# SPEAKER VERIFICATION USING THE DYNAMIC TIME WARPING

Svetlana SEGĂRCEANU[1], Tiberius ZAHARIA[2]

*Majoritatea schemelor comerciale de autentificare recente se bazează pe doi factori, de pildă o parolă şi un cod de securitate, sau codul PIN şi informaţii personale. Aceasta cu scopul de a spori nivelul de securitate în comparaţie cu cel asigurat de sistemele cu o parolă. Prezentăm în continuare modul în care metoda alinierii dinamice în timp poate fi adaptată pentru a putea fi aplicată corespunzător la verificarea vorbitorului folosind doi factori. În etapa de extragere a trăsăturilor caracteristice am utilizat două abordări perceptuale: PLP versus analiza în scara Mel, cu un al doilea scop, acela de a le evalua performanţele. Participanţii în experiment au rostit anumite texte obligatorii şi numele lor. Au fost testate mai multe combinaţii de vocabular*

*Most of the nowadays commercial authentication schemes are based on two factors: for instance a password and a security ID, or the PIN code and personal data. This is meant to improve the security in comparison with the systems based on one only factor. We present a way to adapt the Dynamic Time Warping approach in order to apply it suitably to a two factors scheme. For feature extraction in speaker verification experiments we used two perceptual approaches: the PLP versus the Mel-scale methodology, with a second purpose of assessing their performance. The participants in the experiment uttered some compulsory sentences in Romanian or their names. Several combinations of vocabulary were tested.*

**Keywords:** speaker verification, dynamic time warping, threshold setting, weighting, perceptual analysis of speech, biometric measures

## 1. Introduction

The speaker verification issue is to decide on the invoked identity of a client. Two decisions are possible: client and impostor. As any pattern recognition problem, it involves two aspects: training and testing (the verification itself). In the training phase the user must pronounce a number of utterances in order to create her or his model. In the verification process the user's processed signal output is compared to the model of the invoked speaker S. Furui ([1]) proposed

[1] PhD student, Depart. of Electronic and Telecommunication Engineering, University POLITEHNICA of Bucharest, Romania, e-mail: svet_segarceanu@yahoo.com
[2] PhD student, Depart. of Electronic and Telecommunication Engineering, University POLITEHNICA of Bucharest, Romania, e-mail: tezeu2000@gmail.com

that the utterance should also be compared to a model of the impostor, obtained by training with several "impostor" users.

Most of the latest commercial security schemes use the authentication based on two factors such as, for example, the password and the security ID, or the PIN code and the security ID, or a password and personal information. Although this would improve the security as compared to the security provided by only one factor, it does not guarantee that the pretended identity is the true one, as the PIN codes, personal data, the security ID, can all be obtained by the fake ([2]). The biometric technologies, unlike all other authentication methods, should prove that the users are what they pretend they are.

We used a variant of the Dynamic Time Warping approach, which, besides evaluating the alignment between two time sequences, produces new reference templates by applying a sort of averaging of the warped sequences. We show how we established the specific individual thresholds used in the verification phase. Finally we show how to improve the DTW performance by applying a weighted variant. As the speech material was limited we used in the training phase only utterances obtained in the first two recording sessions, and tried to squeeze as much information as possible out of them. In our research we show the advantage of a two factor scheme over a one factor scheme and try to evaluate various alternatives to the perceptual analysis: different perceptual scales, different rules. We used some highly corrupted records of 21 Romanian speakers. The speech signal was sampled at 11.125 kHz, each sample represented on 8 bits. The length of the speech frame was set to 22ms and the frame rate to half the length of the frame. A Hamming window, and a pre-emphasis filter ($\mu$=0.95) were applied. The speech database contains compulsory text, used in training and verification phases. The users also uttered arbitrary text, but we did not use this material in the verification trials. We evaluated the power of the perceptual analysis by means of the within-class and between-class scatter matrices of the perceptual features.

## 2. Speaker Verification

By speaker verification the pretended identity of a speaker is either accepted or rejected. This involves comparing the set of extracted features at the test trial, to the model of the assumed speaker. The score acquired at testing is compared to a threshold, specific for each individual, and if it is under that threshold the speaker is accepted. In classical approaches, the threshold is established based on the inter/intra-speaker scores. For instance the score could be established as the average score obtained at training.

The evaluation of a speaker verification system assumes the assessment of two aspects: the accuracy of the system and the required resources. The accuracy is expressed with the help of the following indicators:

FAR (False Acceptation Rate) – the probability of the false acceptance:

$$FAR = \frac{\text{number of false rejections}}{\text{number of impostors}} \tag{1}$$

FRR (False Rejection Rate) – the probability of the false rejection:

$$FRR = \frac{\text{number of false rejections}}{\text{number of speakers invoking their identity}} \tag{2}$$

These two indicators are used to set the thresholds. EER (Equal Error Rate) is attained for a threshold value FAR and FRR are approximately equal.

