ODYSSEY04 - The Speaker and Language Recognition Workshop
Toledo, Spain
May 31 - June 3, 2004

Text-Dependent Speaker Verification using Feature Selection with Recognition Related Criterion

Yaniv Zigel[(1)] *and Arnon Cohen*[(2)]

(1) NICE Systems Ltd., P.O.B 690 Ra'anana 43107, Israel,
(2) Electrical and Computer Engineering Department, Ben-Gurion University, Beer-Sheva, Israel.
yanivz@nice.com, arnon@ee.bgu.ac.il

Abstract

Speaker verification and identification systems most often employ HMMs and GMMs as recognition engines. This paper describes an algorithm for the optimal selection of the feature space, suitable for these engines. In verification systems, each speaker (target) is assigned an "individual" optimal feature space in which he/she is best discriminated against impostors. Several feature selection procedures were tested for the selection process. A Recognition Related Criterion (RRC), correlated with the recognition rate, was developed and evaluated.

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

1. Introduction

Today, Automatic Speaker Verification (ASV) systems [1 – 5] use a common feature space for all speakers. Moreover, this common set of features is most often the set of cepstral and delta-cepstral coefficients, used in speech recognition tasks. Very little work has been done on selecting the optimal feature space for speaker verification/identification tasks [6 – 14].

The motivation for the research is the assumption that every speaker has his own 'optimal' feature space, which optimally discriminates him against other speakers. This was supported by preliminary past work [6].

The goal of this paper is to demonstrate the significance of employing an individual feature space in modern Continuous Density Hidden Markov Model (CD-HMM) [15] or Gaussian Mixture Model (GMM) [16] based verification systems.

In this paper, a new criterion for feature selection was developed, which is suitable for speaker verification tasks. A text-dependent speaker verification system, based on CD-HMM was developed with individual feature selection procedure. The system was evaluated on a text-dependent database, using several feature selection procedures along with the new feature selection criterion, named "Recognition Related Criterion" (RRC).

Figure 1 shows a general scheme of the proposed speaker verification system.

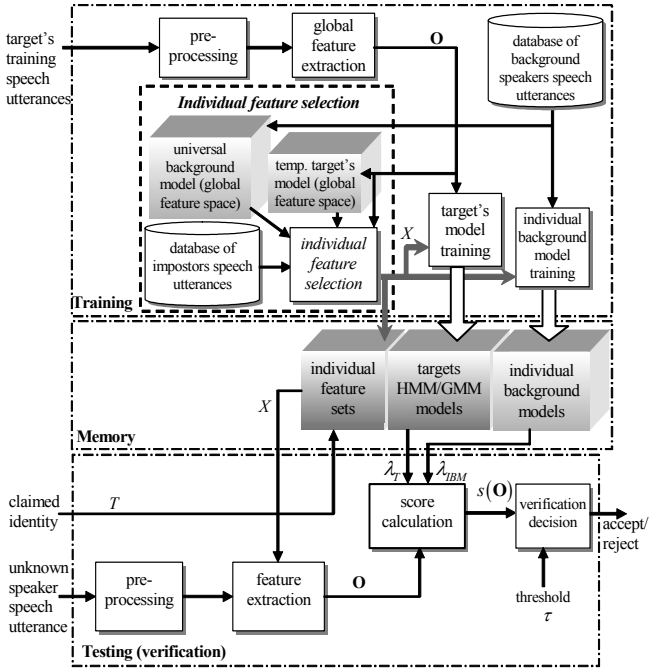

Figure 1: The proposed speaker verification system

The proposed speaker verification system consists (in the training stage) of the speaker's (target) training speech utterances, pre-processing and global feature extraction (set of all pre-determined features), The training of each speaker involves the extraction of a high dimensional feature space (termed here "global feature space"), from which the individual optimal feature sub-space will be extracted.

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

In the test stage, an unknown speaker's claimed identity T and test utterance are presented to the system. A verification algorithm, (to be discussed in section 3) employing the individual feature space, is used to provide an accept/reject decision.

2. Feature selection

Often, in pattern recognition problems, there is large number of features that may be used. Usually, it is not possible to work with very large dimensional feature space due to the fact that the recognition error may increase ("the curse of

dimensionality" [17]). Very often also the application forces constraints in memory and computation power, which limit the dimension of the feature space.

