Combining Four Text Independent Speaker Recognition Methods

P. Thévenaz, H. Hügli
Institut de microtechnique Uni-NE
Rue Abraham-Louis Breguet 2
CH-2000 Neuchâtel

Abstract

This paper deals with automatic text independent speaker recognition in a telephone bandwidth context. First, the meaning of text independence is reviewed; then, we present our solution to this problem.

Our aim is to get a sufficient number of different methods, in order to fruitfully combine them. Hence we present four methods of text independent speaker verification. Algorithms and performances are individually analyzed before we attempt to combine them. These methods are essentially statistical in nature; they make use of cepstral vectors obtained by LPC analysis.

The first method simply characterizes the speaker by his mean cepstrum. The second method is based on the accumulation of vector quantization error of a locution by the speaker's codebook. The third method is derived from the second one by using differential cepstral vectors instead. The fourth and last method exploits the histogram of entries in a universal cepstrum codebook, according to a vector quantization technique.

The combination of the resulting distances given by these four methods is achieved by a Fisher linear discriminant analysis, which provides a great improvement in performances over any single method. The performances achieved are compared to what can be found elsewhere in the literature.

Introduction

A speaker recognition method is aimed at discovering or confirming the identity of people on the mere basis of their speech (for a review, see [1, 3, 18]). Human beings currently solve this problem quite easily in everyday life, while the machine runs properly only when the conditions are good enough (no bandwidth limitation), and the task is simple enough (password detection). However, an automatic solution would be interesting not only in restricting access to security areas, but also when conditions are less favorable, like in authenticating transactions over a telephone network, in forensic applications, or in any kind of application involving vocal print.

Text independence addresses the very text to be spoken by the user. It is usually useful to discern some gradations between text independent and text dependent mode. The most constrained type is total text dependence, often referred to as password. The text is fixed and the machine has already heard it several times in a training phase. Then comes constrained text, where the speaker has no choice about what to say (the machine dictates the phonetical content of the locution), but maybe never trained the machine before with these particular sounds, or strings of sounds. Constrained vocabulary is next, where the speaker is free to select words in a limited vocabulary known by the machine, forming any sentence he likes within the given vocabulary. Text independence is achieved by letting the speaker talk with no constraints about the content of the locution. With total independence, we would accept to release the remaining constraints, that is, accept simultaneous speech from several speakers, different emotional content (cries, laughs, tears, songs, etc.), acoustical pollution from other sources (music, traffic jam, etc.), or environmental peculiarities like diver's helium atmosphere.

A speaker recognition method would come in two forms. The first one is speaker identification, where one simply talks, and the machine tells who is speaking. The second form is speaker verification, where one has to both give an identity claim and speech input. The machine's answer may be "registered user" or "impostor".

From now on, we will restrict ourselves to text independence (vs total independence) and speaker verification (vs identification).

Furthermore, we have to define a figure of merit, by using some way of (usually a posteriori) error probability estimation [11]. Known are Minimum Average False Reject and false Acceptance rate (MAFRA), Equal Error Rate (EER), and false reject rate for a given Constant False Acceptance rate (CFA). The first one (MAFRA) sets the speaker verification trade-off (rejecting authorized users vs acknowledging impostors) such that the sum of the false reject and the false acceptance rates is minimized. The second one (EER) sets the trade-off such that both types of error are equally treated, assuming that there is a priori as many correct users as incorrect ones. The existence of the last figure of merit (CFA) pinpoints the fact that, in practice, people are mostly interested in securing an access, at the cost of some casual discomforts for the user (unmotivated rejects safer than wrong approvals). We have selected the equal error rate as basis of confrontation between methods.

Paper outline

The introduction presents the problem to be solved, addressing the text independence concern. Then, the parameter space is reviewed, leading to LPC cepstral computation. The description of four methods of speaker recognition comes next, presenting the average cepstrum, the accumulation of cepstral vector quantization error, the accumulation of differential cepstral vector quantization error, and the distribution of entries in a universal codebook. Some experimental results confirm the usefulness of each individual method. The combination of these methods is then achieved by Fisher linear discriminant analysis, leading to enhanced performances. A summary, where the results obtained are compared to published ones, precedes the conclusion.