## 3. The Method

The Dynamic Time Warping algorithm measures the similarity, or distortion, between two sequences, which may vary in time. DTW is thus a method that allows a computer to find an optimal match between two given time sequences, applying certain restrictions, and furnishes a measure of their similarity, and an optimal path. The minimum distortion is always based on the previous step and should satisfy the following condition:

$$DTW(i, j) = \min_{k}(DTW(i-1, k) + d(k, j) \tag{3}$$

where d is a metric defined on the space $\Phi$, where the time sequences take values:

$$d : \Phi \times \Phi \to R \geq 0 \tag{4}$$

The general form of this algorithm involves the construction of a cumulated cost matrix *dtw*, representing the cost of the time warping of two time sequences s and *t*, of lengths *m* respectively *n*, based on the local cost matrix([3]):

$$C \in R^m \times R^n : c(i, j) = \left\| d(s_i, t_j) \right\| \tag{5}$$

Given two time sequences, *s* and *t*, with $m = |s|$, $n = |t|$, the algorithm *dtw(s,t,m,n)* involves the next steps:

Initialization:

$dtw(0,0) = d(s[0], t[0])$

for $i = 1$ to $m$ $\quad dtw(i,0) = dtw(i-1,0) + c(i,0)$ $\quad$ end for

for j = 1 to $n$ $\quad dtw(0, j) = c(0, j) + dtw(i, j-1)$ $\quad$ end for

The iterative process:

for $i = 1$ to $m$

for j = 1 to $n$

$dtw(i, j) = c(i, j) + \min(dtw(i-1, j), dtw(i, j-1), dtw(i-1, j-1))$

end for

end for

return dtw(m - 1, n - 1)

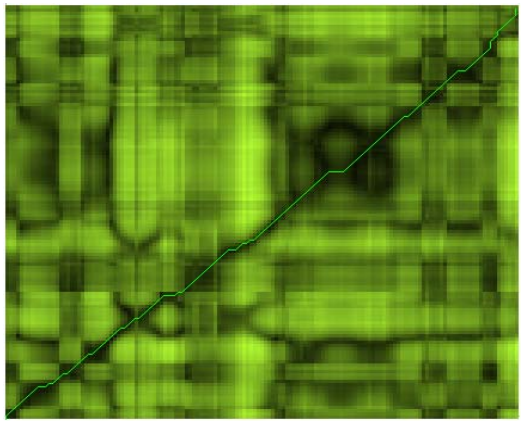

Fig. 1. The local cost matrix and the optimal path obtained from two sequences of PLP coefficients

Backtracking: Once generated the accumulated cost matrix, the optimal path warped by the DT Fig. 1 presents the local cost matrix and the optimal paths obtained by applying the DTW algorithm for two sequences of PLP coefficients. The algorithm meant to generate the optimal path, *optimalPath(s,t,m,n)*,is based on the accumulated cost matrix *dtw(i,j)* of two time sequences, *s* and *t*, with $m = |s|, \quad n = |t|$ :

```
path = new Vector()
i = m - 1; j = n - 1;
while (i > 0) & (j > 0) do
 if i == 0 then    j = j - 1;
    else if j == 0 then    i = i - 1
       else if dtw (i - 1, j) == min (dtw (i - 1, j); dtw (i, j - 1); dtw (i - 1, j - 1))
                  then      i = i - 1;
            else if dtw (i, j - 1) == min (dtw (i - 1, j); dtw (i, j - 1); dtw (i - 1, j - 1))
                     then    j = j - 1;
                else      i = i - 1; j = j - 1;
end if
 path.add((  i; j))
end if
end while
```

### 3.1 DTW in Speaker Recognition

Applying the DTW technique in speech recognition involves reporting the test utterance to one or more reference templates, usually representing one or more pronunciations of the invoked speaker's model. The feature sequence obtained from the test utterance is dynamically warped with the sequences of parameters contained in the reference templates. If we use two reference templates, the DTW algorithm generates two distortions *dtw1* and *dtw2*. These are compared to a threshold established in the training stage. If in the training phase, for a certain speaker, we use two templates, we may set one speaker's threshold to the value obtained from warping these two sequences, $dtw = dtw(coeff_1, coeff_2)$, where $coeff_1$, $coeff_2$ are the two reference feature sequences.

An alternative for the choice of the reference templates would be the generation at training, of a reference sequence in which each sample would be the average of the samples that are "aligned" by applying the DTW algorithm to two feature sequences obtained from two training utterances. By this we are able to use one reference model instead of two, thus sparing some memory resources. Fig. 2 presents the sequence obtained by warping two first order PLP coefficients sequences, the outer in red and green. The sequence generated by the algorithm is the middle yellow one.

We present subsequently the formalized algorithm ***dtwa(coeff$_1$, coeff$_2$,n$_1$,n$_2$),*** which uses as input parameters two ***dim***-dimensional sequences, $coeff_1$ and $coeff_2$, with $n_1 = |coeff_1|$, $n_2 = |coeff_2|$. In the first steps, the algorithm performs the calculation of the DTW distortion and the optimal path of the two sequences.

