Frame Length Selection in Speaker Verification Task

DONATO IMPEDOVO, MARIO REFICE
Human-Machine Interaction System Lab.
DEE – Dept of Electrical and Electronic Engineering
Politecnico di Bari
Via Orabona 4, 70125 Bari
ITALY
impedovo@deemail.poliba.it, refice@poliba.it

Abstract: - In this paper an approach based on the use of different frame lengths for the feature extraction process in the training and recognition phases of a speaker verification system is presented. It is able to sensibly reduce the ER and the degradation on performance related to the training/verification feature mismatch. The potentiality of the approach are investigated in an a-posteriori search of the best combination to be adopted. A prototype of an expert system is also presented in order to automatically search, in real time operating conditions, the optimal combination of parameters. Tests have been performed on a set of speakers whose speech productions were spanned over approximately 3 months.

Key-Words: - Speaker Verification, Text Dependent, Mismatch, Frame Length, CD-HMM

1 Introduction

In everyday life, it is a common experience for people to be able to identify speakers by their voices, and even "distinguishing between voices they have heard only one or two times" [1].

In speech technology, many attempts have been made aiming at modelling such human ability for a number of applications, such as in security access control systems, or in specific investigation fields like computational forensics. Such a task is particularly challenging because, differently from fingerprints or DNA sequence, a person's voice can change strongly, depending on several factors like state of health, emotional state, familiarity with interlocutors [1], and also along the time.

In voice identification, several linguistic-phonetic parameters have been investigated and proposed as representatives of the individual vocal tract characteristics. Average fundamental frequency, formants frequency values at some specific points, intonation patterns are, among others, the most popular features for which a certain degree of reliability have been experimentally demonstrated.

In Forensic applications, for example, a common used approach is the computation of the so called "Likelihood Ratio" which proves to provide reliable results when two speech samples are to be compared in order to quantify the probability of belonging to the same speaker or to two different ones. The most common acoustic features used in these kind of applications are Vowel Formants, extracted in some specific context. Although Cepstral analysis has been found more powerful than formant analysis since it is more representative of the global characteristics of speech production, in forensic applications this approach has not been fully exploited yet.

One of the main reasons is that judges and lawyers, usually not familiar with speech analysis techniques, understand the connection between formants and physiology better than the relation between the latter and the abstract concept of Cepstrum. [2]. It is also obvious that any judgment in a court is normally based on a number of proves coming from different sources and speaker identification/verification is only one of these.

In a security access systems however, like those we are looking at in this work, there is no need to convince people and the only requirement is the capability of the system to provide a reliable decision with as much accuracy as possible.

In speech technology applications, the Mel Frequency Cepstrum Coefficients (MFCC) are widely recognised to be a good set of acoustic parameters as they encode vocal tract and some source information, even though a reduced set of phonetic features has also been demonstrated to be effective in text-independent speaker identification performance [3].

In this paper we report on the influence of the frame length on the computation of MFCC in a text-

dependent speaker recognition system. The approach uses different frame lengths for the speech signal processing in the training and in the verification phases. The aim is to identify optimal parameter combination in order to improve performances.

The approach here proposed shows its theoretical potentialities in the a-posteriori observation of performance.

The system here proposed in details is an additional component of an expert system we are developing aiming at identifying a person in a control security access application

2 Speaker Verification process

Speaker verification is defined as the process of deciding if a speaker is who she/he claims to be Text-dependent applications are the ones with the highest performances and can be applied successfully in real situations. In these kinds of systems, the speaker is recognized through an uttered phrase known by the system as for instance a password.

This procedure implies at least a double security level: the first consists in the secrecy of the chosen password while the second is represented by the vocal characteristics of the speaker.

From a general point of view, the process of speaker verification consists in a decision derived by a comparison between the features extracted by a recognition engine and those stored in a database as schematized in Fig. 1.

The state of the art for the recognition engine is based on statistical classifiers such as Hidden Markov Models (HMM) [4] or Gaussian Mixture Models (GMM) [5]. These systems work in two phases: enrolment and recognition. During the training (enrolment) phase, the system learns the generative models for each user, while during the recognition phase the unknown input is classified according to the known models, or possibly rejected.