Parameter space

The cepstrum has been chosen as main parametric representation of speech, on the basis of previous studies demonstrating its good properties for speaker characterization [5, 9, 17, 21, 22, 23].

The speech signal is acquired by a consumer-quality microphone and stored on tape. It is then low-pass filtered (8th order RC, fc=3,400 [Hz]) and sampled (fs=8,000 [Hz], 12 [bit] linear resolution). These values have been selected in order to be compatible with telephone bandwidth applications [7, 8, 10]. The gain of the system is such that no saturation occurs, while assuring a dynamic of about 10 to 11 [bit]. Every ensuing operation will be conducted using a 4 bytes floating-point representation of numbers.

First, we segment speech in frames of 30 [ms], the start of two consecutive frames being delayed by 10 [ms].

$$(1)\quad s(n) \qquad n \in [0, N-1] \qquad N = 240 \qquad \Delta N = 80$$

Then, we do some preemphasis over the signal,

$$(2)\quad y(n) = \begin{cases} 0 & n = 0 \\ s(n) - \mu \cdot s(n-1) & n \in [1, N-1] \end{cases}$$

$$\mu = 0.95$$

multiply it by a symmetrical Hamming window,

$$(3)\quad h(n) = y(n) \cdot \left(\alpha + (1-\alpha) \cdot \mathrm{COS}(\pi \cdot \frac{2n+1-N}{N+1})\right)$$

$$\alpha = 0.54$$

compute biased autocorrelation,

$$(4)\quad R(k) = \sum_{n \in [0,N-1-k]} h(n) \cdot h(n+k) \qquad k \in [0, P]$$

$$P = 14$$

and finally run LPC analysis by the Levinson algorithm, yielding inverse filter coefficients a_k satisfying:

$$(5)\quad a_j = \begin{cases} 1 & j = 0 \\ \sum_{k \in [1,P]} a_k \cdot R(|j-k|) = R(j) & j \in [1, P] \\ 0 & j > P \end{cases}$$

The regressive cepstrum conversion

$$(6)\quad c(j) = \begin{cases} 0 & j = 0 \\ -a_1 & j = 1 \\ -a_j - \sum_{k \in [1,j-1]} \frac{k \cdot c(k) \cdot a_{j-k}}{j} & j \in [2, Q-1] \end{cases}$$

$$Q = 20$$

results in a cepstrum vector of dimension Q, obtained by the usual way of LPC computation [5, 24], portraying each frame of the speech signal. Note that the cepstrum vector is redundant, as Q>P. The best choice of Q and P has not yet been investigated, because we're more interested in the magnitude of the efficiency gain between single and combined methods, than in the efficiency itself. Furthermore, it is significant of this careless approach that it takes into account every frame of the speech record, wether it is actually speech or silence.

Methods

a) <cpt>: average cepstrum

If the hypothesis of a time invariant channel between speaker and digitizing device is verifyed, then the temporal average of the cepstral vectors issued from a single speaker does represent the cascade of his average vocal tract cepstral response and the channel cepstral response. By virtue of homomorphic analysis, subtracting two averaged cepstra leads to terms which cancel out (same channel) and terms which possibly do not (vocal tract), resulting in a cepstral vector whose magnitude hopefully represents some kind of discrepancy between speakers [18].

Let $\mathbb{S}$ be a set of S speakers i and let them pronounce some locution $\mathbb{L}_i$ characterized by a set of cepstral vectors $\mathbf{C}_{i,k}$

$$(7)\quad \mathbb{S} = \{i \mid i \in [1, S]\} \qquad S = 10$$

$$\mathbb{L}_i = \{\mathbf{C}_{i,k} \mid k \in [0, L_i-1]\} \qquad L_i = 1534$$

Compute an average cepstrum for speaker i

$$(8)\quad \langle \mathbf{C}_i \rangle = \frac{1}{L_i} \cdot \sum_{k \in [0,L_i-1]} \mathbf{C}_{i,k} \qquad i \in [1, S]$$