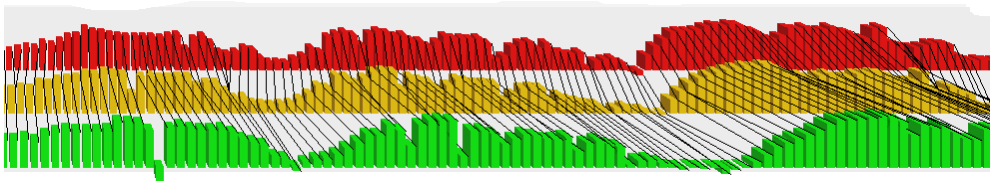

Fig. 2. The sequence obtained by the alignment of two sequences of PLP coefficients. Each sample in the middle sequence is the average of matching samples warped by the DTW algorithm

```
dtw(coeff1, coeff2, n1, n2);
 optPath = optimalPath(coeff1, coeff2, n1, n2);
 coeffTemp = new double [max(n1, n2)][dim];
    for i = 0 to dim coeffTemp[0][i] = 0 end for
    iteratii = 0; l = 0;
 pathOpt = optPath(0); temp1 = pathOpt[0]; temp2 = pathOpt[1];
u = 1;
while u < |optPath| & pathOpt[0] >= 0 & pathOpt[1] >= 0 do
  while pathOpt[0] = temp1 || pathOpt[1] = temp2 do
     if pathOpt[0] = temp1 then
        for i = 0 to dim coeffTemp[l][i] += coeff2[pathOpt[1]][i]; end for
        iteratii++;
     end if
     if pathOpt[1] = temp2 then
       for i = 0 to dim coeffTemp[l][i] += coeff1[pathOpt[0]][i]; end for
       iteratii++;
     end if
    pathOpt = optPath(u); u++;
  end while
    for s = 0 to nbDim coeffTemp[l][s] = coeffTemp[l][s] / iteratii; end for
 u--; l++;
 if pathOpt[0] >= 0 & pathOpt[1] >= 0 then
    for s = 0 to nbDim coeffTemp[l][s] = 0; end for
    temp1 = pathOpt[0]; temp2 = pathOpt[1]; iteratii = 0;
 end if
end while
n = l; coeff1 = new float[max(n1, n2)][dim];
for i = 0 to n
    for u = 0 to dim coeff[i][u] = coeffTemp[l-i-1][u]; end for
 end for
```

The procedure outputs, besides the newly sequence, a distortion measure

$$dtwa = (DTW(coeff_1, coeff, n_1, n) + DTW(coeff_2, coeff, n_2, n))/2 \quad (6)$$

In the above algorithm given a pair of aligned features in the optimal path sequence, ($coeff_1(u)$ , $coeff_2(v)$), the two sequences of feature vectors are averaged as follows:

1..if $coeff_1(u)$ (respectively $coeff_2(v)$ ) occurs in more than one optimally aligned pairs, all these pairs generate one point equal to the average of all the components of these pairs, and $coeff_1(u)$ respectively $coeff_2(v\ ())$ is counted once.

2. in particular, if $coeff_1(u)$ and $coeff_2(v)$ occur only once, this pair generates one vector equal to their mean.

### 3.2 Weighting the DTW Distortion

Formally speaking, the alignment path built by DTW is a sequence of points, which must satisfy the following constraints:

1. Boundary condition: the starting and ending points of the warping path must be the first and the last points of the aligned sequences ( *pathOpt[1] = (1; 1)* and *pathOpt[K] = (n1;n2),* where *K* is the length of the warping path).

2. Monotonicity condition, which preserves the time ordering of points.

3. Step size condition: this criterion restricts the warping path from long jumps (shifts in time) while aligning sequences. With the above notations

*pathOpt [t+1] - pathOpt[t+1]* $\in \{(1; 1); (1; 0); (0; 1)\}$.

In [8] Sakoe and Chiba introduce one more constraint:

4. The slope constraint condition based on the fact that too steep or too mild a gradient may reflect unrealistic correspondence between the aligned sequences. For instance a short pattern A, warped with a relatively long pattern B may generate a steep gradient. They account for the slope of the warping path measured as the local ratio of the numbers of leaps in the two directions of each of the warped sequences. They propose adding weights to each of the DTW distances, to penalize or favor certain types of point-to point correspondence. Thus the DTW distortion between two sequences turns into:

$$\mathrm{DTW_w}(\mathrm{s_1},\mathrm{s_2}) = \sum_{k=1}^{K} w_k d(pathOpt(k)[0], pathOpt(k)[1]) \tag{7}$$

where, if we denote *s(k)[i] = pathOpt(k)[i], i=1,2*:

$$w_k = (s(k)[0] - s(k-1)[0] + s(k)[1] - s(k-1)[1])/(n_1 + n_2) \tag{7a}$$

## 4. Perceptual Analysis

### 4.1 The Bark scale

The use of filter banks in speech processing for speech or speaker recognition systems is meant to model the human auditory apparatus, which behaves as if composed of a series of superposed filters. The pass band of each filter is called the critical band. Two pure tones lay in the same critical band if their frequencies are close enough to meet a certain degree of superposition of their amplitude envelopes in the basilar membrane. The bark scale, so called by the name of Heinrich Barkhausen, was among the first attempts to describe the effects of the critical bands. The nonlinear spacing between the critical bands matches a psycho acoustical scale proposed by Zwicker in 1961 ([4], [5], [6]):

$$f_{bark}(f) = 13 * \arctan(0.00076 \cdot f) + 3.5 \cdot \arctan\left(f/7500\right) \tag{8}$$