Feature selection is the process of selecting a features subset, which is most effective for preserving class separability. The feature selection method can be specified in terms of two components:

1) ***Performance criterion***, $J(\bullet)$

2) ***Selection procedure***,

The problem of feature selection can be described as follows:

Given a set Y of K features $Y=\{y_i \mid i=1,2,...,K\}$ select a subset X (of $k<K$ features) $X=\{x_i \mid i=1,2,...,k,\ x_i \in Y\}$ such that the performance criterion $J(\bullet)$ is optimized.

In speaker verification/identification tasks, the aim of this selection is to determine the feature space of size $k<K$ for which the recognition error is minimized. Minimizing the recognition error is not always easy to implement; hence separability measures are often introduced as criteria.

Several selection procedures are discussed in the pattern recognition literature [17, 18], such as:

Exhaustive search - an optimal method of feature selection. It considers all the combinations of *k* out of *K*. Implementation of such a search requires an enormous amount of computation, namely $\binom{K}{k}=\frac{K!}{k!(K-k)!}$ searches. For example, with *k* = 24 and *K* = 120, the number of searches is $\sim 10.872\times 10^{24}$ (!).

K-best Method - this method is probably the simplest one. The best subset of *k* features is composed of the *k* best features considered one at a time. However, a set of the best **individual** *k* features is not necessarily the best set of *k* features.

Forward Selection - This method sometimes called "*bottom-up*" [11], "*ascendant selection*" [8], or "*add-on*". The *Forward selection* procedure starts with the empty set and adds features iteratively. Initial tests are done with each of *K* features, one at a time, selecting the best single feature. Then, tests with two features, including the best one selected at the previous stage, and each (one at a time) of the remaining *K* – 1 features. The cycle is repeated until the desired number of features has been chosen.

Backward Selection - This method is a simple stepwise search technique, sometimes called "*knock-out*" strategy [14] or "*top-down*" [11]. The *Backward selection* procedure starts from the full set of *K* features. All *K* subsets of $K-1$ features are used in the performance criterion calculation to determine the best subset (of $K-1$ features). The feature not used in this best subset is "knock-out" of consideration. The process is repeated with $K-1$ subsets of $K-2$ features, etc.

The l-r Algorithm - The l-r algorithm [12] uses the forward and the backward selection in order to yield a better performance selection procedure. For every iteration, the algorithm uses the forward procedure in order to add l features, and uses the backward procedure in order to remove the r worst features from the augmented subset.

The Sequential Floating Forward Sequence (SFFS) - The Sequential Floating Forward Sequence (SFFS) [19] can be seen as a "dynamic" l-r algorithm. The SFFS procedure consists of applying, after each forward step, a number of backward steps as long as the resulting subsets are better than the previously evaluated ones at that level. Consequently, there are no backward steps at all if the performance cannot be improved.

Dynamic Programming (DP) - dynamic programming is utilized to find an optimal set of features with much fewer calculations than *exhaustive search*. The dynamic programming is a multistage optimization technique that makes use of the principle of optimality which states: whatever the initial state and decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision. When applied to the selection of features, the principle in conjunction with a functional equation permits the choice of attributes that have the maximum effectiveness [7]. One may view the dynamic programming procedure as a tree search method as shown in figure 2. In this representation, the features x_j ($j=1,2,...,K$) are depicted by the nodes of the tree. Subsets can be interpreted as paths or branches joining the nodes of subsequent stages. There are *k* stages in this iterative algorithm, as the number of features in the optimal subset. Let $\mathbf{q}_n^j=\left(q_1^j,q_2^j,...,q_n^j\right)$ ($j=1,2,...,K$) be one of the *K* possible subsets selected after *n* stages and q_n^j represents a feature in *X*. For every x_j ($j=1,2,...,K$) at the *n*th stage, the subset $\mathbf{q}_n^j$ is picked such that:

$$J\left(\mathbf{q}_n^j\right)=\max_i J\left(\mathbf{q}_{n-1}^i,x_j\right);\ i=1,2,...,K\ ;\ x_j\notin\mathbf{q}_{n-1}^i \qquad (1)$$

where *J* is defined as a feature performance criterion. For a detailed discussion of the DP algorithm see for example [7].