We can now measure in an Euclidean space the squared distance between two locutions j and i

$$(9)\quad d_1^2(j,i) = |\langle \mathbf{C}_j \rangle - \langle \mathbf{C}_i \rangle|^2 \qquad (i, j) \in [1, S]$$

If this distance is small enough, we will pretend that speaker i and speaker j are but the same person. If the distance is bigger than some threshold T_1, we will pretend that two different people spoke.

b) ΣVQ: accumulation of cepstral vector quantization error

Looking at the average cepstrum exemplifies the interest we have in the steady-state part of vocal production. The speaker jumps indeed from one of these states to another while producing any utterance; this is reflected in the fact that we can classify the whole range of cepstra into a reduced set of representatives, named a codebook. This codebook can then be used to reconstruct the whole utterance by replacing each original cepstrum by the nearest representative from the codebook, leading to a small error between original and reconstructed speech. The next step is simply to accumulate this vector quantization error for all cepstra of a given utterance. We hope that a codebook tailored to a single speaker will yield small accumulated errors for utterances pronounced by himself, and bigger ones for locutions spoken by any other speaker [12, 17, 21, 22].

We used a technique named H-means in order to build the codebook [13]. This technique is recurrent; an initial condition is refined until certain criteria are met. Let $\mathbb{P}_i^t$ be a speaker's partition at step t, and match each cepstrum $\mathbf{C}_{i,k}$ to the nearest representative;

$$(10)\quad \mathbb{P}_i^t = \{\mathbf{P}_{i,g}^t \mid g \in [0,G-1]\} \qquad G = 32$$

$$q_k = \mathrm{ArgMin}_{g \in [0,G-1]} |\mathbf{C}_{i,k} - \mathbf{P}_{i,g}^t|^2 \qquad k \in [0, L_i-1]$$

compute the new partition representatives by averaging cepstra,

$$(11)\quad \mathbf{P}_{i,g}^{t+1} = \frac{1}{\sum_{k \in [0,L_i-1]} \delta(g,q_k)} \cdot \sum_{k \in [0,L_i-1]} \delta(g,q_k) \cdot \mathbf{C}_{i,k}$$

$$\delta(g,q) = \begin{cases} 1 & g = q \\ 0 & g \neq q \end{cases} \qquad g \in [0, G-1]$$

and stop when no more change occurs, the last partition obtained being the final result $\mathbb{P}_i$.

$$(12)\quad \begin{cases} \mathbb{P}_i = \mathbb{P}_i^t & \mathbb{P}_i^t = \mathbb{P}^{t+1}{}_i \\ t = t+1 & \mathbb{P}_i^t \neq \mathbb{P}^{t+1}{}_i \end{cases}$$

The distance obtained for speaker j quantized by speaker i's codebook is now:

$$(13)\quad d_2^2(j,i) = \frac{1}{L_j} \cdot \sum_{k \in [0,L_j-1]} |\mathbf{C}_{j,k} - \mathbf{P}_{i,q_k}|^2 \qquad (i, j) \in [1, S]$$

Again, we can define a threshold T2 deciding wether or not the distance is small enough to pretend that i and j are two instances of a same person.

c) Σ∂VQ/∂t: accumulation of differential cepstral vector quantization error

Until now, we have investigated two methods dealing with stable segments of speech. However, a human being never truly switches from one state to another; there is some

gradual transition instead, which has not yet been taken into account [6]. To improve this matter of fact, we devise a third method simply by replacing in equation (10) each cepstrum $C_{i,k}$ with its time differential $\partial C_{i,k}/\partial t$.

$$(14)\quad d_3^2(j,i) = \frac{1}{L_j} \cdot \sum_{k \in [1, L_j - 1]} |\partial C_{j,k}/\partial t - \dot{P}_{i,q_k}|^2$$

$$\partial C_{j,k}/\partial t = C_{j,k} - C_{j,k-1} \qquad (i, j) \in [1, S]$$

Of course, we define a new distance and a new threshold T3 as well. By doing so, we hope that each speaker has his own characteristic way of walking from one state to another; furthermore, the information available in the transient parts of the speech should be independent of that found in stable parts. The drawback is that if we find G stable states, there are G·(G-1) possible transitions to look for, urging for the need of a much bigger codebook size. Nonetheless, our careless approach leads us to keep the same value for G in method ΣVQ and Σ∂VQ/∂t, arguing that no language ever uses all possible transitions.

