Usefulness of the LPC-Residue in Text-Independent Speaker Verification

Philippe Thévenaz
National Institutes of Health
Bethesda MD 20892-5766, USA

Heinz Hügli
Institut de microtechnique, Université de Neuchâtel
CH–2000 Neuchâtel, Switzerland

Keywords

Speech processing, speaker recognition, speaker verification, text-independence, open-test methodology, natural speech database, multi-session database, linear prediction analysis, synthesis filter, residue, complementary features.

Abstract

This paper is a contribution to automatic speaker recognition. It considers speech analysis by linear prediction and investigates the recognition contribution of its two main resulting components, namely the synthesis filter on one hand and the residue on the other hand. This investigation is motivated by the orthogonality property and the physiological significance of these two components, which suggest the possibility of an improvement over current speaker recognition approaches based on nothing but the usual synthesis filter features. Specifically, we propose a new representation of the residue and we analyse its corresponding recognition performance by issuing experiments in the context of text-independent speaker verification. Experiments involving both known and new methods allow us to compare the recognition performance of the two components. First, we consider separate methods; then we combine them. Each method is tested on the same database and according to the same methodology, with strictly disjoint training and test data sets. The results show the usefulness of the residue when used alone, even if it proves to be less efficient than the synthesis filter. However, when both are combined, the residue shows its true relevance. It achieves a reduction of the error rate which, in our case, went down from 5.7% to 4.0%.

Résumé

Cet article présente une contribution au domaine de la reconnaissance de locuteurs. Il traite de l'analyse de la parole par prédiction linéaire et examine la contribution en reconnaissance de ses deux composantes principales, le filtre de synthèse d'une part et le résidu d'autre part. Cette étude se fonde sur la propriété d'orthogonalité ainsi que l'importance physiologique de ces deux composantes, qui suggèrent que la reconnaissance du locuteur se basant exclusivement sur le filtre de synthèse peut être améliorée. En particulier, nous proposons une nouvelle représentation du résidu et nous examinons ses propriétés de reconnaissance au moyen d'expériences conduites dans un contexte de vérification du locuteur indépendante du texte. Ces expériences, utilisant à la fois des méthodes connues et nouvelles, nous permettent de comparer les contributions des deux composantes au succès de la reconnaissance. Nous commençons

par comparer les méthodes séparément, puis conjointement. Nous conduisons ces expériences en utilisant la même base de données et la même méthodologie, caractérisée par la stricte séparation des ensembles d'apprentissage et de test. Les résultats obtenus démontrent l'utilité propre du résidu, même si elle apparaît moindre que celle du filtre de synthèse. Cependant, le résidu se montre particulièrement utile quand ces deux composantes sont combinées. Dans le cas reporté ici, un taux d'erreur de 5.7% a pu être réduit à 4.0%.

Zusammenfaßung

Dieser Artikel ist ein Beitrag zur automatischen Sprechererkennung. Er widmet sich der linearen prädiktiven Sprachanalyse und untersucht den Beitrag zur Erkennung der resultierenden zwei Hauptkomponenten, namentlich des Synthesefilters einerseits und des Residuums andererseits. Diese Untersuchung ist durch die Orthogonalitätseigenschaft beider Komponenten sowie deren physiologischer Bedeutung motiviert, welche darauf hinweisen, daß übliche, nur auf Merkmale des Synthesefilters basierte Sprechererkennung verbessert werden kann. Insbesondere schlagen wir ein neues Merkmal zur Beschreibung des Residuums vor, und analysieren danach die entsprechenden Erkennungseigenschaften durch praktische Experimente im Rahmen der textunabhängigen Sprecherverifizierung. Wir vergleichen die Beiträge zur Erkennung der beiden Komponenten durch Versuche mittels bekannten sowohl originalen Methoden. Zuerst werden die Methoden einzeln verglichen, dann kombiniert. Alle Versuche werden mittels derselben Datenbank und nach dem selben Testverfahren, mit getrennten Trainings- und Testdaten, durchgeführt. Die Resultate zeigen, daß das Residuum ein recht nützliches Merkmal ist. Allein betrachtet ist es zwar weniger effizient als das Synthesefilter. Das Residuum zeigt aber seine echte Wirkungsweise im kombinierten Einsatz mit dem Synthesefilter. Es bewirkt eine Reduzierung der Fehlerrate, welche zum Beispiel von 5.7% auf 4.0% gelangt.