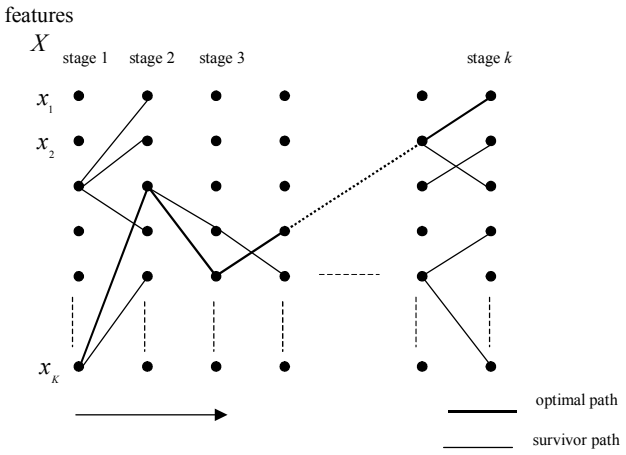

Figure 2: Feature subset selection using dynamic programming

To guarantee optimal results, the performance criterion, *J*, has to be monotonic, non-decreasing function of *n* and can be separated into two parts, one corresponding to the history of the process up to the *n*-1 stage and the other corresponding to the behavior of the process at the *n*th stage [20]. Most of the criteria used in practice, cannot guarantee these characteristics, especially when the features are dependent, so the DP becomes a sub-optimal selection method.

2.1. Performance criterion for Speaker Verification

In verification systems, the decision to accept or reject an identity claim is based on the comparison of a score with a threshold, τ. In this paper the score, $s(\mathbf{O})$, of utterance's observations, **O,** is the log likelihood ratio, $s(\mathbf{O}) = \log p(\mathbf{O} \mid \lambda_T) - \log p(\mathbf{O} \mid \lambda_{IBM})$, where λ_T is the target speaker's model, and λ_{IBM} is the individual background model.

A common and very often used evaluation measure for testing performances of speaker verification systems is the Equal Error Rate (EER), denoting the case when the false accept error is equal to the false reject (miss) error. It is therefore logical to use the EER (or some function thereof), as the performance criterion.

The use of the EER as a criterion becomes impractical due to the large computational load, since feature selection search algorithms require the estimation of the criterion at each step. Moreover, and more important, given a relatively small amount of training data available, the calculated EER yields very low resolution to be used as a criterion.

The proposed performance criterion, presented here, is the estimation of a function of the EER, based on the assumption that the scores' Probability Density Function (PDF) of the target $\left(f\left[s(\mathbf{O}) \mid \mathbf{O} \in \mathbf{O}_T\right]\right)$ and impostors $\left(f\left[s(\mathbf{O}) \mid \mathbf{O} \in \mathbf{O}_I\right]\right)$, may both be assumed Gaussians. Here $\mathbf{O}_T$ and $\mathbf{O}_I$ are the observations uttered by the target and impostors respectively. Figure 3 schematically describes the estimation of the EER.

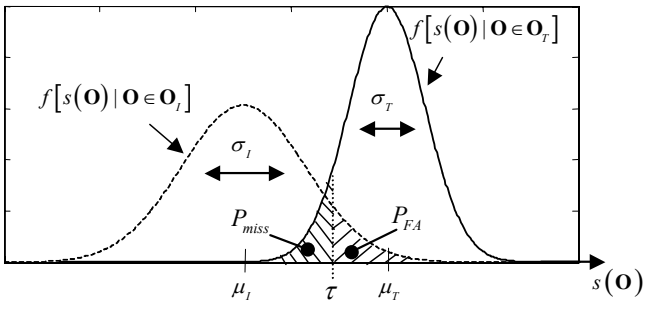

Figure 3: Estimation of verification errors from target and impostors Gaussian-like PDF's.

The PDF of the target's score is assumed to be Gaussian:

$$f\left[s(\mathbf{O}) \mid \mathbf{O} \in \mathbf{O}_T\right] = \frac{1}{\sqrt{2\pi}\sigma_T} \exp\left[-\frac{\left(s(\mathbf{O}) - \mu_T\right)^2}{2\sigma_T^2}\right]$$

and the PDF of the impostors is similarly assumed to be Gaussian with parameters (μ_I, σ_I). Given a threshold, τ, the False Accept (P_{FA}) and False Reject (or "miss" P_{miss}) errors may be calculated by the areas under the appropriate curves as shown in figure 3.