d) p(P): distribution of entries in a universal codebook

The vector quantization technique has two inputs and two outputs. One input is the cepstrum to be quantized, the other is the codebook; one output is the quantization error, the other is the code selected, which yet remains to be exploited. With this in mind, we design a universal codebook with the same technique as developed in ΣVQ, but with a greater codebook size, in order to represent well enough every speaker in the world.

$$(15)\quad \mathbb{P} = \{\mathbf{P}_u \mid u \in [0, U-1]\} \qquad U = 256$$

Now, let us vector quantize a locution $\mathbb{L}_i$ by this codebook, in order to estimate the probability of use of any one entry in the codebook;

$$(16)\quad p_i(\mathbf{P}_u) = \frac{1}{L_i} \cdot \sum_{k \in [0, L_i - 1]} \delta(q_k, u) \qquad u \in [0, U-1]$$

the distance obtained by this fourth method is linked to a threshold T4, and reads

$$(17)\quad d_4^2(j,i) = \sum_{u \in [0, U-1]} (p_j(\mathbf{P}_u) - p_i(\mathbf{P}_u))^2 \qquad (i, j) \in [1, S]$$

With this procedure, we pretend to discriminate between people who would have almost identical vocal tract response, but would differentiate in the global amount of time spent in any given stable state; for example, a French speaker will spend no time at all in the English sound /Θ/, like in "three". In this sense, the last two methods are more behavioral than physical, and as such, their correlation should be low. The correlation between <cpt> and ΣVQ should be low as well, because of the different time scale; the same should be true of Σ∂VQ/∂t and p(P).

Experimental results

a) experimental setup

We have already seen in equation (7) that a set of 10 speakers pronounced some 15-second-long locutions. Among these speakers are 9 men and 1 woman, between 20 and 40 years of age; each uttered, in a single session, eight sentences built up with twenty different French numerals. No two sentences were the same. The recordings took place in a quiet room, without any other special acoustic care.

We estimate the error probabilities by the leave-one-out method. We select each locution in a row and build a reference (mean cepstrum, codebooks and entries' distribution) for it, considering then all other locutions as claims for the same identity. When this claim is founded, we get an intra speaker distance; when not, an inter speaker one. Finally, we put a speaker dependent single threshold Tn yielding EER for the whole set of data pertaining to method n and speaker i.

b) <cpt>: results

Table 1 shows the confusion matrix for false acceptance, speaker by speaker. Each entry is at most 64, this value meaning that no intruder is ever rejected. Table 2 shows the false reject array. Each entry is again at most 64, this value meaning in this case that a legitimate user has no access at all. Zeros represent ideal cases.

```
--> Reference X          X
| Speaker Y          Y  Z = false_accept (Y by X)
V
```

Y\X	0	1	2	3	4	5	6	7	8	9	sum
0	0	0	0	0	0	0	0	32	0	0	32
1	0	0	0	19	1	15	13	3	14	4	69
2	0	0	0	1	0	0	0	0	0	0	1
3	0	37	21	0	4	2	0	5	2	26	97
4	0	20	17	6	0	1	0	0	39	0	83
5	0	28	0	1	0	0	0	0	20	0	49
6	0	30	0	0	0	2	0	19	0	7	58
7	14	12	5	1	0	0	11	0	0	5	48
8	0	18	0	0	13	11	0	0	0	0	42
9	0	2	2	1	0	0	0	1	0	0	6
sum	14	147	45	29	18	31	24	60	75	42	485

Table 1: false accept for <cpt>; EER = 8.7%

X\X	0	1	2	3	4	5	6	7	8	9	sum
sum	2	16	6	4	2	4	4	6	8	6	58

Table 2: false reject for <cpt>; EER = 8.7%

One can see that our <cpt> method is never totally perfect, whoever the speaker is. It is nonetheless interesting to see some individual differences; compare for example speaker 0 and speaker 1.