The first macro-step both for the training and the recognition phases (Fig.1) is the Features Extraction from the speech input files. Features Extraction is the process through which a reduced set of parameters is derived from the input speech in order to characterize and represent it. The utterance recognition component uses a set of phonemes and sub-words speaker independent models in order to recognize the pronounced password ("name surname").

As already mentioned, spectral based features, like Mel Frequency Cepstral Coefficients (MFCCs) and their derivatives are widely used and accepted as the basic features able to give a suitable representation of the vocal tract [6, 7, 8].

The process for extracting MFCCs from the input speech emulates the way human ears capture and process sounds with different accuracy levels over different frequency bandwidths. Unfortunately, since the vocal tract characteristics tend to vary along the

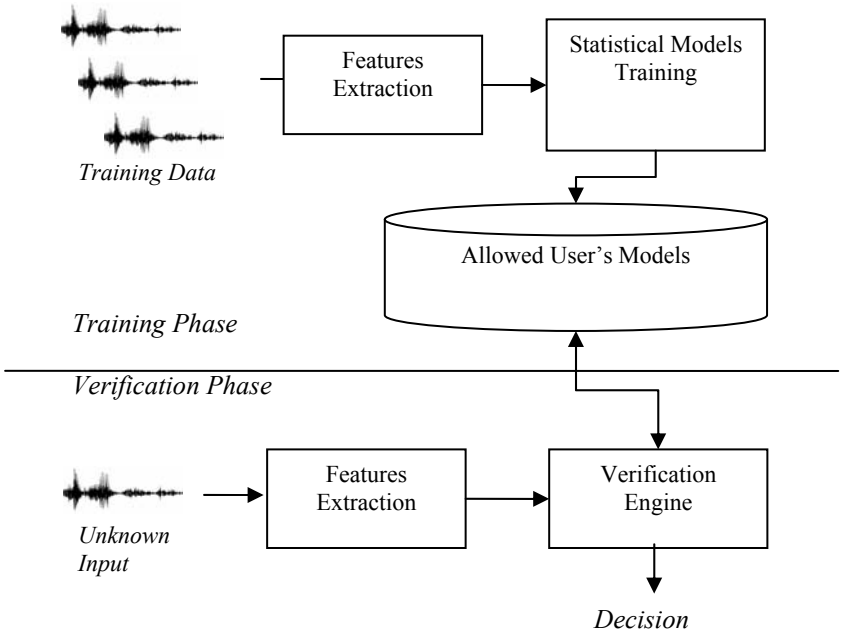

Fig. 1. Overall model of a speaker verification system

time, whereas the identification process is based on models trained on features extracted during the enrolment phase, the performance of the recognition process results in a significant degradation.

A reason of this degradation may be due to the so called "pitch-mismatch" [9]: the change of the pitch cycle along the time, even though both the speaker and the spoken utterances remain the same. In this paper we refer to a system we are developing, where the password used by the speaker for verification simply consists in his/her "name surname".

In order to cope with the above mentioned problem of performance degradation, we carried out a set of tests by varying the frame length of speech analysis window.

Tests have been carried out by observing verification performance on speakers who have been asked to perform different authentication session over a period of three months.

In order to cope with the huge amount of data, and with the needed flexibility of the system in dealing with the experimental adjustment of the varying analysis windows, a fast prototyping system has been used [10, 11].

The use of different frame lengths for the speech signal processing in the training and in the recognition phases can cope with the kind of degradation already mentioned since a specific frame length could be better than another in a particular case to better solve the natural pitch's cycle.

The aim is to identify optimal parameter combination in order to improve performance.

3 System Description

The system consists of two components, the first aims at checking the validity of the uttered password, and the second at verifying the genuine identity of the speaker that has pronounced the password.

3.1 The speech recognition engine

The utterance recognition component uses a set of phonemes and sub-words speaker independent models in order to recognize the pronounced password ("name surname").

The speech recognition system is based on a very simple grammar with few syntactic rules for the dynamic concatenation of the basic units. At the end of the process, the recognition hypothesis is parsed by using the knowledge of the known and allowed identities. The output is the transcription of the identity stored in the database or, if the utterance has been judged as not belonging to the defined sets, a rejection message is provided. The complete description of the utterance recognition system is beyond the aims of this paper and will not be illustrated in more details here.

3.2 The speaker verification engine

The speaker verification engine is here described according to its main components: the features extraction for the speech signal and the recognition engine.