To check the Gaussian assumption, a χ^2 goodness-of-fit test was successfully performed on the targets' scores as well as on the impostors' scores, using 0.05 significance level and 9 degrees of freedom [24].

Figure 4 shows an example of a target's score histogram and its impostors' score histogram, with the best fitted Gaussians. The scores were calculated in the target's selected feature space (24 features).

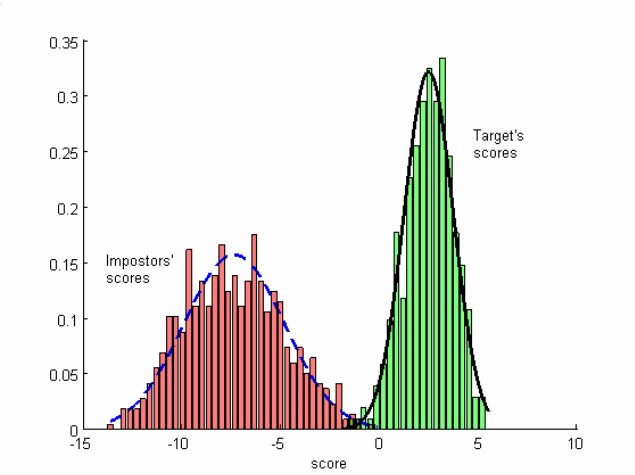

Figure 4: Gaussian fit for the histogram of target (#3) and impostors' scores.

2.1.1. The Recognition Related Criterion (RRC) [13]

Under the Gaussian assumption, the false reject, or P_{miss} errors and the false accept P_{FA} may be written:

$$P_{miss} = \int_{-\infty}^{\tau} f\left[s(\mathbf{O}) \mid \mathbf{O} \in \mathbf{O}_T\right] ds = \int_{-\infty}^{\tau} \frac{1}{\sqrt{2\pi}\sigma_T} \exp\left[-\tfrac{1}{2}\left(\frac{s-\mu_T}{\sigma_T}\right)^2\right] ds = \quad (2)$$

$$= \operatorname{erf}\left(\frac{\tau - \mu_T}{\sigma_T}\right) + \tfrac{1}{2}$$

$$P_{FA} = \int_{\tau}^{\infty} f\left[s(\mathbf{O}) \mid \mathbf{O} \in \mathbf{O}_I\right] ds =$$
$$= \int_{\tau}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_I} \exp\left[-\frac{1}{2}\left(\frac{s-\mu_I}{\sigma_I}\right)^2\right] ds = \quad (3)$$
$$= -\mathrm{erf}\left(\frac{\tau-\mu_I}{\sigma_I}\right) + \frac{1}{2}$$

where τ is the threshold for which $P_{miss} = P_{FA}$ $(= EER)$ (figure 3), and:

$$\mathrm{erf}(x) = \frac{1}{\sqrt{2\pi}} \int_0^x \exp\left[-\frac{1}{2}t^2\right] dt$$

$$\mu_T = E\left\{s(\mathbf{O}) \mid \mathbf{O} \in \mathbf{O}_T\right\},\ \mu_I = E\left\{s(\mathbf{O}) \mid \mathbf{O} \in \mathbf{O}_I\right\}$$
$$\sigma_T = \mathrm{std}\left\{s(\mathbf{O}) \mid \mathbf{O} \in \mathbf{O}_T\right\},\ \sigma_I = \mathrm{std}\left\{s(\mathbf{O}) \mid \mathbf{O} \in \mathbf{O}_I\right\}$$

The value of τ for which $P_{miss} = P_{FA}$ may be calculated by:

$$P_{miss} = P_{FA} \Rightarrow \mathrm{erf}\left(\frac{\tau-\mu_T}{\sigma_T}\right) + \frac{1}{2} = -\mathrm{erf}\left(\frac{\tau-\mu_I}{\sigma_I}\right) + \frac{1}{2}$$
$$\Rightarrow \mathrm{erf}\left(\frac{\tau-\mu_T}{\sigma_T}\right) = \mathrm{erf}\left(-\frac{\tau-\mu_I}{\sigma_I}\right)$$

