. Under the same conditions, a system based on the selected feature space yielded an EER of only 4.15%.

It was found that most of the selected features in the text-dependent tests belong to the ΔMFCCs, especially the highest order coefficients $\Delta m_4 \div \Delta m_{12}$. In the text-independent tests, most of the selected features belong to the dynamic (delta, transitional) coefficients. This is somewhat similar to the finding of [Charlet and Jouvet, 1997] that: "much of speaker-dependent information is contained in transitional coefficients." It has also been demonstrated that the SFFS selection procedure is preferable to the other selection methods tested here.

Very few published papers deal with individual feature space (section 3.3). These papers were employed on relatively simple speaker recognition methods (quadratic classifier, VQ) and not with the state-of-the art HMM/GMM based speaker recognition systems. Moreover, they have tested relatively few features, in the task of speaker identification. The main innovation in this research was the combination of individual feature selection algorithm with the state-of-the art HMM/GMM based speaker verification systems (chapter 4). A novel criterion (RRC) was employed in the individual feature selection algorithm. This criterion was proved (chapter 3) to correlate with verification results.

The global feature set was chosen to contain K = 120 features from 10 groups of 12 order features. Among them: LPCs, MFCCs, PARCORs, LARs, LPCCs, and their derivatives (delta). These features were chosen due to computation considerations. In future work, we plan to include other features, such as PLPs [Hermansky et al., 1992] and prosodic features. Because of limitations in computation power and time, only ten target speakers were considered in the text-independent test. The results have shown to be statistical significant. We feel however, that tests with larger number of target speakers should be considered. Work is under way to employ more target speakers, as well as employing female speakers.

Work is under way to apply the algorithm to the problem of identification rather than verification. For the identification problem, we plan to use a common ‘optimal’ feature space rather than individual feature space.