3.2.1 Features Extraction

In this work, the Mel Frequency Cepstral Coefficients (MFCCs), their time derivatives and the respective energy parameter have been considered. MFCC are obtained from the power spectrum of the speech signal. Since speech is a non stationary signal, in order to perform the Discrete Fourier Transform (DFT) a short time analysis is performed. Figure 2 shows the framing process.

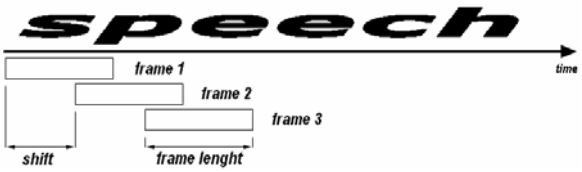

Fig. 2. The framing process

In speaker verification applications, this step is usually performed using 20÷30ms frames and a fixed shift of 10÷15ms, with the assumption that the signal is quasi-stationary within a frame interval.

For each frame the DFT is computed as follow:

$$X(k) = \sum_{n=0}^{N-1} w(n)x(n)\exp(-j2\pi kn / N)$$

for $k = 0,1,...,N-1$, where:

- $x(n)$ is the time discrete signal in the frame with length N,
- k corresponds to the frequency $f(k) = kf_s/N$,
- f_s is the sampling frequency in Hertz,
- $w(n)$ is the Hamming time window given by

$$w(n)=0.54-0.46\cos(\pi n/N).$$

The magnitude spectrum $|X(k)|$ is then scaled in frequency and magnitude. The frequency is scaled using the Mel filter bank $H(k,m)$ and then the logarithm is considered:

$$X'(m) = \ln\left(\sum_{k=0}^{N-1} |X(k)| \cdot H(k,m)\right)$$

for $m = 1, 2, \ldots, M$, with

- M is the number of filter banks,
- $M << N$.

The Mel filter bank is a collection of triangular filters defined by the center frequencies $f_c(m)$ and defined as follow:

$$H(k,m) = \begin{cases} 0 & for \quad f(k) < f_c(m-1) \\ \dfrac{f(k) - f_c(m-1)}{f_c(m) - f_c(m-1)} & for \quad f_c(m-1) \le f(k) \le f_c(m) \\ \dfrac{f(k) - f_c(m+1)}{f_c(m) - f_c(m+1)} & for \quad f_c(m) \le f(k) < f_c(m+1) \\ 0 & for \quad f(k) > f_c(m+1) \end{cases}$$

and the center frequencies of the filter banks are spaced logarithmically on the frequency axis. Finally, the MFCCs are obtained by computing the Discrete Cosine Transform (DCT) of $X'(m)$:

$$c(i) = \sum_{m=1}^{M} X'(m) \cos\left(i \frac{\pi}{M}\left(m - \frac{1}{2}\right)\right)$$

for $i = 1, 2, \ldots, M$, and $c(i)$ is the ith MFCC.

The Mel warping transforms the frequency scale to place less emphasis on high frequencies: it is based on the non linear human perception of the frequency sounds. For each frame, over the MFCC, the delta cepstrum coefficients (time derivates of the MFCC) and the respective power parameters have been also considered.

Unfortunately, the alignment between the position of the frame under analysis and the stationary part of the signal cannot be guaranteed in a uniform way. As a consequence, artefacts are introduced in the power spectrum and in the information related to the fundamental frequency (pitch) conveyed by the MFCC. It has been observed that for high pitched speakers and for those characterised by average pitch variations between enrolment and testing phases ("pitch mismatch"), the fine spectral structure related to the pitch causes degradation on speaker recognition performance [12, 13]. In the last years, different approaches have been proposed, which are largely based on pitch synchronous methods by windowing the Cepstrum coefficients after their evaluation [14], or on the attempt of calculating features less sensitive to pitch changes yet capable of retaining good discriminative properties [15], aligning each individual frame to its natural cycle [9, 16].

Other approaches have considered modified two dimensional root Cepstral analysis [17, 18], or on multi-scale fractal analysis [19].

In this work, an approach that uses different frame lengths is investigated. Recognition performances have been computed by using different frame lengths to extract features from the speech signal in the training and in the recognition phases.