Since $\mathrm{erf}(\bullet)$ is monotonically injected (one to one) function, the last equation yields

$$\frac{\tau-\mu_T}{\sigma_T} = -\frac{\tau-\mu_I}{\sigma_I}$$

hence, the value of τ for which $P_{miss} = P_{FA}$ is given by:

$$\tau = \frac{\mu_I\sigma_T + \mu_T\sigma_I}{\sigma_I + \sigma_T} \quad (4)$$

Introducing the value of τ (4) in the P_{miss} $(= EER)$ equation (2):

$$EER = \mathrm{erf}\left(\frac{\mu_I - \mu_T}{\sigma_I + \sigma_T}\right) + \frac{1}{2} \quad (5)$$

Since we are interested in minimizing EER, the constant ½ is irrelevant. Moreover, since erf(·) is a monotonically injected function, its argument may be used as a criterion. Thus, the proposed performance criterion, *RRC*, is

$$RRC = \frac{\mu_T - \mu_I}{\sigma_I + \sigma_T} \quad (6)$$

The criterion of equation (6) is to be maximized. Note that this criterion is somewhat similar to the F-ratio for the two Gaussian curves.

3. Experimental setup

The experiment was set 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 use of background models for score normalization instead of cohorts is because of computation speed considerations.

3.1. The database

The algorithm was evaluated with utterances of the Hebrew word /hamesh/ (five), taken from the Hebrew Isolated Digits (HID) database. The database contains high quality speech, recorded over a six months period; sampled at 16KHz with 12 bits resolution.

Ten male speakers from this database, which have the highest number of utterance repetitions, were chosen to be target speakers. For each target there are 39 impostors. The number of utterances (repetitions of the word 'five') for each target speaker is between 70 to 400, and the number of utterances for each impostor is 45. The first 20 utterances for each Target speaker are used for training, the next 20 (targets' utterances) for feature selection procedure (evaluation), and the rest utterances for testing.

3.2. Front-end processing and the global feature set

A conventional front end processing is employed in the system. First, the speech windowed by a 30 ms Hamming window with 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 are extracted from the speech frames. The global feature set was chosen to contain $K = 120$ features from 10 groups of 12 order features. Table 1 shows the overall set of features and their assigned symbols.

Table 1: The features and their symbols.

#	Feature name	Order	Symbols
1	Mel Frequency Cepstral Coef. (MFCC) [21]	12	$m_1 \div m_{12}$
2	Linear Prediction Cepstral Coef. (LPCC) [15, 22]	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 [21] were chosen since they are most often used in speaker verification / recognition systems. In our work, no Cepstral Mean Subtraction (CMS) was added. The other features [22] were chosen due to ease of estimation. Since the goal of the paper is proof of concept, only 120 features were used, to reduce the calculation time. In future work other features, such as for example PLP's will be included. The features have been normalized to their standard deviation in the feature extraction process to improve results.

3.3. The verification system and the feature selection procedure

The verification system is shown in figure 1 and described shortly in the introduction. The individual feature selection blocks are shown in the training stage of the system. The individual feature selection procedure requires a set of impostor utterances. Obviously, one cannot use all the impostor utterances in the database – a small set of impostors' utterances must be selected. These were chosen as follows:
For each target - (*T*), a CD-HMM was trained with 20 of the target's utterances, yielding a target model λ_T. These models were defined for the full (global) 120-feature space. For each one of the target models, six "selected" impostors (cohort speakers - *C*) were determined using the Close Impostors Clustering (CIC) method [23] with the divergence-like criterion:

$$d_D(C \mid T) = \frac{1}{N_{\mathbf{O}_T}} \sum_{i=1}^{N_{\mathbf{O}_T}} \left[\log p\left(\mathbf{O}_T^i \mid \lambda_T\right) - \log p\left(\mathbf{O}_T^i \mid \lambda_c\right) \right] - \frac{1}{N_{\mathbf{O}_c}} \sum_{j=1}^{N_{\mathbf{O}_c}} \left[\log p\left(\mathbf{O}_c^j \mid \lambda_T\right) - \log p\left(\mathbf{O}_c^j \mid \lambda_c\right) \right] \quad (7)$$

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's utterances and candidate impostor's utterances respectively. The cohorts selected in the 120-feature space were used for all sub spaces required by the feature selection algorithm.

