Proceedings of the First Swiss Symposium on Pattern Recognition and Computer Vision, Lausanne, Switzerland, January 30-31, 1992

A Residue-Based Approach to Text-Independent Speaker Recognition

Philippe Thévenaz, Heinz Hügli
Institut de microtechnique
Abraham-Louis Breguet 2, CH–2000 Neuchâtel, Switzerland

Abstract

The subject of this paper is speaker recognition, which aims at asserting the identity of people on the basis of their voice only. Several techniques are already available; our contribution is to be found in the investigation of features new to this domain, named residue of the linear predictive coding analysis. Our experiments in text-independent mode show that the results obtained by using these features are as good as those obtained by some other classical technique making use of the pitch, while requiring less fine tuning of parameters; hence they can be considered more appealing, because of a more direct implementation.

Introduction

The identity of people can manifest itself through many channels; among them one can find body aspect, personality traits, fingerprints, etc. The aim of speaker recognition is to use voice as only basis for guessing the identity of the talker.

Several techniques are already available [3, 4, 9, 14, 20, 25, 28]; some of these, named text-dependent, take advantage of speaker cooperation by making use of a password [2, 7, 18, 19, 22]. In the training phase, the speaker teaches the machine his very own way of pronouncing a fixed sentence; in the recognition phase, the actual pronunciation of the password is matched against the learned one, and a score is established which gives the similarity of the two sentences. A threshold decides if they are issued from the same speaker or not. This class of techniques proves to be the most efficient and the most constrained at the same time.

The other class of techniques, named text-independent, do not make use of the signal's near-term temporal evolution. Long-term statistics of the speaker's voice are computed instead, and the matching process is based upon these statistics [5, 6, 8]. The user has more freedom at the cost of some loss of efficiency. A very relevant aspect of this class of techniques is the proper choice of features to be used [1]. Ideally they should:

i) occur naturally and frequently in normal speech,
ii) be easily measurable,
iii) vary as much as possible among speakers, but be as consistent as possible for each speaker,
iv) not change over time or be affected by speaker's health,
v) not be affected by reasonable background noise, nor be dependent on specific transmission characteristics,
vi) not be modifiable by conscious effort of the speaker, or at least, be unlikely to be affected by attempts to disguise the voice.

Until now, much attention has been paid to features obtained through the linear prediction analysis of speech. This kind of analysis is frame-based; it transforms the input signal into a set of three outputs, sufficient to exactly reconstruct each frame of the signal. These are:

i) a P-component vector named the linear prediction coefficients (LPC),
ii) a scalar gain factor,
iii) a time signal named excitation, or equivalently, residue.

The LPC have been the most extensively studied of these features, particularly in a transformed version akin to the real cepstrum, considered to be a form well adapted to speech processing (speaker as well as speech recognition); less attention has been paid to the excitation. We conjecture however that each speaker has his own and possibly characteristic residue, in which we hope to find clues which are not correlated with the standard techniques. In fact, a good reason for considering the excitation is that every bit of information relevant to speaker recognition may be shared not only by the LPC and the gain factor, but also by the excitation. We want to retrieve this information hidden within the residue part, and we speculate that the resultant orthogonality with the LPC will allow to enhance the overall efficiency of a speaker recognition system which would make simultaneous use of both techniques.

Further, the residue satisfies some of the basic requirements for a "good" feature to try, as it is:

i) available at any time,
ii) easily measured and uniquely defined; no parametrization is required,
iii) independent of any linear transmission characteristics, since these are already reflected in the LPC.

In this paper, we will present our attempts to use the residue as main feature for text-independent speaker recognition. We will first briefly review the actual way of computing this residue, then we will discuss some convenient way for its representation. Next, two different techniques for its exploitation will be presented. Finally, we will compare our results to those of a third experiment, based on speaker's F0. A conclusion will put an end to this paper.