The use of frames having different sizes between the two different phases has been already demonstrated to cope with the pitch mismatch, reducing false rejections for the subset of high pitched speakers [20]. The frame lengths here considered are 22, 25, 28 and 31ms: they equally divide the range of the most used lengths, while the fixed size of 10ms has been adopted for the shift in the framing process.

3.2.2 The verification step

The system is based on HMMs with continuous observation densities. In text independent applications, GMMs are generally adopted [5] even though recently discriminative approaches have been proposed [21]. A GMM can be considered as a special case of continuous observation densities HMM [4, 22], where the number of states is one.

An HMM can be characterized by a triple:

- the state transition probabilities matrix A,
- the observation densities matrix B,
- the initial state probabilities Π,

through the following notation:

$$\lambda = \{A, B, \Pi\} = \{a_{i,j}, b_i, \pi_i\}$$

with $i, j = 1, \ldots, N$, where:

- N is the total number of states in the model,
- $a_{i,j}$ is the transition probability from the state i to j.

Given an observation sequence (features vectors) $O=\{o_t\}$ with $t=1, \ldots, T$, the continuous observation

probability density for the state *j* is characterized as a mixture of Gaussian probabilities:

$$b_j(o_t) = \Pr(o_t \mid j) = \sum_{m=1}^{M} c_{jm} P(o_t; \mu_{jm}, R_{jm})$$

with

$$P(o_t; \mu_{jm}, R_{jm}) = (2\pi)^{-d/2} |R_{jm}|^{-1/2} \exp\left\{\frac{1}{2}(o_t - \mu_{jm})^T R_{jm}^{-1}(o_t - \mu_{jm})\right\}$$

where:

- M is the total number of the Gaussian components in the mixture,
- μ_{jm} is the d-dimentional mean vector of the mth component at state j
- R_{jm} is the d-dimentional covariance matrix of the mth component at state j
- c_{jm} are the mixture weights which satisfy the constraint $\sum_{m=1}^{M} c_{jm} = 1$.

The mentioned model parameters have been estimated by the Baum-Welch iterative methods (also known as the expectation-maximization EM algorithm), in order to maximize *Pr(O| λ)* [23, 24].

A continuous HMM is able to keep information and to model not only the sound, but even the articulation and the temporal sequencing of the speech.

In text independent speaker verification tasks, the sequencing of sound in the training data does not necessarily represent the sound sequences of the testing data, so the state transition probabilities have a little influence, while in text-dependent application they play a fundamental role.

The Hidden Markov Models considered in all our experiments adopt a left-to-right no skip topology: a transition may only occur from one state to the immediate following one or to the same state. Figure 3 shows a 4 states left to right no skip model.

The choice of this topology is justified by the fact that speech production is a sequential phenomenon: in each state a new symbol is emitted, the one given at time *t* is temporally antecedent to the one which will be emitted at the time *t+1*.

For each state, the Gaussian observation probability-density function (pdf) is used to statistically characterize the observed speech feature vectors.

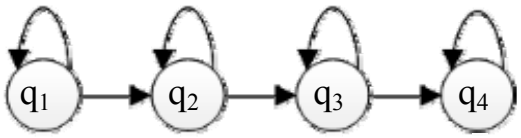

Fig. 3 : Four state left to right no skip model

For each speaker 4 different models were trained on the same speech data but adopting one of the four different parameters of 22, 25, 28 or 31ms for the framing process. In the verification phase they were used separately in order to better analyze the results. For each speaker an anti-model has been trained in order to represent impostors using the genuine author's password.

Give *O* the input file to be verified, λ_k the model for the *k*-th genuine speaker, and λ_k^I the impostor model for the *k*-th identity, the following value is computed:

$$S_k = \log\left(p(O \mid \lambda_k)\right) - \log\left(p\left(O \mid \lambda_k^I\right)\right)$$

where the logarithmic probabilities $\log\left(p(O \mid \lambda_k)\right)$ and $\log\left(p\left(O \mid \lambda_k^I\right)\right)$ have been evaluated by the Viterbi algorithm. In the decision process, if S_k is over a threshold and the output of the utterance recognition system corresponds to the *k*-th identity verified by the speaker verifier, then the output is accepted, otherwise it is rejected. This allows the rejection of all unknown (i.e. not included in the database) utterances from the inset and outset speakers.