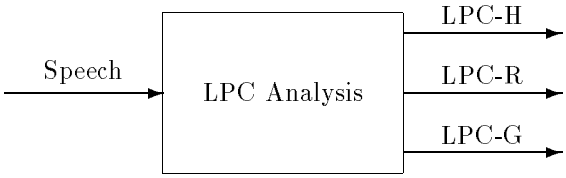

Figure 1: *Block diagram of the linear prediction analysis.*

1 Introduction

Automatic speaker recognition comes traditionally in two flavours and two colours. The two flavours are speaker identification (SI) and speaker verification (SV), and the two colours are text dependence and text independence. Many reviews of these concepts have already been published [1, 5, 12, 13, 20, 21, 22, 24, 25, 30, 36]. There are also two main trends for automatic SV pre-processing: filter-bank (FB) and linear prediction analysis (LPC) [13, 38].

About the latter, Figure 1 shows its three features, which are usually extracted from speech on a frame by frame basis. In the context of SV, the synthesis filter parameters (LPC-H) are used most often; the information available in the residue (LPC-R), if any, is usually left apart, as well as the gain signal (LPC-G). Recently, SV research has progressed in improving old and finding new methods using FB and LPC-H, see e.g. [3, 4, 6, 7, 32, 35, 37, 44], but little research has been done over the use of LPC-R since the time of [13] (1985).

Now, an interesting property of LPC is to render LPC-H orthogonal to LPC-R in some sense (that orthogonality would be lost if models of the excitation more elaborate than the strict LPC-R would be used). Because of that fundamental orthogonality, which holds up to the analysis order, it is fruitful to combine complementary information upon the user identity extracted from

both LPC-H and LPC-R. Thus, if one accepts the model that associates a vocal tract to LPC-H and an excitation to LPC-R, then one should also accept that the independence between these two physical processes is reflected upon their contribution within the model. We may then hypothesise that the vocal tract excitation differs among speakers and stays stable within a given speaker; it draws to the conclusion that LPC-R has to be investigated in order to see if the information pertaining to the speaker's identity may be extracted and made useful. The novelty of our investigation comes from the fact that the residue has been largely ignored so far in automatic speaker recognition.

In practice, informal experiments have already shown that the residue carries significant speaker specific information, for it is known that human beings listening to LPC-R find enough clues to recognise people [17]. Even if the residue is generally considered to be only a coarse approximation of the true glottal flow [16], some of its features have been shown to correlate with a subjective evaluation of voice properties [15, 34]; for example, LPC-R has been considered useful for diagnostic purposes [27]. A parent feature, namely the fundamental frequency (F0), has also been made useful in the SV context [9, 14].

In the coding domain, the use of LPC-R in multi-pulse LPC (MPLPC) [2, 31] and in its extensions like code-excited linear prediction (CELP) or glottal excitation linear prediction (GELP) [11], has been successful in enhancing the subjective quality of speech synthesisers, thus proving its relevance to speech processing. More accurate models of the glottal source (as opposed to the strict LPC-R) are also known to enhance this subjective quality [8, 10, 26, 33]. However, it is important to note, in the context of automatic SV, that these substitute for LPC-R lack the property of orthogonality to LPC-H.

Here, we investigate the usefulness of LPC-R in text-independent speaker verification by proposing a representation of the residue, and by evaluating its practical impact in a series of experiments. First, we perform experiments with new and known methods in order to establish links with already published results; then, we combine the methods together and compare the joint and separate performances. All comparisons are performed on the same database, and with the same open-test methodology. Although we restrict our experiments to a text-independent, speaker verification framework, the principle given is also valid for a more general speaker recognition task.

$\mathcal{V}(\omega, \lambda_0) = \alpha$	$\alpha = a$	$\alpha = r$
$\omega = \omega_0$	A	$\bar{R}$
$\omega \neq \omega_0$	$\bar{A}$	R

Table 1: *Verification cases.*

In the rest of this paper, we present briefly in section 2 a verification task formalism. After that, we present in section 3 the database used for our experiments and in section 4 the associated pre-processing steps, followed in section 5 by the recognition error rates observed by using features based on the synthesis filter. In section 6, we present the results obtained while using residue based features. We discuss then the combination of methods in section 7. Finally, we summarise and conclude this paper in section 8. The discussion and formalism of our residue comparison process is deferred in annex.