c) ΣVQ: results

Table 3 and 4 are equivalent to table 1 and 2, in the case of vector quantization. One can see that the average magnitude of efficiency is quite similar, although individual results may be quite different. For example, the efficiency of the ΣVQ method for speaker 0 and 1 is almost the same, albeit it was not true for the <cpt> method.

```
--> Reference X          X
| Speaker Y          Y  Z = false_accept (Y by X)
V
```

Y\X	0	1	2	3	4	5	6	7	8	9	sum
0	0	1	0	0	0	0	0	0	0	3	4
1	10	0	24	31	12	0	1	2	9	9	98
2	0	0	0	3	0	0	0	0	0	0	3
3	0	0	0	0	0	0	0	0	0	0	0
4	0	6	45	11	0	0	0	0	0	0	62
5	0	0	0	0	0	0	0	0	0	0	0
6	17	14	16	15	0	2	0	6	0	19	89
7	22	10	36	14	0	7	10	0	0	20	119
8	0	9	13	5	0	0	0	0	0	0	27
9	8	6	22	30	0	0	6	8	0	0	80
sum	57	46	156	109	12	9	17	16	9	51	482

Table 3: false accept for ΣVQ; EER = 9.1%

X\X	0	1	2	3	4	5	6	7	8	9	sum
sum	7	6	18	15	2	2	3	2	2	6	63

Table 4: false reject for ΣVQ; EER = 9.1%

To confirm that the errors done by one method can be corrected by the help of the other, we construct the scatter graph of figure 1 (speaker 0), where one can see two almost non-overlapping domains.

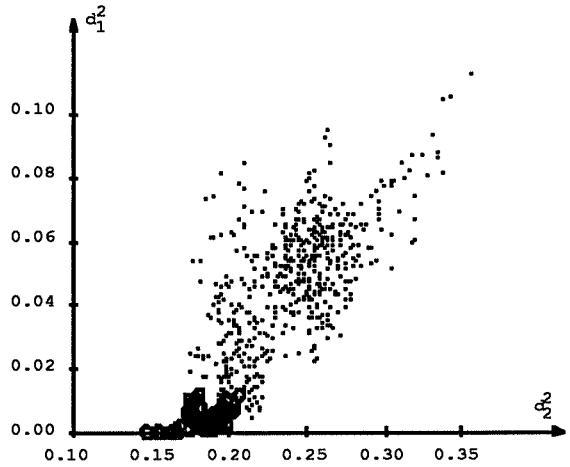

Figure 1: scatter graph <cpt> vs ΣVQ

The inter speaker domain is characterized by points, while the intra speaker domain is constituted by circles.

d) ΣδVQ/δt: results

As shown by tables 5 and 6, the results obtained by the third method are not excellent. Notwithstanding this lack of efficiency, we ascertain that this method is useful, due to its good error decorrelation property with regard to ΣVQ (scatter graph not shown).

--> Reference X X
| Speaker Y Y Z = false_accept (Y by X)
V

Y\X	0	1	2	3	4	5	6	7	8	9	sum
0	0	10	0	3	0	23	24	14	0	3	77
1	20	0	8	22	17	39	29	16	8	8	167
2	16	17	0	59	17	32	16	17	17	19	210
3	0	0	2	0	0	1	0	0	0	0	3
4	39	47	41	57	0	49	39	33	43	27	375
5	0	0	0	0	0	0	0	0	0	0	0
6	22	18	3	17	3	24	0	19	1	2	109
7	16	10	8	10	8	41	10	0	5	8	116
8	29	32	32	40	32	34	29	20	0	17	265
9	38	20	40	64	14	44	27	21	6	0	274
sum	180	154	134	272	91	287	174	140	80	84	1596

Table 5: false accept for ΣδVQ/δt; EER = 29.0%

X\X	0	1	2	3	4	5	6	7	8	9	sum
sum	22	18	15	31	12	33	23	16	14	11	195

Table 6: false reject for ΣδVQ/δt; EER = 29.0%

e) p(P): results

Tables 7 and 8 condense the results obtained for the distribution of entries in a universal codebook method.