4 Experimental Results

4.1 Experimental Setup

The experiments have been carried out on a corpus specifically collected which currently includes 34 speakers (16 female and 18 male) aged between 20 and 50 (we are still in the process of improving the size of this corpus). Each of them repeated his/her "name surname" in 4 or 5 recording sessions during a time span of 3 months. The time between sessions, and the number of recordings in each session is variable across speakers, making speech data much closer to real situations. In all the recordings sessions, a typical PC sound card has been used in an office environment with 22.050 Hz sampling rate, 16-bit quantization level and single channel.

Once the signal was framed, for each frame a filter bank of 24 Mel filters was used in order to extract 19 MFCC coefficients, their time derivatives (Δ-Cepstral, used to model trajectory information) and two energy values related to the MFCCs and to their derivatives [22, 25, 26, 27, 28]: this makes a total of 40 parameters per vector. In the training process, these parameters remain the same while the extracted features vectors differ by varying the frame length within the values 22, 25, 28 and 31 ms. No score normalisation was applied.

Table 1 summarizes the parameters adopted to extract features from the speech signal.

For each speaker in the corpus, four HMMs have been trained (one per frame length) using recordings of the first session (between 30 and 40 seconds of training data per speaker).

Parameters	MFCCs – Energy – ΔMFCC – Energy
Frame Lengths	22, 25, 28, 31 [ms]
Frame Shift	10 [ms]
No. of frame for Δ calculation	5
No. of vector element per frame	(19+1)*2=40
No. of MEL filters	24

Table 1 - Features parameterization details

Each model has 8 states and 3 Gaussian components in the mixture per state (this has been determined experimentally). The relative little amount of training data simulates real applications where an exhaustive training session cannot be performed and imposed on the users.

The system was tested on the remaining recordings of the following sessions even considering trials from impostors: the testing data are about a mean of 240 seconds of recordings for each speaker.

4.2 Performances

In a verification task, two typical errors can occur: the rejection of a genuine speaker (FR – False Rejection) since it has been considered as an impostor by the system, and the acceptance of an impostor (FA – False Acceptance) when it has been considered as a genuine speaker by the system. The Error Rate (ER) has been evaluated as the sum of these two terms.

Table 2 reports the results related to a classic approach where the performances of a baseline system are reported. In this case the frame length used for features extraction is the same value both for training and testing. In the training phase one genuine model for each speaker is trained: this is the way speaker verification systems usually work. As can be observed there is a significant gap in performance when comparing results obtained for the male set versus the female set. For the male subset the best classic approach is the one related to the use on 31ms: in this specific case the ER of 1.59 is observed. For the female subset the best classic approach is the one related to the use of 25ms: 2.57% in ER is observed.

System Configuration	Males ER %	Female ER %
22-classic	1.69	3.13
25-classic	1.62	2.57
28-classic	1.74	3.01
31-classic	1.59	3.35

Table 2 - Classic Approach Performances

Many tests have been performed by considering all the possible combination of HMM models (22, 25, 28 and 31ms) with one of the four different lengths to be used for the framing process in the verification phase. In the following, the most significant results are reported.

Table 3 reports the results obtained by using a frame length of 22ms in the framing process of the verification phase on the different models.

Model	Males ER %	Female ER %
22-ms	1.69	3.13
25-ms	1.46	2.56
28-ms	1.65	3.10

Table 3 - 22ms in the verification phase vs. 22, 25 and 28-ms models

The first row in Table 3 refers to the classic approach. The second row refers to the use of 22ms frame length in the verification phase with models trained on features obtained using 25ms: in this configuration an ER reduction of about 13% for the male subset and of about 18% for the female subset is observed. Improvements are also observed when considering 28ms models.

Table 4 shows the results obtained by using a frame length of 28ms in the framing process of the verification phase on the models trained using frame lengths of 25, 28 and 31ms.

Model	Males ER %	Female ER %
25-ms	1.61	2.53
28-ms	1.74	3.01
31-ms	0.88	3.60

Table 4 - 28ms in the verification phase vs. 25, 28 and 31-ms models

As can be observed, the use of 25ms frame length in the verification phase produces an ER reduction of about 16% for the female subset when compared with the classic approach (second row), while the use of 31ms produces about 50% of improvements for the male subset if compared with the classic approach. In this case, it must be observed that the use of the combined approach with 31ms models produces a degradation for the female subset.