2 Verification

Let $\mathcal{V}$ be a verification task having an implicit access to a set of references. A first input is ω, an identity claim of a true speaker ω_0, while a second input is a speech sample λ_0. The output is a decision $\alpha \in D = \{a, r\}$ of the acceptance or reject of the claim. Formally, Ω being a set of allowed identities and L a set of speech samples, the verification task is

$$\mathcal{V} : (\Omega \times L) \rightarrow D \mid (\omega, \lambda_0) \mapsto \alpha \qquad (1)$$

If many verification experiments are carried out, then one may construct Table 1 where the observed cases are accumulated. The number of

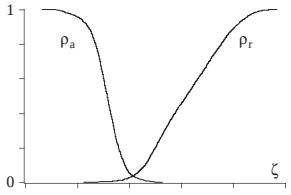

Figure 2: *Verification diagram ρ_a vs. ζ and ρ_r vs. ζ.*

correct acceptances of the claim is A, the number of false acceptances is $\bar{A}$, the correct rejects are R and the false ones $\bar{R}$. The false-acceptance and false-reject rates ρ_a and ρ_r are given by

$$\rho_a = \frac{\bar{A}}{\bar{A}+R} \quad \text{and} \quad \rho_r = \frac{\bar{R}}{A+\bar{R}} \tag{2}$$

It is frequent for the decision α to be based on the comparison of a threshold value ζ and a dissimilarity δ computed between incoming speech and some representative belonging to the references set. For example, we may have

$$\mathcal{D}(\delta,\zeta) = \begin{cases} a & \delta < \zeta \\ r & \delta \geq \zeta \end{cases} \tag{3}$$

The diagram of Figure 2 shows the behaviour of the error rates ρ_a and ρ_r with respect to the threshold ζ. The peculiar threshold value ζ_e at which both error rates have the same value yields the equal error rate $\rho_e = \rho_a(\zeta_e) = \rho_r(\zeta_e)$.

3 Database

Our database consists of French speech obtained from radio broadcasting over three consecutive days, each one being named session till the end of this paper. It follows that we do not have any control, or even knowledge, of the recording conditions. A given speaker may have experienced microphone changes between sessions, the acoustic conditions may be not the same, and the type of background noise may also differ. The topics involved are various, ranging from weather forecasts to sport events through political debates, not to mention sociological dissertations or spoken news; thus, a great amount of variability is present in our database. The number of male and female speakers is balanced.

	Sessions				Sessions		
Female	I	II	III	Male	I	II	III
F1	•	•	•	M1	•	•	•
F2	•	•	•	M2	•	•	•
F3	•	•	•	M3	•	•	•
F4	•	•	•	M4	•	•	•
F5	•	•	•	M5	•	•	•
Fa			•	Ma			•
Fb			•	Mb			•
Fc			•	Mc			•
Fd			•	Md			•
Fe			•	Me			•
Ff			•	Mf			•

Table 2: *Speakers and sessions.*

Sessions	I	II	III
References	Representatives	Thresholds	
Tests			Tests

Table 3: *Use of the sessions.*

Table 2 shows with alphabetical characters the labels associated with each speaker and with Roman numerals the sessions involved. In the training phase of our methodology, session I is used for building representatives and session II is used for estimating thresholds yielding equal error rates in a verification task conducted with test material coming again out of session II. A reference consists then in two parts, firstly a representative holding the relevant statistics for the speaker, and secondly a threshold. In testing phase, session III is used for independently testing the classifiers with the aid of the thresholds previously estimated. Table 3 presents a sum up of this procedure.

Our methodology is said to be open, or equivalently U-type, because the data set used in the learning phase (here: sessions I and II) is strictly

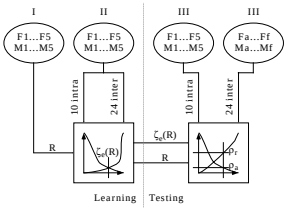

Figure 3: *Overview of the open-test methodology.*

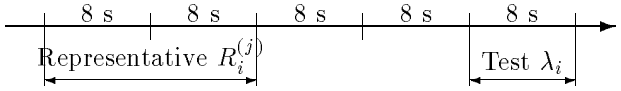

Figure 4: *Speech segments (pauses are not removed).*

disjoint from the data set used in the testing phase (here: session III). In the learning phase, we determine entirely the classifiers, thresholds included. All of the data in session III is strictly ignored up to the test phase; for example, none of it participates in the estimation of the weights which may intervene in distance computations (Mahalanobis, or weighted Euclidean), or is used in any other way before the true test of the classifier. This precaution assures one, under very mild statistical assumptions [18], that the final error rate is an upper bound of the real error rate of the SV method.