Residue extraction

We will use a speech production model which is very common in speech processing domain. It consists of an all-pole filter driven by an excitation (residue) part and scaled by a gain factor. The all-pole filter simulates the vocal tract, and the excitation the vocal folds. For analysis, the excitation is considered to be of minimal energy (which is in some sense equivalent to be flat spectrum). The normalized all-pole filter coefficients $a(i)$ are then computed according to this criterion. In short, if s is the signal, N the window length, P the analysis order and μ a preemphasis coefficient, then the Levinson algorithm solves for

$$a(i) = \begin{cases} 1 & i = 0 \\ \sum\limits_{j=1}^{P} a(j)\cdot \mathrm{R}(|i\text{-}j|) = \text{-}\mathrm{R}(i) & i \in [1, \mathrm{P}] \\ 0 & i > \mathrm{P} \end{cases}$$

where the biased autocorrelation R(i) is

$$\mathrm{R}(i) = \sum_{n=i}^{\mathrm{N}\text{-}1} b(n)\cdot b(n\text{-}i) \qquad i \in [0, \mathrm{P}]$$

where the Bartlett windowed signal b is

$$b(n) = y(n)\cdot(1 - \frac{2}{\mathrm{N}}\cdot \left| n - \frac{\mathrm{N}}{2} \right|) \qquad n \in [0, \mathrm{N}+\mathrm{P}[$$

and where the preemphasized signal y is

$$y(n) = \begin{cases} 0 & (n \le 0) \vee (n \ge \mathrm{N}) \\ s(n) - \mu\cdot s(n\text{-}1) & n \in [1, \mathrm{N}[\end{cases}$$

The residue u is the signal which would exactly generate the speech signal s if it were submitted to the all-pole filter, up to a scale factor G. In fact we have

$$s(n) = \sum_{i=1}^{\mathrm{P}} a(i)\cdot s(n\text{-}i) + \mathrm{G}\cdot u(n) \qquad n \in [\mathrm{P}, \mathrm{N}[$$

hence, after windowing, the residue becomes

$$u(n) = \begin{cases} 0 & (n < 0) \vee (n > \mathrm{N}+\mathrm{P}) \\ \frac{1}{\mathrm{G}}\cdot\left(b(n) - \sum\limits_{i=1}^{\min(n,\mathrm{P})} a(i)\cdot b(n\text{-}i) \right) & n \in [0, \mathrm{N}+\mathrm{P}] \end{cases}$$

which minimizes E with respect to $a(i)$

$$\mathrm{E} = \sum_{n=-\infty}^{\infty} u(n)\cdot u(n)$$

Residue representation

For a classical voice coding application, it is advantageous to describe the residue u as belonging to one of two states, namely voiced or unvoiced. The latter one is characterized by a noise-like structure, and a random generator can be used to approximate u for speech synthesis. The first one is characterized by pulses corresponding to vocal folds closures or openings, and a comb signal can be used whose Dirac spikes are spaced by a time interval corresponding to the pitch period. Finally, a state (voiced, unvoiced) and a scalar (F0) are sometimes considered sufficient to encode u for such an application.

However, the voice quality obtained may be still enhanced by the use of a technique named multipulse excited linear predictive coding, where the coarse residue description (state, F0) is replaced by some frame dependent short excitation pattern, selected among many candidates in a codebook, and repeated with pitch period over the whole frame [10, 17]. The description becomes (excitation codebook entry, F0). The fact that an enhancement is made possible by this technique buttresses our guess that some valuable information lays hidden within the residue.

Now, the literature on speaker recognition abounds in cases where the LPC only are emphasized, the residue being simply put aside, if ever mentioned. The reason is that a great part of the speech signal's information content is already stored in the $a(i)$, as these coefficients are used with success by virtually every study on speech or speaker recognition [7, 8, 13, 15, 16, 18, 19, 20, 24, 26, 27, 29]. A contrario, we want to examine the excitation itself in order to discover there the speaker characterizing information which may not be included in the $a(i)$, if any. First of all, we have to find a representation of the residue finer than (state, F0), and more handy than (excitation codebook entry, F0).

A matter which has to be considered is the framing process, which comes in two flavors, namely pitch synchronous or pitch asynchronous. In the first case, each voiced frame is synchronized with pitch epoches, hence such for the residue. In the other case, which is our working paradigm, the synchronization is lost. It follows that we have to find a way to reject any effect dependent on the linear phase of the residue. We decide indeed to simply get rid of any phase at all, by transforming u into its power spectrum U

$$U(k) = \left| \sum_{n=0}^{N-1} u(n)\cdot\exp(-j\cdot 2\pi\cdot\frac{nk}{N}) \right| \qquad k \in [0, N[$$

where $j=\sqrt{-1}$, $|\cdot|$ stands for the complex norm of its argument, and where the duration of u has been limited to the part driven only by s (or more to say, its windowed counterpart b).