In order to summarize, a comparison of the best combined and best classic approach having the lowest ER, leads to the following considerations:

- female subset: an ER reduction of about 2% has been observed with the combination 28V-25T (28ms for the Verification phase and 25ms for the Training one) if compared with classic 25V-25T;
- male subset: an ER reduction of about 40% has been observed when using the combination 28V-31T if compared with the classic 31V-31T.

4.3 Punctual Speaker Analysis

Results observed and discussed in the previous paragraph, refer to performance's mean values obtained over the whole female and male subsets. In order to explore the theoretical potentialities of the proposed approach a speaker by speaker test was successively performed. Test searched for the best a posteriori combination to be used for each speaker in order to minimize the ER. The final values obtained are the following:

- ER - Male subset = 0.66%
- ER – Female subset = 1.59%

The reduction for the male subset from the best classic approach is of about 58%, while for the female subset is of about 40%.

The frame length combinations able to perform for each speaker the lowest ER are reported in Fig. 4. The dotted squares refer to female speakers, while the vertical lined ones refer to male speakers. As can be observed there are many users that gain the minimum ER in all the different combination, bat there are also many others that gain the lowest ER just in one specific configuration.

5 Multiple Models

When a real time application is considered, the approach proposed could be applied performing an a posteriori investigation on the best frame length combination for each single speaker on the current trial in order to adopt it in the following one. On the other hand periodical investigation over specific and pre-determined testing data set could be performed: this second approach would have the advantage of a statistical value. Of course many other protocols could be introduced. Unfortunately there is evidence that the best frame length combination varies not only among speakers, but even among trials, so that the best combination determined in a specific trial could be different from the a-posteriori best one of the next trial.

Moreover these kind of protocols need a specific testing data set that should be different for each speaker, thus resulting in long, expansive and non fully automated process, especially when applications involving an high number of speakers are considered.

In order to solve this problem, and to exploit the potentialities of the approach, as already showed in paragraph 4.3, the needing is for a system able to check at each trial all the possible combinations and to evaluate which the best one could be. Of course the evaluation cannot have the certainty of the a posteriori one. One simple way to implement such kind of system is to identify the combination, among that considered, that generates the highest