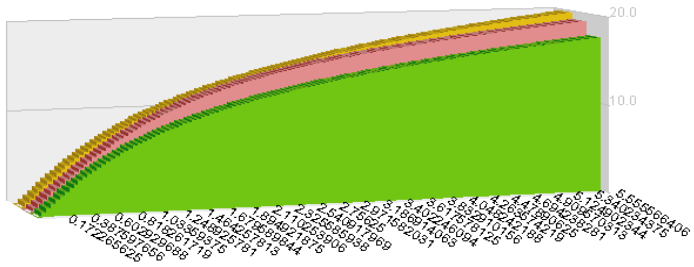

Fig. 3. Graphical representations of three expressions of the Bark scale of (8),(8a),(8b).

Alternative expressions for the bark scale are (see [5], [6]):

$$f_{bark}(f) = 6\ln\left[ f/600 + \left(f^2/600^2 + 1\right)^{0.5} \right] \tag{8a}$$

$$f_{bark}(f) = 7\ln\left[ f/650 + \left(f^2/650^2 + 1\right)^{0.5} \right] \tag{8b}$$

In the above relations f expressed in Hz. Fig. 3 shows the three representations where (8b) seems a good approximation of (8).

### 4.2 The Mel Scale

The Mel scale was proposed by Stevens *et al*. (1937) to model the characteristics of the nonlinear perception of the pitch by human ear. The Mel frequency as expressed as below ($f$ is measured in Hz):

$$f_{mel}(f) = 1125\ln(1 + f/700) \tag{9}$$

### 4.3 Perceptual Feature Extraction

#### 4.3.1 Linear predictive perceptual analysis

The linear predictive perceptual analysis makes use of Durbin's recursive algorithm ([6], [7]), to calculate the prediction coefficients, based on the autocorrelation coefficients. However, the autocorrelation coefficients are calculated as the Inverse Fourier Transform of a perceptually motivated power spectrum X(·). The algorithm involves the following steps ([4], [5]):

-Computation of the windowed power spectrum

-Critical band integration through a filter-bank, defined by:

$$C_k(f_{bark}) = \begin{cases} 10^{f_{bark} - f^k_{bark}} & f_{bark} \le f^k_{bark} - 0.5 \\ 1 & f^k_{bark} - 0.5 < f_{bark} < f^k_{bark} + 0.5 \\ 10^{-2.5(f_{bark} - f^k_{bark}) + 05} & f_{bark} \ge f^k_{bark} + 0.5 \end{cases} \tag{10}$$

where $f_{bark}$ is defined by either of (2), (3) or (4), and $f^k_{bark} \approx k$.

-Equally loudness pre-emphasis:

$$E(\omega) = \sqrt{\frac{(\omega^2 + 1.44 * 10^6)\omega^2}{(\omega^2 + 1.6 * 10^5)(\omega^2 + 9.61 * 10^6)}} \tag{11}$$

where $\omega=2\pi f$.

-Intensity to loudness compensation and re-sampling:

$$Q(k) = F^{1/3}(k) = (\int_0^{\pi} E(\omega) C_k(\omega) S(\omega) d\omega)^{1/3} \tag{12}$$

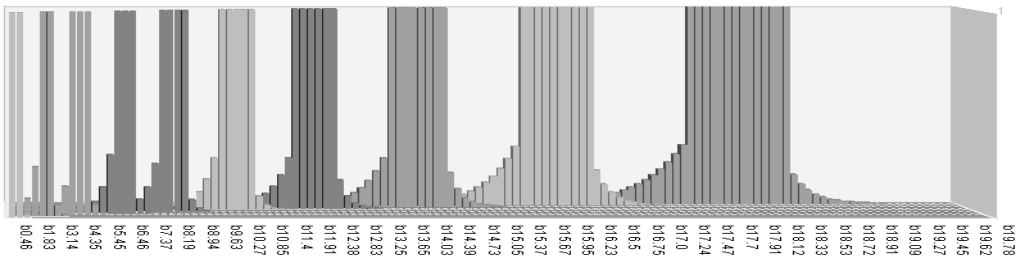


Fig. 4. Representation of the trapezoidal filters (10)

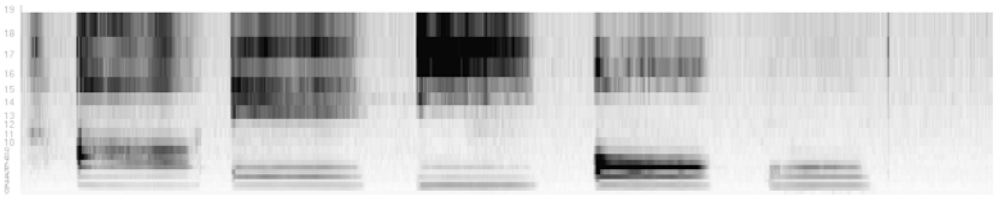

Fig. 5. The PLP spectrum calculated using the relation (12), the bank of filters (10), and the bark scale (8b)

-Inverse Fourier transform of the power spectrum

-Calculation of the perceptual prediction coefficients by the Levinson–Durbin recursion ([7]) and, based on these, of cepstral coefficients. Fig. 5 presents the PLP spectrogram with 20 bands, calculated using (12), the filter- bank (10), and the Bark scale (8b).