The feature selection procedure was executed for each target with the *RRC* (6) criterion using the evaluation database: 20 target's utterances, and 10 utterances from each one of the six cohort impostors (C = 6). 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.

In the test stage, an unknown speaker's claimed identity and test utterance are presented to the system. From the identity claim, the appropriate feature space is drawn and feature extraction is made on the pre-processed utterance to yield 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, **O,** is the speech utterance's observations, λ_T is the target speaker's model, and λ_{IBM} is the individual background model. This model is trained for each target in its individual feature space, using the same background speakers.

4. Results and discussion

Figure 5 shows the maximum value of the criterion (*RRC*), 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. These curves indicate that the worst selection procedure is, as expected, the k-best. The next is the forward selection procedure. The two best selection procedures are the DP and the SFFS. The SFFS yields similar results, it is however more efficient than the DP in terms of calculations load. The SFFS was thus chosen as the selection procedure in our individual feature selection system.

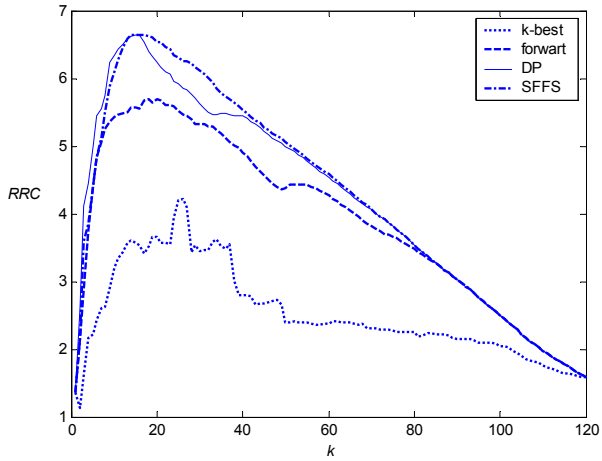

Figure 5: Maximum RRC criterion as a function of the feature space dimension, k, for several feature selection procedures (for speaker #3).

Figure 6 shows EER test results of the various selection methods as a function of the feature space dimension, for speaker #3. From this figure one can see that the dimension of k=33 yields best results (for the SFFS). For dimension sizes above 35, the EER increases probably due to the "curse of dimensionality". The results shown in figure 6 are posterior results, based on the test database. In practical cases, one would like to determine the order of the feature space, from the training/evaluation data, during the training process.

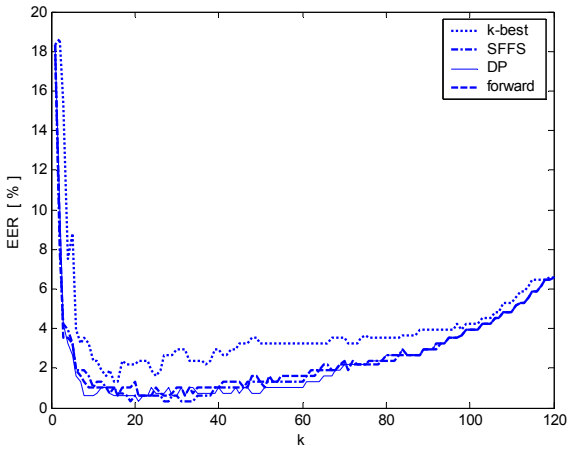

Figure 6: Real EER test results of the different feature selection procedures in different feature space dimension (for speaker #3).

Table 2 shows the 24 selected feature subsets for each one of the first five target speakers, using the SFFS feature selection procedure. The SFFS used the *RRC* criterion (6). From this table one can see that different feature spaces were selected for the different target speakers. One can see also that the dominant features in the optimal sets belong to the MFCC family.

Table 2: Selected features for the (first 5) target speakers.

Sp #	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 7 shows the histogram of feature appearance in the individual selected feature subsets (from the ten targets).