But, as mentioned above, the spectrum of u tends to be the most flat possible, up to a spectral tilt governed by the preemphasis factor μ; this means that the values taken by U are not of direct interest. Therefore we go one step further and take v the log spectrum of U as our final representation for u, in a manner equivalent to the real cepstrum of a signal

$$v(n) = \frac{1}{N}\cdot \sum_{k=0}^{N-1} \ln(U(k))\cdot\exp(j\cdot 2\pi\cdot\frac{nk}{N}) \quad n \in [0, N/2]$$

where almost half of the values (N/2-1) are redundant and may be dropped, due to symmetries in U and v.

Finally, we are left with a representation of the residue which highlights its periodicities, while forgetting about its phase. The main difference between an F0 single scalar and this (N/2+1)-component vector is that every periodicity is taken into account, not just the fundamental one. For example, our residue representation should allow us to extract the speech signal state, and if voiced, should allow us to tell if the pitch is regular or not within a given frame.

Experiment setup

We have a set of ten speakers (one female, nine males) who produced each a set of eight utterances of 15 s duration, for a total of eighty utterances and 1200 s. The acquisition has taken place in a normal office room environment, within a single session. There were no two same sentences, so we can pretend to text independence; but they were all built with the same set of twenty randomized french digits, none of them being present twice in any single utterance. This speech has been low-pass filtered at 3.4 kHz, 48 dB/oct, and sampled at 8 kHz. The frame duration has been set to 30 ms (N=240), with a frame step of 10 ms. The preemphasis factor is μ=0.95, and the LPC order is P=14. No speech/silence detection has been used, which means that every frame is taken into account, even if it is actually mere background noise.

We select the leave-one-out kind of experiment as way of error rate estimation; that is, we select an utterance as reference and then match every other utterance against this reference. Hence, we can make a total of 560 intra-speaker tests and 5760 inter-speaker tests. The resulting efficiency estimation is given in term of equal error rate for a verification task, where each of the eighty references has its own a posteriori threshold.

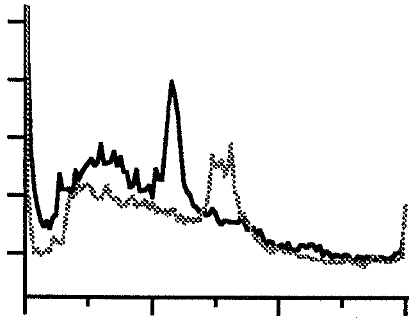
Average residue

The first experiment we will do is very simple. The basic idea is that each speaker has his own modus operandi for the generation of the excitation signal u, and further that it stays stable within time. By this we mean for example that a talker who would show a wide F0 range for a given utterance, would also show this behavior for any other utterance. According to this stationarity hypothesis, we compute a vector V representing the value of v averaged over the whole utterance duration

$$V = \frac{1}{T}\cdot \sum_{t=0}^{T-1} v(t)$$

where T is the number of frames of the given utterance and $v(t)$ is the residue cepstrum at frame t.

One can see below a plot of two superimposed curves V, corresponding respectively to the speaker labeled 0 and the speaker 1. The peaks of the two curves can easily be separated by the naked eye, in a subjective comparison process.



Graph 1: Two different average residues. *The dark curve corresponds to V of one of the eight utterances available for the speaker labeled 0, while the light curve corresponds to some V for the speaker labeled 1. The horizontal axis is the periodicity; the vertical axis denotes the values taken by V.*

In order to make an objective comparison of two V, we choose to compute a distance belonging to the class of Mahalanobis distances [12]

$$dM = \sqrt{(V''-V')\cdot M\cdot (V''-V')^{T}}$$

where M stands for the inverse covariance matrix of all the V. However, we introduce some simplification by deciding a priori not to take into account the very first component of V, which only codes for the mean log power of u, and by considering the V components as uncorrelated. Finally, we let M be the unit matrix with its first element put to zero. Our distance becomes indeed equivalent to a very classical distance

$$d(V',V'') = \|V''-V'\|$$

where $\|\cdot\|$ stands for the Euclidean norm, and where the vectors have lost their first component.

The results of this experiment are tabulated in figure 1 and figure 2, where each column collects the references per speaker, and each line collects the tests per speaker. The numbers given do reflect the errors made over 64 tests, respectively 56 tests for the intra-speaker case.