### 4.3.2 The Mel- perceptual feature extraction

The Mel perceptual feature extraction is also accomplished with the help of a filter bank, defined by M triangular filters, with the role of averaging the spectral energy around each central frequency. A Mel filter is defined by ([6]):

$$H_m'[k] = \begin{cases} 0 & k < f[m-1] \\ \dfrac{k - f[m-1]}{(f[m] - f[m-1])} & f[m-1] \le k < f[m] \\ \dfrac{f[m+1] - k}{(f[m+1] - f[m1])} & f[m] \le k < f[m+1] \\ 0 & k \ge f[m+1] \end{cases} \tag{13}$$

where $\sum_{m=1}^{M} H_m[k] = 1$ and $f[m] = \frac{N}{F_s} f^{-1}(f_{mel}(f_l + m\frac{f_h - f_l}{M+1}))$

In the above equations, $f[m]$ is calculated based on the lowest and highest frequency values of the filter bank, the sampling frequency and the length of a speech frame. These filters increase in spacing and decrease in height, although certain implementations make use of equal height filters. The Mel cepstrum is derived from the power spectrum calculated in the classical frequency range:

$$X[k] = \sum_{n=0}^{N-1} x[n] \cdot e^{-2\pi kn/N},\ 0 \le k < N \tag{14}$$

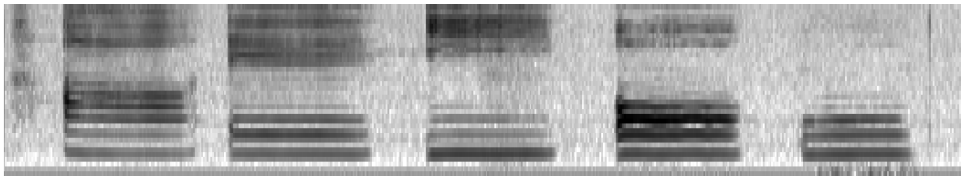

Fig. 6. Log-energy in the Mel scale filtered through a bank of 36 filters

Log-energy as filtered by one filter is defined as:

$$S[m] = \ln\left[\sum_{k=1}^{N} (X[k])^2 H_{m]}[k])\right], 0 \le m < M \tag{15}$$

Fig. 6 provides a spectrogram in the Mel scale, using 36 filters.

## 5. Experiments and Results

The goal of our research was to assess the performance of a speaker verification system based on the DTW approach and a two-factor scheme. On the other hand to investigate how speaker verification performance relies on the type of features extracted.

In our experiments we used speech samples of 9 male and 12 female speakers. Each speaker recorded between 4 and 11 sessions, uttering each time some required and arbitrary text. The total number of recorded utterances by the 21 speakers was 2737, varying from 62 to 195 per speaker. The compulsary text included six Romanian sentences: “Eu iau nouă ouă moi”, “Meniul moliei e lâna”, “Aureola e o lumină”, ”Lamiia ia anemia unui om”, “Ei au o inimă imună”, “Eu ii iau o anemonă”, pronounced once in each recording session. In the two-factor approach framework, in the training and verification processes the speakers uttered two sentences: one compulsory sentence, and a second one which was either their names or a certain one, depending on the speaker and the sentence used as the first password, of the following: “Eu iau nouă ouă moi”, “Meniul

moliei e lâna", "Aureola e o lumină", "Lamâia ia anemia unui om", "Ei au o inimă imuna". We tested the combinations using as first "password" "Eu iau nouă ouă moi", and "Meniul moliei e lina". Because the total number of utterances uttered was quite large, in order to assess the False Acceptance Rate we tested a limited number of combinations uttered by some "impostors", containing the first compulsory text and a number of sentences as the second "password", among which the above mentioned ones, depending on the reference speaker. At training and at recognition we extracted 14 cepstral coefficients on each speech frame, derived either from the PLP or from Mel-scale analysis. We used a spectrum – based criterion to remove the non-voiced frames. The first two sessions were used for training.