In some more details, a given reference speaker ω_i possesses 9 different representatives $R_i^{(j)}, j \in [1,9]$ estimated from his speech segments $\lambda_i^{(k)}, k \in [1,10]$ in session I. For each representative, we build a verification diagram with the aid of 10 intra-speaker distances and 24 inter-speaker distances computed from speech in session II out of the set of reference speakers $\{\mathrm{F1}\ldots\mathrm{F5}, \mathrm{M1}\ldots\mathrm{M5}\}\backslash\{\omega_i\}$. Each verification diagram in turn allows us to associate to the current representative an estimated equal error rate threshold $\hat{\zeta}_e(R_i^{(j)})$, thus creating a complete reference (note again that each speaker ω_i possesses now 9 references which are independent from one another, but all pertain to himself).

So far, we have just completed the learning phase of the acquisition procedure without using any of the test speakers $\{\mathrm{Fa}\ldots\mathrm{Ff}, \mathrm{Ma}\ldots\mathrm{Mf}\}$. We proceed with the test phase by considering the now built reference as a classifier with an *a priori* threshold $\hat{\zeta}_e(R_i^{(j)})$. Then, we estimate $\hat{\rho}_a^{(j)}$ and $\hat{\rho}_r^{(j)}$ by applying this classifier in session III to 10 intra-speaker tests and 24 inter-speaker tests using $\{\mathrm{Fa}\ldots\mathrm{Ff}, \mathrm{Ma}\ldots\mathrm{Mf}\}$. Figure 3 illustrates the whole process. The global error rates for a speaker ω_i is computed by averaging its 9 $\hat{\rho}_a^{(j)}$ and $\hat{\rho}_r^{(j)}$; the final error rates ρ_a and ρ_r are obtained by pooling the speakers together.

We then double the number of individual experiments by permuting the roles of session I and session II; the number of tests done in assessing the error rates is hence quite high: about twice those found in similar studies, e. g. [39] where 3456 verification comparisons pro method are conducted, against 6120 in our case, or 12240 if one counts the computations needed by the threshold estimation step.

This open-test methodology results in a false-acceptance error rate generally different from the false-reject error rate; the equal error rate is *not* reported since it would break down the methodology to a closed-test one, due to the *a posteriori* threshold it would imply. Instead, the overall quality is measured as the arithmetic mean of the two values ($\rho = \frac{1}{2}(\rho_a + \rho_r)$). We use our methodology under the very same conditions for all the methods at hand in order to allow their fair comparison.

4 Pre-processing

Speech is cut into contiguous non-overlapping segments of 8 s duration, without any respect to text and without pause removal. As we retain 10 consecutive segments pro speaker and pro session, it follows that our database amounts to 56 minutes of natural speech. We build each representative with a pair of consecutive segments, which corresponds to 16 s of speech. A test sample consists of a single segment, that is, 8 s. Figure 4 shows an example of this process.

Speech is low-pass filtered with $f_c = 3.4$ KHz; it is then sampled with $f_s = 8.0$ KHz and quantified with $q = 16$ bit resolution. It is cut in overlapping frames of 0.030 s duration stepped each 0.010 s. After pre-emphasis with $\mu = 0.95$, each frame is multiplied by a Bartlett window and fed to LPC with $p = 14$ as analysis order [28]. The resulting LPC-H coefficients are transformed into p cepstral coefficients.

The original LPC-R signal is first truncated to exactly the same duration as a frame's one, and then transformed to its cepstral representation through a pair of discrete Fourier transforms (note that *no* de-emphasis operation occurs). Finally, because of symmetries due to the real nature of the residue, its cepstral duration is truncated to that of just more than half a frame.