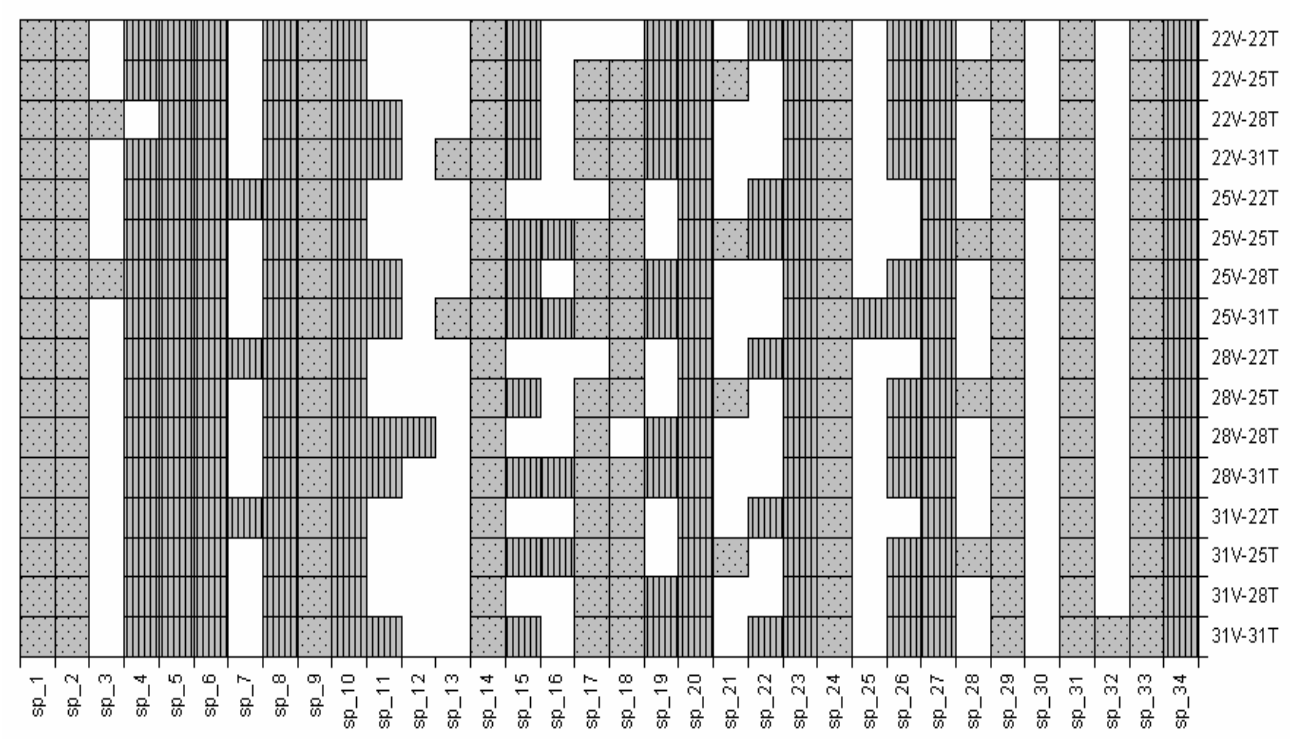

Fig. 4 – "Best" Frame length combination

probability in the verification process. In this formulation S_k (par. 3.2.2) depends by the specific *fl* (frame length) parameter related to the speaker model considered, moreover it would depend even by framing length parameter adopted in the framing process of the verification phase.

In order to investigate this kind of approach a preliminary test was performed. For each genuine speaker 4 different models were trained (22ms, 25ms, 28ms and 31ms) at the same time. In the verification phase the unknown input to be verified is framed using just one of the different frame lengths, successively the value S_k is computed for each model and the decision related to the highest value is given as final output decision.

Table 5 reports results related to the use of 25ms frame length in the verification process vs. 25ms models (first row) and vs. the multiple models (second row).

Model	Males ER %	Female ER %
25-ms	1.62	2.57
ALL	1.53	2,51

Table 5 – Classic approach vs. a ALL models

Improvements were observed both for the male subset (an ER reduction of about 5%) and for the female subsets (an ER reduction of about 2%), but results are still far from that obtained in the a-posteriori selection of the best combination.

6 Conclusion

This paper reports on the investigation of the influence of frame length for the computation of MFCC on the performance of a text-dependent speaker verification system.

For each speaker specific combinations of frame length to be adopted in the training and in the verification phase can be searched in order to minimize the ER. Combinations are speaker-dependent, and for each speaker the best combination could be found in a particular session and then applied in the following. Unfortunately a specific combination which is optimal for a session is not necessary the optimal even in the following one. Preliminary tests based on the use of multiple speaker models have showed the potentiality of performing an a priori selection of combinations, but performance are still far from the best (a-posteriori) already observed.

Future work will be focused on the mechanisms of variations of the optimal combination and on the

development of an intelligent system able to automatically detect it.

References:

[1] F. Nolan, "Dynamic Variability in Speech (DyViS). A forensic phonetic study on British English", http://www.ling.cam.ac.uk/dyvis/

[2] T. B. Alderman, "Forensic Speaker Identification", Lincom Europa, 2005, ISBN 3 89586 715 2.

[3] C.Y. Espy-Wilson, S.Manocha, S.Vishnubhotla, A new set of features for text-independent speaker identification, Proceedings of Proceedings of *International Conference on Spoken Language Processing,* ICSLP 2006, pp.1475-1478, 2006.

[4] L. R. Rabiner, B.H. Juang, "An Introduction to Hidden Markov Models", *IEEE Acoustics, Speech, and Signal Processing (ASSP) Magazine* 3(1), 1986, pp. 4-16.

[5] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[6] G. R. Doddington, "Speaker Recognition-Identifying People by their Voices", *Proceedings of IEEE*, Vol. 73, No. 11, November 1985, p.1651-1664.

[7] R. J. Mammone, X. Zhang, R. P. Ramachandran, "Robust Speaker Recognition, A Featurebased Approach", *IEEE Signal Processing Magazine*, 1996, pp. 58-71.

[8] S. Furui, *Digital Speech Processing, Synthesis, and Recognition*, New York and Basel: Marcel Dekker, 1989.