We assessed the averaged approach (DTWA) presented above, as compared to the classical dynamic time warping (DTW). On the other hand we show how the performance can be improved by weighting the DTW distortion, as specified by (7). $DTW_w$ denotes the weighted variant. The performance of the two methods was based on the evaluation of certain scores. In the two-factor approach, the score is derived from two different utterances of the first "password" and other two of the second "password". The utterances of the first "password" are warped resulting a distortion measure $DTW[coeff_{11}, coeff_{12}]$, and similarly the utterances of the second password produce a distortion measure $DTW[coeff_{21}, coeff_{22}]$. The score is set to:

$$S = DTW(coeff_{11}, coeff_{12}) + DTW(coeff_{21}, coeff_{22}) \tag{16}$$

For a certain client we used as reference templates the feature vectors obtained from the four recordings of the two "passwords", $coeff_{11}$, $coeff_{12}$, respectively $coeff_{21}$, $coeff_{22}$, uttered in the first two sessions. To compute the impostor model associated with a speaker we used the first two recording sessions and evaluated the scores obtained from reporting each impostor's combination of utterances of the two "passwords" ($coeff_1$, $coeff_2$) to the four reference templates ($coeff_{11}$, $coeff_{12}$, respectively $coeff_{21}$, $coeff_{22}$), obtaining the scores:

$$S_j = DTW(coeff_{1j}, coeff_1) + DTW(coeff_{2j}, coeff_2) \quad j = 1,2 \tag{17}$$

The total score obtained by the impostor utterance is set to $S_1 + S_2$. We averaged the scores of the impostors using the recurrent formula:

$$\hat{\mu}_{N+!} = \frac{1}{N+1}\sum_{i=1}^{N+1} x_i = \hat{\mu}_N + \frac{1}{N+1}(x_{N+1} - \hat{\mu}_N) \tag{18}$$

where $\hat{\mu}_N$ is the estimate of the average of $x_1, x_2, \ldots, x_N$.

The threshold was set to a weighted sum of the impostors average score and the speaker's score S (16). ($S_{1i}$, $S_{2i}$, computed as in (17), concern impostor utterance $i$, $N$ is the number of impostor utterances):

$$thr_1 = w_1 S + w_2 \sum_{i=0}^{N-1} (S_{1i} + S_{2i}) \quad (19)$$

Analogously, in the DTWA approach, the score is derived from two different utterances of the first "password" and other two of the second one. There is only one pair of "average" reference templates, ($coeff_1$, $coeff_2$), furnished by the DTWA algorithm, for each "password". The reference template is warped with the two feature vectors for each of the "password", extracted from the first two sessions. According to (6) two scores are produced by the DTWA procedure:

$$Sj = (DTW(coeff_{1j}, coeff_j) + DTW(coeff_{2j}, coeff_j))/2 \quad j = 1,2 \quad (20)$$

As above, we evaluated the scores obtained by reporting each set of two impostor utterances of the "passwords" ($coeff_{10}$, $coeff_{20}$), to the reference templates ($coeff_1$, $coeff_2$):

$$S = DTW(coeff_{10}, coeff_1, n_1, n) + DTW(coeff_{20}, coeff_2, n_{10}, n) \quad (21)$$

These impostor scores are averaged and the threshold is set to a weighted sum of this average and the speaker's score obtained as in (20):

$$thr_2 = w_1 (S_1 + S_2) + w_2 \sum_{i=0}^{N-1} S_i \quad (22)$$

*Table1*

**Two Factor Verification Rates For The PLP and Mel Approaches**

| | PLP | | MEL | |
|---|---|---|---|---|
| | FRR | FRR | FAR | FAR |
| DTWA | 26.72% | 18.40% | 30.18% | 28.56% |
| DTW | 24.75% | 17.33% | 30.77% | 23.78.% |

In the verification process, for the utterances of the two compulsory texts, we evaluated the score ($S_1$ + $S_2$) in the case of DTW approach, and $S$ in the DTWA approach and compared it to $thr_1$, $thr_2$ respectively. We compared different approaches to the perceptual analysis. The results obtained are presented in Table I. We used the weight values $w_1$=0.7625 and $w_2$ = 0.2375.

In our experiments we found useful to apply a weighted Euclidean metric:

$$d(x, y) = \sum_{i=1}^{\dim} w_i (x_i - y_i)^2 \quad (23)$$

For instance, for the evaluation based on the Mel-scale analysis we used low weights for the first cepstral coefficients, mainly for the first one ($w_1 \approx 0.01052$, $w_2 \approx 0.27027$, $w_3 \approx 0.526316$). We applied sub-unitary weights for the first three PLP-coefficients as well ($w_1 \approx 0.08333$, $w_2 \approx 0.27027$, $w_3 \approx 0.666667$). The results obtained using the two methods are presented in Table II.

*Table II*

**Verification Rates for the PLP and Mel approaches using the weighted Euclidean distance**

| | PLP | | MEL | |
|---|---|---|---|---|
| | FRR | FRR | FAR | FAR |
| DTWA | 22.21% | 22.21% | 17.49% | 16.8% |
| DTW | 20.8 | 20.8 | 15.51% | 16.6% |

To appraise the performance of the verification system based on one factor, we performed similarly the threshold setting for each client, by computing an average score for the impostors with regard to the respective client. We investigated both the DTW and the DTWA algorithms. In the DTW approach we set the score of the two utterances of the password to:

$$S = DTW(coeff_1, coeff_2) \tag{24}$$

where $coeff_1$, $coeff_2$, are the feature vectors extracted from the two utterances of the "password". The threshold was set to:

$$thr_3 = w_1 S + w_2 \sum_{i=0}^{N-1} (S_{1i} + S_{2i}) \tag{25}$$

The scores $S_{ji} = DTW(coef_j, coeff_i)$, $j$=1,2 belong to the impostor $i$ with respect to the reference templates, $coeff_1$, $coeff_2$, and $S$, estimated by (24).