Formally, if $s(n)$ is the pre-emphasised signal, windowed to an even length N, if G is the LPC gain and if a_k is the synthesis filter coefficient of order k, then the truncated residue $u(n)$ is given by

$$u(n) = \frac{1}{G}\left(s(n) - \sum_{k=1}^{\min(p,n)} a_k \cdot s(n-k)\right) \forall n \in [0, N[\quad (4)$$

Its amplitude spectrum is given by

$$|U(k)| = \left|\sum_{n=0}^{N-1} u(n) \cdot e^{-j2\pi nk/N}\right| \forall k \in [0, N[\quad (5)$$

The LPC-R real cepstrum then reads

$$v(n) = \frac{1}{N}\sum_{k=0}^{N-1} \ln|U(k)| \cdot e^{j2\pi nk/N} \forall n \in [0, N/2] \quad (6)$$

Figure 5 presents an example of an unvoiced and of a voiced residue. The spike present in the lower right part of the figure can be clearly associated with the fundamental frequency (F0) of the corresponding speech sample. Some authors often summarise the whole residue in just one number representing F0 [29, 40]. We feel however that the residue as a whole carries richer information than the fundamental frequency alone.

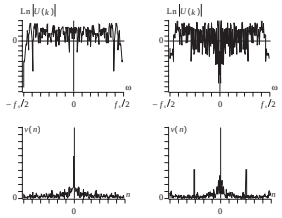

Figure 5: *Amplitude spectrum (top) and real cepstrum (bottom) of a residue. The left part is unvoiced, the right part is voiced.*

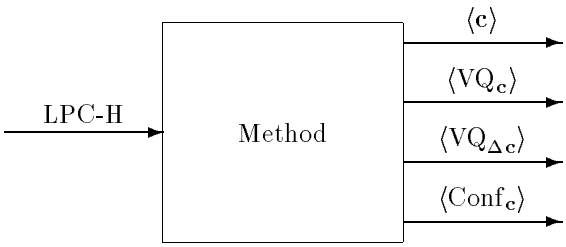

Figure 6: *Methods based on LPC-H.*

5 SV by LPC-H complex cepstrum

The representation of LPC-H is usually done in terms of its complex cepstrum $\mathbf{c}$, which possesses alleged good properties for SV when used in conjunction with (sometimes weighted) Euclidean distance or with Mahalanobis distance [19]. We discuss here the four recognition methods given in Figure 6; their results are given in Table 4.

The first method is $\langle\mathbf{c}\rangle$ the long-term average of the complex LPC-H cepstrum. It is a well known text-independent recognition method, see e. g. [39]. We compared three distances (Euclidean d_e, weighted Euclidean d_w and Mahalanobis d_M) [18].

The second method is $\langle\mathrm{VQ}_\mathbf{c}\rangle$ the long-term average error of the vector quantization (VQ) distortion, which is another well known text-independent recognition method [41]. Here, the codebook size is $K = 32$.

The third method $\langle\mathrm{VQ}_{\Delta\mathbf{c}}\rangle$ applies to differen-

Method	Dist.	ρ_a %	ρ_r %	$\frac{1}{2}(\rho_a+\rho_r)$ %
$\langle \mathbf{c} \rangle$	d_e	39.8	55.9	47.9
	d_w	10.4	21.9	16.2
	d_M	6.2	25.7	15.9
$\langle \mathrm{VQ}_{\mathbf{c}} \rangle$	d_e	4.6	10.3	7.5
	d_w	3.3	8.2	5.7
$\langle \mathrm{VQ}_{\mathbf{\Delta c}} \rangle$	d_e	42.5	25.2	33.8
$\langle \mathrm{Conf}_{\mathbf{c}} \rangle$	d_e	9.0	10.7	9.9
	d_w	7.3	11.6	9.4
	d_M	0.3	60.9	30.6

Table 4: *Error rates of LPC-H methods.*

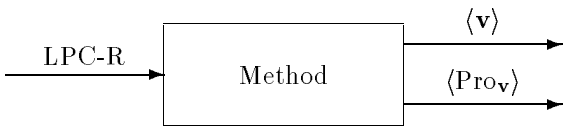

Figure 7: *Methods based on LPC-R.*

Method	Dist.	ρ_a %	ρ_r %	$\frac{1}{2}(\rho_a+\rho_r)$ %
$\langle \mathbf{v} \rangle$	d_e	16.3	14.8	15.6
	d_w	14.0	18.6	16.3
	d_M	0.4	84.6	42.5
$\langle \mathrm{Pro}_{\mathbf{v}} \rangle$		11.4	14.3	12.9

Table 5: *Error rates of LPC-R methods.*

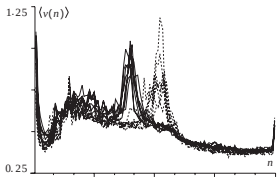

Figure 8: *Average residue real cepstra of two speakers. Speaker A is shown with full lines and speaker B with dashed lines.*

tial complex LPC-H cepstra the VQ distortion recognition method [41]. Here, the codebook size is again $K = 32$. We observe that our results obtained using this method are very bad; preliminary experiments discouraged us to attain a better success with the two other distances d_w and d_M. It appears then that, on our data, the $\langle \mathrm{VQ}_{\mathbf{\Delta c}} \rangle$ method performs worse than in the case reported in [41].