Y\X	0	1	2	3	4	5	6	7	8	9	Σ
0	◊							3		8	11
1		◊	35						10	27	72
2		44	◊						10	31	85
3		1	4	◊			21			6	32
4		33	29	2	◊				4	16	84
5						◊					0
6		1	1	13			◊			9	22
7	52							◊	1	19	72
8		33	29						◊	58	120
9	1	15	10						13	◊	39
Σ	53	127	108	13	0	0	21	3	38	174	537

Figure 1: Inter-speaker confusion matrix. *Number of errors made while comparing any utterance from speaker X (column, reference) with any utterance from speaker Y (row, test), over a corpus of 64 tries pro entry. The resulting average equal error-rate is e=9.3% for the average residue experiment*

X	0	1	2	3	4	5	6	7	8	9	Σ
	6	12	10	1	0	0	2	0	3	17	51

Figure 2: Intra-speaker confusion matrix. *Number of errors made while comparing all utterances from speaker X with themselves, autotest excluded, over a corpus of 56 tries pro entry. The resulting average equal error-rate is e=9.1% for the average residue experiment*

Finally, the global equal error rate of the technique we just presented lays around 9.2%. This result has been obtained from data acquired in a single session, with a posteriori set thresholds, and where these thresholds are reference dependent, that is, even more than speaker dependent.

Vector quantized residue

In this section, we apply to the residue a technique which is quite popular in the case of text-independent speaker recognition based on LPC cepstrum features [11, 13, 15, 16, 24, 26, 27, 29]. We want to see if this technique does also fit our present case.

The basic idea is to construct a personal codebook adapted for each speaker, in such a way that any part of the speech he produces is ever close to some entry in his own codebook, while being as far as possible from the entries of everybody else's codebook. Once obtained, the speaker's codebook can be used to vector quantize the speech to be analyzed; the raw version and the quantized version are then compared. If they are issued from the same speaker, then they will tend to be close to one another. If not, their distance will be greater, as no entry in the considered codebook has a good match to the analyzed speech.

The codebook construction is a classification problem. We select a clustering technique named K-means for its solution, where the chosen intra as well as inter-class distance is the same as in the first section, and where the prototype for each class is its gravity center. The number of classes is set to K=32. Each utterance gives rise to a single codebook, so that we have several codebooks per speaker in order to be able to make enough intra-speaker tests.

The speech is then processed, and the error of vector quantization is computed according to the distance measure presented at the preceding section. The accumulation of this error along the whole utterance leads to a scalar dVQ, compared to a threshold for the final decision of authentication or of reject. We have

$$\mathrm{dVQ} = \frac{1}{\mathrm{T}} \sum_{t=0}^{\mathrm{T}-1} \|v(t)\text{-}\mathrm{VQ}(v(t))\|$$

where the quantized vector VQ(v) is

$$\mathrm{VQ}(v) = w\left(\underset{k=0}{\overset{\mathrm{K}-1}{\operatorname{argmin}}} \|v\text{-}w(k)\| \right)$$

where $\{w(k) \mid k \in [0, \mathrm{K}[\}$ is the set of vectors found in the codebook, and where argmin(·) returns the argument which minimizes its operand.

The results of this experiment are given at figures 3 and 4, in the same format as for the preceding section.

Y\X	0	1	2	3	4	5	6	7	8	9	Σ
0	◊									1	1
1	8	◊	61		12				59	36	176
2	13	11	◊		1				10	8	43
3	31	4	17	◊	7	19	1	5	8	10	102
4	52	45	64	10	◊	6			56	48	281
5						◊					0
6	64	46	64	64	50	64	◊	60	62	64	538
7	64					1		◊	1	19	85
8	9		3						◊	27	39
9	16	3	10						16	◊	45
Σ	257	109	219	74	70	90	1	65	212	213	1310

Figure 3: Inter-speaker confusion matrix. *Number of errors made while comparing any utterance from speaker X (column, reference) with any utterance from speaker Y (row, test), over a corpus of 64 tries pro entry. The resulting average equal error-rate is e=22.7% for the vector quantized residue experiment*

X	0	1	2	3	4	5	6	7	8	9	Σ
	24	11	24	8	7	8	0	8	22	21	133

Figure 5: Intra-speaker confusion matrix. *Number of errors made while comparing all utterances from speaker X with themselves, autotest excluded, over a corpus of 56 tries pro entry. The resulting average equal error-rate is e=23.7% for the vector quantized residue experiment*

Finally, the resulting global equal error rate lays around 23.2%, which is much worse than the result obtained with the technique based upon the average residue analysis.

We explain this bad result by a possible mismatch between the analyzed features and the clustering technique, because we find that roughly more than 50% of the signal is represented by as few as 25% of codebook entries, while less than 25% of the signal requires as much as 50% of codebook entries. However, we decided not to undertake the systematic exploration of parameter configurations necessary to optimize our results.