In the DTWA approach the threshold is set to:

$$thr_4 = w_1 S + w_2 \sum_{i=0}^{N-1} S_i \tag{26}$$

where $S$ is the client's score:

$$S = (DTW(coeff_1, coeff) + DTW(coeff_2, coeff))/2 \tag{27a}$$

and $S_i$ is the impostor $i$ feature vector score reported to $coeff$.

$$S_i = DTW(coeff_i, coeff) \tag{27b}$$

The performance obtained using one factor is presented in table III.

*Table III*

**One factor verification rates for PLP and Mel approaches using the weighted Euclidean distance**

| | PLP | | MEL | |
|---|---|---|---|---|
| | FRR | FRR | FAR | FAR |
| DTWA | 25.41% | 23.03% | 16.59% | 24.94% |
| DTW | 22.08% | 19.45% | 17.73% | 18.59% |

We examined the weighted approach in both DTW, and DTWA, for the one-factor and the two factors schemes. The results are presented in tables IV and V. They demonstrate a significant improvement of the performance for the sheer DTW approach, from an EER about 16% to around 8.5%, using the Mel cepstral

coefficients and from 18.7% 10.7% for the PLP coefficients. The same improvement is noteworthy for the DTWA method as well, as the results reveal.

*Table IV*

**Two Factors verification rates for PLP and Mel approaches using the weighted DTW**

| | PLP | | MEL | |
|---|---|---|---|---|
| | FRR | FRR | FAR | FAR |
| DTWAw | 8.03% | 15.66% | 7.36% | 11.70% |
| DTWw | 12.04% | 9.43% | 10.70% | 6.31% |

*Table V*

**One factor verification rates for PLP and Mel approaches using the weighteing approach and the weighted Euclidean distance**

| | PLP | | MEL | |
|---|---|---|---|---|
| | FRR | FRR | FAR | FAR |
| DTWAw | 16.06% | 17.33% | 11.04% | .13.55% |
| DTWw | 18.07% | 15.66% | 15.72% | 10.82% |

To assess the class separability power of the feature sets we calculated the within-class scatter and the between-class scatter matrices $S_w$ and $S_b$ ([5]):

$$S_w = \frac{1}{K}\sum_{m=1}^{M}\left[\sum_{k=1}^{K_m}(y_{m,k}-\mu_k)(y_{m,k}-\mu_k)^T\right] \tag{28a}$$

$$S_b = \frac{1}{K}\sum_{m=1}^{M}K_m(\mu_m-\mu)(\mu_m-\mu)^T \tag{28b}$$

where $K$ is the total number of features, $K_m$ the number of features in class $m$, $M$ the number of classes, $\mu_\mu$, $\mu$ the feature mean in class $m$, and the overall feature mean, respectively. Figs. 7, 8 present images of within-class and between-class scatter matrices obtained with PLP and Mel cepstral coefficients us.

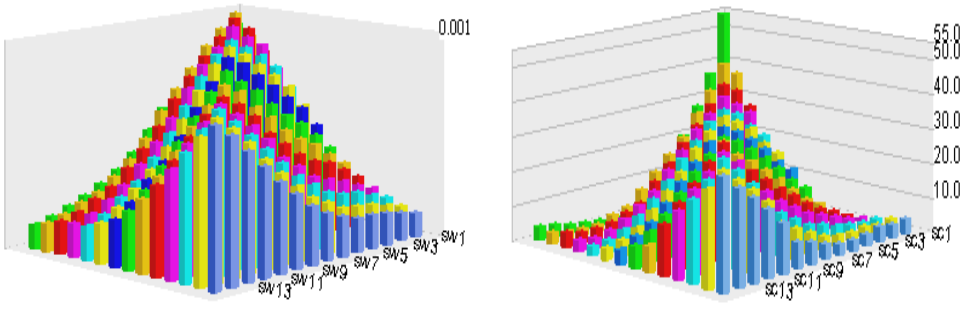


Fig. 7. The within- class scatter-matrix of the Mel cepstral coefficients (at right) and PLP cepstral coefficients (left) sets

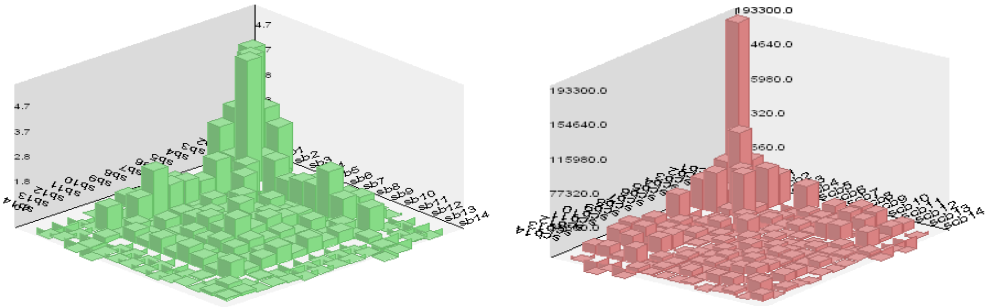

Fig. 8. The between- class scatter-matrix for the sets of Mel cepstral coefficients (at right) and PLP cepstral coefficients (a left)