The fourth text-independent speaker recognition method $\langle \mathrm{Conf}_{\mathbf{c}} \rangle$ is called conformity [42]. It aims at recovering the information unused by VQ distortion, by looking specifically at the frequency of selection of the codebook entries. The compared features correspond then to a histogram; looking at such a histogram as a vector, the comparison is made using d_e, d_w or d_M distances. The codebook is speaker-independent; its size is $K = 128$.

6 SV by LPC-R real cepstrum

We will use LPC-R in a feature called LPC-R real cepstrum. The reason for the choice of this LPC-R representation is that the original residue is primarily a time-domain feature, but as we don't want to consider synchronisation with pitch epochs and simultaneously want to get rid of its phase contribution, we compute and retain only its amplitude spectrum (which tends to be flat). Now, this intermediate feature is meaningful only when transformed back from the frequency domain to the time domain. Furthermore, we arbitrarily decide to introduce a logarithmic nonlinearity; it follows from these considerations that we investigate here the properties of the LPC-R real cepstrum $\mathbf{v}$ as the final representation for the residue. Since the phase information is no more present, that representation is lossy; however, it tends to better satisfy the set of requirements given in [43]. We have tried the two recognition methods presented in Figure 7; their results are given in Table 5.

The first method is $\langle \mathbf{v} \rangle$ the long-term average of the LPC-R real cepstrum. It is illustrated in Figure 8, where one can see a set of eight samples

for each of two male speakers. It is quite apparent from the figure that speakers A and B show both a low intra-speaker variability and a high inter-speaker variability.

The second method $\langle \mathrm{Pro}_{\mathbf{v}} \rangle$ is new. We call it prominence. It is an attempt at a better use of $\mathbf{v}$ based on the observation that its most important feature is the presence of a cepstral peak when faced to a voiced frame. Hence, we retain as significant data only those test residue cepstral values which exceed the reference average residue cepstrum for the considered component, and we weight these prominent peaks by their variance square root. A global distance measure is then based on this kind of clipped data. This new way of comparing two speech samples, while exacerbating the importance of these prominent peaks, is still richer than the mere use of the fundamental frequency because the whole range of frequencies contributes to the comparison, and because each data frame may influence the comparison result, be it voiced or not.

Comparing results of Table 4 with those of Table 5, one can see that our residue-based methods are less efficient for SV than some traditional methods based on the use of LPC-H. However, it also appears clearly that our new LPC-R methods are still relevant to the problem, achieving an error rate of 12.9% only. In terms of rank-order, these methods fit the middle range, coming as third ($\langle \mathrm{Pro}_{\mathbf{v}} \rangle$) and forth ($\langle \mathbf{v} \rangle$) out of the six investigated methods. The whole ranking reads, from best to worst, $\langle \mathrm{VQ}_{\mathbf{c}} \rangle < \langle \mathrm{Conf}_{\mathbf{c}} \rangle < \langle \mathrm{Pro}_{\mathbf{v}} \rangle < \langle \mathbf{v} \rangle < \langle \mathbf{c} \rangle < \langle \mathrm{VQ}_{\mathbf{\Delta c}} \rangle$.

So far, we have given a minimal bound for the performance obtained using the residue as a stand-alone feature, and we have been able to compare that performance estimation with that of more traditional methods. But our real aim was indeed to make a joint use of LPC-H and LPC-R; we will see in the next section how we achieve this goal as we experiment several combinations of methods, comparing the results obtained using cross-features to those obtained throughout the use of a single feature.