--> Reference X X
| Speaker Y Y Z = false_accept (Y by X)
V

Y\X	0	1	2	3	4	5	6	7	8	9	sum
0	0	31	0	0	2	0	15	0	0	0	48
1	8	0	5	0	28	3	21	0	27	2	94
2	0	6	0	0	23	0	2	0	13	0	44
3	0	33	17	0	22	3	4	0	24	27	130
4	0	30	24	0	0	0	9	0	23	0	86
5	0	0	0	0	0	0	0	0	0	0	0
6	9	33	4	0	17	4	0	0	12	7	86
7	21	39	6	0	6	5	20	0	11	5	113
8	0	16	0	0	6	0	0	0	0	0	22
9	0	13	9	1	1	2	8	0	6	0	40
sum	38	201	65	1	105	17	79	0	116	41	663

Table 7: false accept for p(P); EER = 14.1%

X\X	0	1	2	3	4	5	6	7	8	9	sum
sum	6	24	8	2	12	10	12	14	14	6	108

Table 8: false reject for p(P); EER = 14.1%

Combination of the previous methods

a) Fisher linear discriminant analysis

Up to now, we have used four different methods to compute a distance between some speech input and a given reference, but we still have to find a way of combining them, at least when they disagree in accepting or rejecting a claim [14, 15, 19]. For example, a slanted line in figure 1 would design a frontier separating intra speaker and inter speaker domains. The scope of this section is to determine such a frontier in our four dimensional space, according to certain optimality criteria. The algorithm chosen is named Fisher linear discriminant [4, 16]; it aims at maximizing the ratio of inter class mean difference to intra class scatter.

Let $\mathbb{D}_\cup$ be the set of intra speakers distance measurements, and $\mathbb{D}_\cap$ the inter speakers set,

(18) $\mathbb{D}_\cup = \{\mathbf{d}_j \mid j \in [0, D_\cup\text{-}1]\}$

$\mathbb{D}_\cap = \{\mathbf{d}_k \mid k \in [0, D_\cap\text{-}1]\}$

where $D_\cup$, $D_\cap$ are the sets' cardinality, and $\mathbf{d}_i$ the quadridimensional distance vectors obtained by the other methods

(19) $\mathbf{d}_i = (d_1, d_2, d_3, d_4)^\dagger$

We describe a straight line through the origin by a vector $\mathbf{w}$. Every $\mathbf{d}_i$ can be projected onto this line, leading to a scalar measurement y_i

(20) $y_i = \mathbf{w}^\dagger \cdot \mathbf{d}_i$

We name the inter class difference $\delta_\cap$

(21) $\delta_\cap = \mathbf{w}^\dagger \cdot S_B \cdot \mathbf{w}$ $\qquad S_B = (\mathbf{m}_\cup - \mathbf{m}_\cap) \cdot (\mathbf{m}_\cup - \mathbf{m}_\cap)^\dagger$

$$\mathbf{m}_\cup = \frac{1}{D_\cup} \cdot \sum_{j \in [0, D_\cup\text{-}1]} \mathbf{d}_j \qquad \mathbf{m}_\cap = \frac{1}{D_\cap} \cdot \sum_{k \in [0, D_\cap\text{-}1]} \mathbf{d}_k$$

We name the intra class scatter $\delta_\cup$

(22) $\delta_\cup = \mathbf{w}^\dagger \cdot S_W \cdot \mathbf{w}$ $\qquad S_W = 0.5 \cdot S_\cup + 0.5 \cdot S_\cap$

$$S_\cup = \frac{1}{D_\cup} \cdot \sum_{j \in [0, D_\cup\text{-}1]} (\mathbf{d}_j - \mathbf{m}_\cup) \cdot (\mathbf{d}_j - \mathbf{m}_\cup)^\dagger$$

$$S_\cap = \frac{1}{D_\cap} \cdot \sum_{k \in [0, D_\cap\text{-}1]} (\mathbf{d}_k - \mathbf{m}_\cap) \cdot (\mathbf{d}_k - \mathbf{m}_\cap)^\dagger$$