For a high class-separablity the within-class matrix should be relatively small while the between-class matrix relatively high. A measure of class separability is:

$$d = trace(S^{-1}S) \tag{29}$$

If $S_w$ is not singular, a good approximation of (29) is:

$$d = trace(S_b)/trace(S_w) \tag{29a}$$

In (29) and (29a) trace is the sum of the diagonal elements of the matrix. We evaluated relation (29a) on 19 classes speakers), and 27374 PLP-cepstral features and (and on 21028 Mel-cepstral vectors. Table VI presents the class separability estimated by (29a), for the Mel-cepstral, and PLP-cepstral coefficients. The results suggest that, in terms of (31a), the PLP based approach has a good separability power, better than the Mel- based approach.

*Table IV*

**Coefficient $d = trace(S_b)/trace(S_w)$ calculated for Several Perceptual Approaches.**

| | PLP | MFCC |
|---|---|---|
| $d = trace(S_b)/trace(S_w)$ | 1443.4 | 966.15 |

## 6. Conclusions

In our research we tried to emphasize the value of the Dynamic Time Warping approach in speaker verification. We devised a new method derived from the DTW meant to spare memory resources without significantly affecting the verification performance. We examined the behavior of the DTW and the DTWA methods in the context of various characteristic feature sets. Overall, the performance of the DTW method was superior to those obtained by applying DTWA. The results improved when using weighted perceptual features sets in both approaches. While applying the non-weighted approach the best results were attained using PLP feature sets, the overall best results were obtained with the weighted mel-cepstral features. Moreover the differences in the performance between the DTW and the DTWA approaches are moderate with the mel-cepstral

coefficients. We also applied the symmetric weighting of the DTW distortion function as pointed out by Sakoe and S. Chiba in [8]. The performance increased by about 8% in all the cases, the two-factors and one-factor schemes, sheer DTW, and the DTWA method. Again the sheer DTW using Mel features achieved the best scores, the Mel-DTWA approach coming next with an about 1% handicap.

Although the performance using the two-factor approach is better than those obtained using one factor, the differences are not significant. We can explain this by the fact that the second "password" was not always pronounced properly: the speakers who used their names not always pronounced their first and last names in the same order. The evaluation based on the scatter matrices of the feature sets suggests better behavior for the PLP feature sets, however this evaluation does not take into account the weighting factor that we applied later. As one can see the large values of first Mel-cepstral coefficient "damages" the aspect of both scatter-within and scatter-between matrices.

In conclusion the DTW approach proved to be a valuable technique to be applied in speaker verification, which is essentially a speech dependant variant speaker recognition. The DTWA derivation, although weakens the DTW results, is still an alternative when the training material is much richer, for instance when adaptive techniques are applied as the verification system is operated for a long time interval. For the future we would like to try to improve the DTW performance by using the dynamic features derived from the Mel and PLP sets and also test the asymmetric an optimized weighting also proposed by H. Sakoe and S. Chiba in [8].

## R E F E R E N C E S

[1] *S. Furui*, "Cepstral analysis technique for automatic speaker verification". IEEE Trans. Acoust., Speech, Signal Processing, **vol. ASSP-29**, pp. 254–272, 1981.

[2] *M. Pawlewski, J. Jones*, Survey. Speaker verification: Part 1, Biometric Technology Today **Vol. 14**, Issue 6, June 2006, pp 9-11

[3*] P. Senin*, *Dynamic Time Warping Algorithm Review*, Information and Computer Science Departament University of Hawaii, Honolulu 2008.

[4] *H. Hermansky*, Perceptual linear predictive (PLP) analysis of speech. Jour. of Acoust. Soc. Am**.Vol**. **87(4)**, pp.1738–1752, April,1990.

[5] *M. C. Woelfel, J. Mc Donough*. Distant Speech Recognition. Chichester, West Sussex, John Wiley & Sons, June 2009, pp.135-179.

[6] *X. Huang, A. Acero, H. W. Hon*. *Spoken Language Processing:A Guide to Theory, Algorithm, and System Development*. Prentice Hall, 2001

[7]. *J.D. Markel, AH Gray Jr*. Linear Prediction of Speech Springer. New York, Verlag, 1976, pp.42-60.

[8] *H. Sakoe and S. Chiba*, "Dynamic programming algorithm optimization for spoken word recognition", Acoustics, Speech and Signal Processing, 19 IEEE Transactions on, **vol. 26**, no. 1, pp. 43{49, 1978. [Online]. Available: http://ieeexplore.ieee.org/xpls/absall.jsp?arnumber=1163055.