Average F0

As a control experiment, we will compute the mean value <F0> of the pitch of each utterance, and use it as feature for speaker verification. To this purpose, we make use of a technique belonging to the class of autocorrelation analysis, which allows us to reject unvoiced frames if their autocorrelation peak is not high enough. The doubled pitch problem is solved by some median filter applied to pitch values before averaging, while the disturbances brought by high frequency noise are alleviated by the preliminary application of a low-pass filter to the speech signal and by center-clipping.

The feature measured then simply reads

$$\langle \mathrm{F0} \rangle = \frac{1}{\mathrm{T}} \sum_{t=0}^{\mathrm{T}-1} \mathrm{F0}(t)$$

where F0(t) is the instantaneous pitch of each frame, if it exists. If undefined, then the actual frame is dropped and the number of frames T is reduced by one.

The results of this experiment are given at figures 5 and 6, in the same format as for the preceding sections.

Y\X	0	1	2	3	4	5	6	7	8	9	Σ
0	◊							38	2	12	52
1		◊	52		1				33	17	103
2		57	◊		2				21	5	85
3				◊			26				26
4		12	15	18	◊						45
5						◊					0
6				48			◊				48
7	64							◊		8	72
8		29	14						◊	49	92
9	11	7	1					4	41	◊	64
Σ	75	105	82	66	3	0	26	42	97	91	587

Figure 5: Inter-speaker confusion matrix. *Number of errors made while comparing any utterance from speaker X (column, reference) with any utterance from speaker Y (row, test), over a corpus of 64 tries pro entry. The resulting average equal error-rate is e=10.2% for the F0 experiment*

X	0	1	2	3	4	5	6	7	8	9	Σ
	8	12	5	8	0	0	2	0	8	9	52

Figure 6: Intra-speaker confusion matrix. *Number of errors made while comparing all utterances from speaker X with themselves, autotest excluded, over a corpus of 56 tries pro entry. The resulting average equal error-rate is e=9.3% for the F0 experiment*

Finally, the resulting global equal error rate lays around 9.7%, which is quite close to the result obtained by the average residue analysis.

Discussion

If the efficiency of the techniques based upon the whole excitation signal or upon the pitch is about the same (at least for the average residue) then the next point is to know wether the nature of errors is different between these techniques. Put in other terms, does the information stored in the whole vector v differ from what one can find in the single scalar F0?

The answer to this question has to be found in speaker to speaker comparisons. If the success of each considered technique is about the same for each selected pair, then it is a clue, although not a proof, that the techniques are indeed equivalent. If not, then these techniques may be combined together in order to achieve a higher efficiency [9, 21].

We want to take into account every possible pair; hence we use the normalized cross-correlation coefficient γ as a similarity measure. Basically, if x and y are two signals, then

$$\gamma = \frac{<(x - <x>) \cdot (y - <y>)>}{\sigma x \cdot \sigma y}$$

where $<x>$, $<y>$ and σx, σy are respectively the mean and the square root of the variance of the signals. The values taken by γ stay in the [-1, 1] range; a zero value means that the two signals are not correlated. We tabulate the result of the comparisons in figures 7 and 8, where each entry gives the normalized cross-correlation coefficient between the errors made by each presented technique.

γ	RSD	VQ RSD	F0
RSD	1.00	◊	◊
VQ RSD	0.33	1.00	◊
F0	0.77	0.32	1.00

Figure 7: Inter-speaker correlation matrix. *Normalized cross-correlation coefficient between the three techniques considered in this paper. RSD stands for residue, VQ RSD stands for accumulation of residue vector quantization error, and F0 stands for pitch.*

γ	RSD	VQ RSD	F0
RSD	1.00	◊	◊
VQ RSD	0.52	1.00	◊
F0	0.62	0.48	1.00

Figure 8: Intra-speaker correlation matrix. *Normalized cross-correlation coefficient between the three techniques considered in this paper. RSD stands for residue, VQ RSD stands for accumulation of residue vector quantization error, and F0 stands for pitch.*

Finally, one can see that the correlations are positive and far from 0.0, which means that the behavior of the techniques considered here is not very different with respect to speaker recognition.

Conclusion

In this paper, we employed the residue as a new feature for speaker recognition. We first tried to find its convenient representation; then we made some experiments with two well known techniques used in text-independent mode. One of these, related to vector quantization, showed not very successful; it may need either a better tuning of the parameters, or a better adapted representation of the features in order to give good results. The other, related to long-term averaging, has to be compared to a control experiment based on the speaker's pitch, which gives close results while needing a higher number of parameters to tune.