The criterion we seek to maximize is

$$(23)\ J(\mathbf{w}) = \frac{\delta_\cap}{\delta_\cup}$$

The solution is the vector

(24) $\mathbf{w} = S_W^{-1} \cdot (\mathbf{m}_\cup - \mathbf{m}_\cap)$

onto which we project every measurement, in order to obtain a scalar global distance resulting from the combination of the distances computed by the four original methods.

b) Fisher: results

Tables 9 and 10 give the results obtained for the combination of the four previous methods. The numerical interpretation of table 9 remains the same as for previous tables, whereas in table 10 we have removed all comparisons of a test locution with the very reference built with. This leads to a maximum intra speaker confusion of 56.

--> Reference X | Speaker Y V								X Y Z = false_accept (Y by X)			
Y\X	0	1	2	3	4	5	6	7	8	9	sum
0	0	0	0	0	0	0	0	2	0	0	2
1	0	0	0	8	2	0	1	2	2	2	17
2	0	0	0	0	0	0	0	0	0	3	3
3	0	0	7	0	0	0	0	0	0	1	8
4	0	0	18	0	0	0	0	0	0	0	18
5	0	0	0	0	0	0	0	0	0	0	0
6	0	11	0	0	0	0	0	6	0	5	22
7	15	20	1	0	0	0	9	0	0	2	47
8	0	0	6	0	0	0	0	0	0	0	6
9	0	7	4	1	0	0	0	1	0	0	13
sum	15	38	36	9	2	0	10	11	2	13	136

Table 9: false accept for Fisher; EER = 3.0%

X\X	0	1	2	3	4	5	6	7	8	9	sum
sum	2	4	5	1	2	0	2	2	1	2	21

Table 10: false reject for Fisher; EER = 3.0%

One can see that the results are far better when combined than when obtained by any single method, the gain being a factor of three.

c) summary

The table 11 summarizes the results obtained for each method, with EER as figure of merit.

method	<cpt>	ΣvQ	$\Sigma \partial vQ/\partial t$	p(P)	Fisher
EER	8.7%	9.1%	29.0%	14.1%	3.0%

Table 11: EER summary

Looking at these results, one has to remember that no special care has been taken to optimize any parameter (P, Q, G, U), that pauses are included in computations, and that distance measurements are simply Euclidean instead of (e.g.) Mahalanobis [20]. Furthermore, our decision is not delayed until sufficient confidence is met, but immediate. The comparison with some other results given in literature suggests that the scores we obtain are good.

Method	EER	Comment
<cpt> [18]	5.8%	Li is 6 times longer
ΣvQ & $\Sigma \partial vQ/\partial t$ [17]	3.0%	More samples for $\partial/\partial t$ Li is 3 times longer G is 2 times bigger Already combined results
p(P)		Our method is new
Fisher [2]	1.9%	Delayed decision More combined methods

Table 12: comparison to literature

Conclusion

We have described four different methods dealing with automatic speaker recognition in a text independent mode. Experiments have shown that these are valuable by themselves, but we were still able to greatly improve the performance of any single method by combining them, using a Fisher linear discriminant technique. On our database, the error rate is divided by a factor of about three.

Bibliography

[1] B. S. Atal, *"Automatic Recognition of Speakers from Their Voices"*, Proc. IEEE, Vol. 64, Nº 4, Apr. 1976, pp. 460-475

[2] J. B. Attili, M. Savic, J. P. Campbell Jr., *"A TMS3220-Based Real Time, Text-Independent, Automatic Speaker Verification System"*, ICASSP 88, New York City, pp. 599-602

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

[4] R. O. Duda, P. E. Hart, *Pattern Classification and Scene Analysis*, Wiley-interscience publication, 1973

[5] S. Furui, *"Cepstral Analysis Technique for Automatic Speaker Verification"*, IEEE Trans. ASSP, Vol. 29, Nº 2, Apr. 1981, pp. 254-272