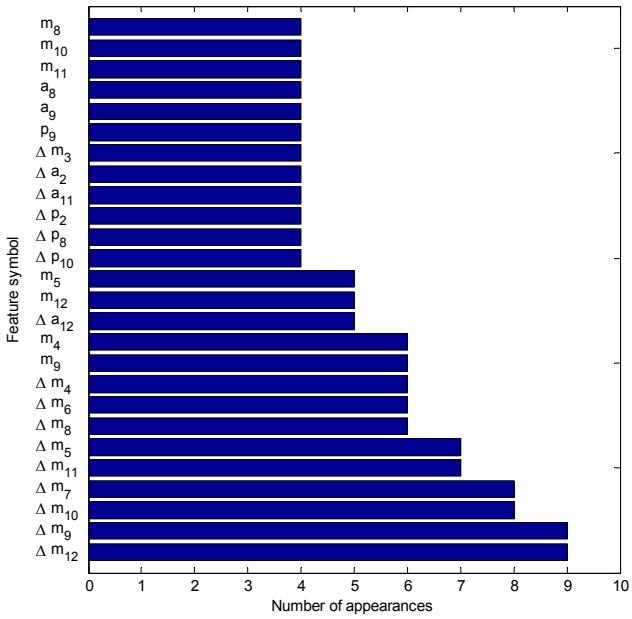

Figure 7: Number of feature appearance in the individual selected feature subsets (ten targets).

From figure 7 one can see that most of the selected features belong to the ΔMFCCs, especially the highest order coefficients $\Delta m_4 \div \Delta m_{12}$.

Figure 8 shows the results of verification experiments obtained with different feature spaces. Results are presented using Detection Error Tradeoff (DET) plots, which show the system tradeoff of misses versus false acceptances. The figure shows the average DET curves of the full set of 120 features and two different (24 dimensional) spaces: 1) the MFCC (12 MFCC + 12 ΔMFCC) feature space, 2) the individual selected feature space. Each curve is an average of ten DET curves of the ten target speakers. 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. The number of target trials is 1734, and the number of impostor trials is 6600.

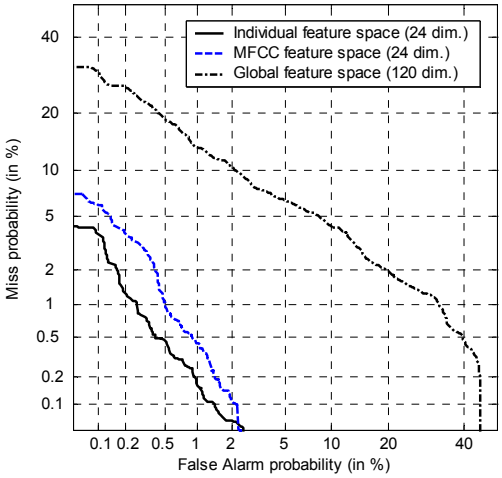

Figure 8: Average DET curves of speaker verification results (feature spaces: global 120 features, 24 MFCC and Del MFCC space, and 24 individual optimal space)

Figure 8 shows that the individual selected feature system yields the best results. The worst results were achieved with the overall 120-feature space, probably due to the "curse of dimensionality".

Table 3 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 3: Average 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. Conclusions

This work has proposed an individual feature selection algorithm for HMMs with a Recognition Related Criterion (RRC). It has shown that employing an individual feature space can significantly improve speaker verification accuracy. It has also been demonstrated that the SFFS selection procedure is preferable to the other selection methods tested here.

The results shown in this paper were for the case of text dependent speaker verification. Work is under way now to apply the feature selection algorithm to the case of text independent speaker verification [24]. For that the GMM is used rather than the HMM. Work is also under way to apply the algorithm to the problem of identification rather than verification. Here we plan to use a common 'optimal' feature space for all the speakers to be identified.

6. References

[1] J.P. Campbell, "Speaker Recognition: A Tutorial," *Proceedings IEEE*, Vol. 85, No. 9, pp. 1437-1462, Sept. 1997.

[2] G.R. Doddington, "Speaker Recognition - Identifying People by their Voices," *Proc. IEEE*, Vol. 73, No. 11, pp. 1651-1664, 1985.

[3] H. Gish and M. Schmidt, "Text-Independent Speaker Identification," *IEEE Signal Pro. Magazine,* pp. 18-32, Oct. 1994.

[4] S. Furui, "Recent Advances in Speaker Recognition," in: Audio- and Video-based Biometric Person Authentication: first international workshop; proc. / AVBPA'97, Switzerland, Berlin: Springer, pp. 237-252, 1997.