[9] R. D. Zilca, B. Kingsbury, G. N. Ramaswamy, "Pseudo Pitch Synchronous Analysis of Speech With Applications to Speaker Recognition", *IEEE Transactions on Audio, Speech, and Language Processing*, 2006, volume 14, no. 2.

[10] D. Impedovo, M. Refice, "Modular Engineering Prototyping Plan for Speech Recognition in a Visual Object Oriented Environment", in *WSEAS Transactions on Information Science and Applications*, Issue 12, vol. 2, 2005, pp. 2228-2234, ISSN 1790-0832.

[11] D. Impedovo, M. Refice, "A Fast Prototyping System for Speech Recognition based on a Visual Object Oriented Environment", Proceedings of *5th WSEAS International Conference on Signal Processing, Computational Geometry and Artificial Vision*, ISCGAV 2005.

[12] T. F. Quatieri, R. B. Dunn, D. A. Reynolds, "On the influence of Rate, Pitch, and Spectrum on Automatic Speaker Recognition Performance", Proceedings of *International Conference on Spoken Language Processing,* ICSLP 2000, 2000.

[13] S. Kim, T. Eriksson, H. G. Kang, D. H. Youn, "A pitch synchronous feature extraction method for speaker recognition", Proceedings of *International Conference on Acoustics, Speech, and Signal Processing* ICASSP 2004, 2004, pp. II-405 II-408.

[14] S. Sae-Tang, C. Tanprasert, "Feature Windowing-Based for Thai Text-Dependent Speaker Identification Using MLP with Backpropagation Algorithm", Proceedings of *International Symposium on Circuits and Systems,* ISCAS 2000.

[15] J. Liu, T. F. Zheng, W. Wu, "Pitch Mean Based Frequency Warping", Proceedings of *International Symposium on Chinese Spoken Language Processing*, ISCSLP 2006, 2006, pp. 87-94.

[16] R. D. Zilca, J. Navratil, G. N. Ramaswamy, "Depitch and the role of fundamental frequency in speaker recognition", Proceedings of *International Conference on Acoustics, Speech, and Signal Processing,* ICASSP 2003, pp. II-81 – II-84

[17] H. Marvi, "Speech Recognition Through Discriminative Feature Extraction", *WSEAS Transactions on Signal Processing*, Issue 10, Vol. 2, October 2006, pp. 1364-1370.

[18] E. Chilton and H. Marvi, "Two-dimensional root cepstrum as a feature extraction method for speech recognition", *Electronics Letters*, 39(10): 815-816, May 2003.

[19] F. Nelwamondo, U. Mahola, T. Marwola, "Multi-Scale Fractal Dimension for Speaker Identification Systems", *WSEAS Transactions on Systems*, Issue 5, Volume 5, May 2006, pp. 1152-1157.

[20] D. Impedovo, M. Refice, "The Influence of Frame Length on Speaker Identification Performance", Proceedings of *International Symposium* on Information *Assurance* and Security, IAS 2007, Manchester, 2007.

[21] V. Wan, S. Renals, *Speaker Verification Using Sequence Discriminant Support Vector Machines*, IEEE Transaction on Speech and Audio Processing, vol. 13, No. 2, March 2005.

[22] S.J. Young, "HTK, Hidden Markov model toolkit V1.4, Technical report", Cambridge University, Speech Group.

[23] L. Baum, T. Petrie, G. Soules, N. Weiss, *A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains*, Annals of the Institute of Statistical Mathematics, vol. 14, pp. 164-171, 1970.

[24] A.P. Dempster, N.M. Laird, D.B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm*, J. Royal Statistical Society, vol. 39, no. 1, pp. 1-38, 1977.

[25] L.R. Rabiner, R. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall Signal Processing Series, 1978. ISBN: 0132136031.

[26] T.Parsons, *Voice and Speech Processing*, McGraw-Hill, 1987.

[27] A.V. Hoppenheim, R.W. Schafer, "Homomorphic Analysis of Speech", *IEEE Transaction On Audio and Electroacustics*, Vol. AU-16, No. 2, pp. 221-226.

[28] J. Deller, J. Hansen, J. Proakis, *Discrete-Time Processing of Speech Signals*, IEEE Press Classic Reissue, 1999, ISBN: 0780353862.