[6] S. Furui, *"Comparison of Speaker Recognition Methods Using Statistical Features and Dynamic Features"*, IEEE Trans. ASSP, Vol. 29, Nº 3, Jun. 1981, pp. 342-350

[7] H. Gish, K. Karnofsky, M. Krasner, S. Roucos, R. Schwartz, J. Wolf, *"Investigation of Text-Independent Speaker Identification Over Telephone Channels"*, ICASSP 85, Tampa, pp. 379-382

[8] H. Gish, M. Krasner, W. Russel, J. Wolf, *"Methods and Experiments for Text-Independent Speaker Recognition Over Telephone Channels"*, ICASSP 86, Tokyo, pp. 865-868

[9] A. L. Higgins, R. E. Wohlford, *"A New Method of Text-Independent Speaker Recognition"*, ICASSP 86, Tokyo, pp. 869-872

[10] M. J. Hunt, *"Further Experiments in Text-Independent Speaker Recognition Over Communication Channels"*, ICASSP 83, Boston, pp. 563-566

[11] K. P. Li, J. E. Porter, *"Normalizations and Selection of Speech Segments for Speaker Recognition Scoring"*, ICASSP 88, New York City, pp. 595-598

[12] K. P. Li, E. H. Wrench Jr., *"An Approach to Text-Independent Speaker Recognition with Short Utterances"*, ICASSP 83, Boston, pp. 555-558

[13] J. Makhoul, S. Roucos, H. Gish, *"Vector Quantization in Speech Coding"*, Proc. IEEE, Vol. 73, Nº 11, Nov. 1985, pp. 1551-1588

[14] N. Mohankrishnan, M. Shridhar, M. A. Sid-Ahmed, *"A Composite Scheme for Text-Independent Speaker Recognition"*, ICASSP 82, Paris, pp. 1653-1656

[15] J. M. Naik, G. R. Doddington, *"High Performance Speaker Verification Using Principal Components"*, ICASSP 86, Tokyo, pp. 881-884

[16] H. Ney, R. Gierloff, *"Speaker Recognition Using a Feature Weighting Technique"*, ICASSP 82, Paris, pp. 1645-1648

[17] A. E. Rosenberg, F. K. Soong, *"Evaluation of a Vector Quantization Talker Recognition System in Text-Independent and Text-Dependent Modes"*, ICASSP 86, Tokyo, pp. 873-876

[18] M. Shridhar, N. Mohankrishnan, *"Text-Independent Speaker Recognition: A Review and Some New Results"*, Speech Comm., Vol. 1, Nº 3-4, Dec. 1982, pp. 257-267

[19] M. Shridhar, N. Mohankrishnan, M. Baraniecki, *"Text-Independent Speaker Recognition Using Orthogonal Linear Prediction"*, ICASSP 81, Atlanta, pp. 197-200

[20] M. Shridhar, N. Mohankrishnan, M. A. Sid-Ahmed, *"A Comparison of Distance Measures for Text-Independent Speaker Identification"*, ICASSP 83, Boston, pp. 559-562

[21] R. Schwartz, S. Roucos, M. Berouti, *"The Application of Probability Density Estimation to Text-Independent Speaker Identification"*, ICASSP 82, Paris, pp. 1649-1652

[21] F. K. Soong, A. E. Rosenberg, *"On the Use of Instantaneous and Transitional Spectral Information in Speaker Recognition"*, ICASSP 86, Tokyo, pp. 877-880

[22] F. K. Soong, A. E. Rosenberg, L. R. Rabiner, B. H. Juang, *"A Vector Quantization Approach to Speaker Recognition"*, ICASSP 85, Tampa, pp. 387-390

[23] G. Velius, "Variants of Cepstral Based Speaker Identity Verification", ICASSP 88, New York City, pp. 583-586

[24] J. J. Wolf, *"Efficient Acoustic Parameters for Speaker Recognition"*, The Journal of the Acoustical Society of America, Vol. 51, Nº 6, Part 2, 1972, pp. 2044-2056