[5] R.J. Mammone, X. Zhang and R.P. Ramachandran, "Robust Speaker Recogntion," *IEEE Signal Processing Magazine*, Vol. 13, No. 5, pp. 58-71, Sept. 1996.

[6] A. Cohen and I. Froind, "On Text Independent Speaker Identification Using a Quadratic Classifier with Optimal Features," *Speech Communication,* Vol. 8, No. 1, pp.35-44, 1989.

[7] R.S. Cheung and B.A. Eisenstein, "Feature Selection via Dynamic Programming for Text-Independent Speaker Identification," *IEEE Trans. Acoust., Speech, Signal Processing,* Vol. ASSP-26, No. 5, pp. 397-403, October 1978.

[8] D. Charlet and D. Jouvet, "Optimizing Feature Set for Speaker Verification," in: Audio- and Video-based Biometric Person Authentication: first international workshop; proc. / AVBPA'97, Switzerland, Berlin: Springer, pp. 203-210, 1997.

[9] S. Furui, "Research on Individuality Features in Speech Waves and Automatic Speaker Recognition Techniques," *Speech Communication*, Vol. 5, No. 2, pp. 183-197, 1986.

[10] A. Haydar, M. Demirekler, and M.K. Yurtseven, "Speaker Identification Through Use of Features Selected Using Genetic Algorithm," *Electronic Letters,* Vol. 34, No. 1, pp. 39-40, Jan. 1998.

[11] A. Jain and D. Zongker, "Feature Selection: Evaluation, Application, and Small Sample Performance," *IEEE Trans. Pattern Analysis & Machine Intelligence,* Vol. 19, No. 2, pp. 153-156, 1997.

[12] M. Pandit and J. Kittler, "Feature Selection for a DTW-Based Speaker Verification System," *Proc. IEEE Int. Conf. Acoust. Speech Sig. Proc.,* pp. 769-772, 1998.

[13] A. Cohen and Y. Zigel, "On Feature Selection for Speaker Verification," *Proceedings of COST 275 workshop on The Advent of Biometrics on the Internet,* pp. 89-92, Nov. 2002.

[14] M.R. Sambur, "Selection of Acoustic Features for Speaker Identification," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-23, pp. 176-182, Apr. 1975.

[15] L.R. Rabiner, B.H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, Englewoof Cliffs, NJ, 1993.
[16] D.A. Reynolds, "Speaker Identification and Verification using Gaussian Mixture Speaker Models," *Speech Communication*, pp. 91-108, 1995.
[17] A.K. Jain, R.P.W. Duin, and J. Mao, "Statistical Pattern Recognition: A Review," *IEEE Trans. Pattern Analysis & Machine Intelligence*, Vol. 22, No. 1, pp. 4-37, 2000.
[18] K. Fukunaga, Introduction to Statistical Pattern Recognition, Academic Press, 1990.
[19] P. Pudil, F.J. Ferri, J. Novovicova, and J. Kittler, "Floating Search Methods for Feature Selection with Nonmonotonic Criterion Functions," *Proc. 12th ICPR,* Jerusalem, 1994.
[20] G.L. Nemhauser, *Introduction to Dynamic Programming*. New York: Wiley, 1966.
[21] S.B. Davis and P. Marmelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-28, No. 4, pp. 357-366, Aug. 1980.
[22] J.R. Deller, J.G. Proakis, and J.H.L. Hansen, *Discrete-Time Processing of Speech Signals*, MacMillan, New-York, NY, 1993.
[23] Y. Zigel and A. Cohen, "On Cohort Selection for Speaker Verification," *Proceedings of Eurospeech 2003,* Geneva, 2003.
[24] Y. Zigel, "Feature Selection for Speaker Recognition," Ph.D. Thesis, Ben-Gurion University, To be published, 2004.
[25] J.P. Campbell, "Features and Measures for Speaker Recognition," *Ph.D. Thesis, Oklahoma State University,* 1992.
[26] J.B. Attili, M.I. Savic, and J.P. Campbell, "A TMS32020-Based Real Time, Text-Independent, Automatic Speaker Verification System," *Proc. IEEE Int. Conf. Acoust. Speech Signal Processing,* pp. 599-602, 1988.