The average residue and the pitch technique both depend on a window length N; (P, μ) are the only additional parameters for the technique based on the average residue. For its part, the pitch technique we made use of needs a tuning at least of the low-pass filter order and characteristic, of the voicing decision procedure, of the center-clipping procedure, of the autocorrelation peak-picking procedure and of the median filter order.

When it may be more appealing to look for optimization in a smaller parameter space, one has to remember that the feature analyzed is still a (N/2+1)-component vector, as compared to the F0 single scalar. However, whatever one can find in the residue is by definition linearly independent of the results obtained by the analysis of the LPC, up to order P. This independency is not as much explicit for the techniques based on the pitch. Thus, by a proper balance of the P order, one can hope to divide the speaker-dependent information between the LPC and the residue, which may then both yield very good candidates to be combined together.

Finally, one should not forget the role of the prediction gain, which basically encodes the speech or silence state. Our future work will be the attempt to integrate all the three outputs of the linear prediction analysis, repectively LPC, residue and gain, into one single system for text-independent speaker recognition.

References

[1] J. J. Wolf, *"Efficient Acoustic Parameters for Speaker Recognition"*, J. Acoust. Soc. Am., Vol. 51, Nº 6, Part 2, 1972, pp. 2044-2056

[2] A. E. Rosenberg, M. R. Sambur, *"New Techniques for Automatic Speaker Verification"*, IEEE Trans. ASSP, Vol. 23, Nº 2, 1975, pp. 169-176

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

[4] A. E. Rosenberg, *"Automatic Speaker Verification: A Review"*, Proc. IEEE, Vol. 64, Nº 4, 1976, pp. 475-486

[5] J. D. Markel, B. T. Oshika, A. H. Gray, *"Long-Term Feature Averaging for Speaker Recognition"*, IEEE Trans. ASSP, Vol. 25, Nº 4, 1977, pp. 330-337

[6] J. D. Markel, S. B. Davis, *"Text-Independent Speaker Recognition from a Large Linguistically Unconstrained Time-Spaced Data Base"*, IEEE Trans. ASSP, Vol. 27, Nº 1, 1979, pp. 74-82

[7] S. Furui, A. E. Rosenberg, *"Experimental Studies in a New Automatic Speaker Verification System Using Telephone Speech"*, ICASSP 80, pp. 1060-1062

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

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

[10] B. S. Atal, J. R. Remde, *"A New Model of LPC Excitation for Producing Natural-Sounding Speech at Low Bit Rates"*, ICASSP 82, pp. 614-617

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

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

[13] 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

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

[15] 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

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

[17] B. S. Atal, *"High-Quality Speech at Low Bit Rates: Multi-Pulse and Stochastically Excited Linear Predictive Coders"*, ICASSP 86, Tokyo, pp. 1681-1684

[18] M. Birnbaum, R. W. Bossemeyer, L. A. Cohen, F. X. Welsh, *"Using Cepstral Features in Speaker Verification"*, Speech Tech'86, pp. 287-290

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

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

[21] 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

[22] J. Mariani, *"Recent Advances in Speech Processing"*, ICASSP 89, Glasgow, pp. 429-440

[23] J. M. Naik, L. P. Netsch, G. R. Doddington, *"Speaker Verification Over Long Distance Telephone Lines"*, ICASSP 89, Glasgow, pp. 524-527

[24] L. Xu, J. Oglesby, J. S. Mason, *"The Optimization of Perceptually-Based Features for Speaker Identification"*, ICASSP 89, Glasgow, pp. 520-523

[25] S. Furui, *"Speaker-Dependent Feature Extraction, Recognition and Processing Techniques"*, ESCA Proc. Speaker Characterization in Speech Technology, 1990, Edimburgh, pp. 10-27

[26] J. Eatock, J. S. Mason, *"Speaker-Dependent Classification in Speaker Recognition"*, ESCA Proc. Speaker Characterization in Speech Technology, 1990, Edimburgh, pp. 94-97

[27] P. Thévenaz, H. Hügli, *"Combining Four Text-Independent Speaker Recognition Methods"*, ESCA Proc. Speaker Characterization in Speech Technology, 1990, Edimburgh, pp. 187-191

[28] R. Boite, *"La reconnaissance de la parole et la vérification du locuteur"*, AGEN Mitteilungen, Nº 52, 1990, pp. 5-13

[29] P. Thévenaz, *"Reconnaissance de locuteurs indépendante du texte"*, AGEN Mitteilungen, Nº 52, 1990, pp. 35-45