	LPC-H				LPC-R	
$\frac{1}{2}(\rho_a + \rho_r)$	α)	β)	γ)	δ)	ϵ)	ζ)
α) $\langle \mathrm{VQ}_{\mathbf{\Delta c}} \rangle$	33.8					
β) $\langle \mathbf{c} \rangle$	<u>13.8</u>	16.2				
γ) $\langle \mathrm{Conf}_{\mathbf{c}} \rangle$	11.6	10.5	9.4			
δ) $\langle \mathrm{VQ}_{\mathbf{c}} \rangle$	11.6	8.2	<u>5.3</u>	5.7		
ϵ) $\langle \mathrm{Pro}_{\mathbf{v}} \rangle$	<u>12.2</u>	<u>9.4</u>	<u>7.0</u>	<u>5.7</u>	12.9	
ζ) $\langle \mathbf{v} \rangle$	18.8	<u>11.9</u>	<u>8.4</u>	<u>5.3</u>	13.7	15.6

Table 6: *Average error rates of combined methods.*

	LPC-H				LPC-R	
$\frac{\min(\rho_1,\rho_2)}{\rho_{12}}$	α)	β)	γ)	δ)	ϵ)	ζ)
α) $\langle \mathrm{VQ}_{\mathbf{\Delta c}} \rangle$	1.00					
β) $\langle \mathbf{c} \rangle$	<u>1.17</u>	1.00				
γ) $\langle \mathrm{Conf}_{\mathbf{c}} \rangle$	0.81	0.90	1.00			
δ) $\langle \mathrm{VQ}_{\mathbf{c}} \rangle$	0.49	0.70	<u>1.08</u>	1.00		
ϵ) $\langle \mathrm{Pro}_{\mathbf{v}} \rangle$	<u>1.06</u>	<u>1.37</u>	<u>1.34</u>	<u>1.00</u>	1.00	
ζ) $\langle \mathbf{v} \rangle$	0.83	<u>1.31</u>	<u>1.12</u>	<u>1.08</u>	0.94	1.00

Table 7: *Relative gains of combined methods.*

7 Combining LPC-H and LPC-R

We combine here pairwise the best results of the six methods observed so far. Out of these, four deal with LPC-H, namely $\langle \mathrm{VQ}_{\mathbf{\Delta c}} \rangle$ (d_e), $\langle \mathbf{c} \rangle$ (d_w), $\langle \mathrm{Conf}_{\mathbf{c}} \rangle$ (d_w) and $\langle \mathrm{VQ}_{\mathbf{c}} \rangle$ (d_w). The other two use LPC-R as characteristic feature, namely $\langle \mathbf{v} \rangle$ (d_e) and $\langle \mathrm{Pro}_{\mathbf{v}} \rangle$. The joint use of all these methods is obtained through a weighted sum of the distances observed individually; the weights are chosen so that the variance contributions are equalised within each pair. We give in Table 6 the observed results. We underlined the entries denoting the existence of a improvement over the best individual method of the considered pair ($\rho_{12} \leq \min(\rho_1, \rho_2)$). The relative value $g = \min(\rho_1, \rho_2)/\rho_{12}$ of this gain is reported in Table 7.

We observe in these tables that complementary techniques appear most often when combining LPC-H with LPC-R. In more details, as much

as 7/8 cases are efficient in an LPC-H–LPC-R combination, none in the pair LPC-R–LPC-R, while 2/6 cases only are efficient in the pair LPC-H–LPC-H. These two last cases are the pair $\langle \mathbf{c}, \mathrm{VQ}_{\Delta\mathbf{c}} \rangle$, which respectively emphasises long- and short-term speech behaviour, and the pair $\langle \mathrm{VQ}_{\mathbf{c}}, \mathrm{Conf}_{\mathbf{c}} \rangle$, where we benefit from the fact that these two methods address different kinds of information.

The next step consists in considering more than two methods together, beginning with three. The number of possible triplets amounts then to $C_3^6 = 20$ if one takes into account the six available methods. Rather than examining each triplet and conducting a separate recognition experiment for each one, we exploited the results obtained so far and retained only those triplets of methods $\langle a, b, c \rangle$ for which simultaneously $\rho_{ab} \leq \min(\rho_a, \rho_b)$ and $\rho_{ac} \leq \min(\rho_a, \rho_c)$ and $\rho_{bc} \leq \min(\rho_b, \rho_c)$. Table 6 or 7 tells us that only three cases have to be considered for efficient triplets of combined methods, namely $\langle \mathbf{c}, \mathrm{VQ}_{\Delta\mathbf{c}}, \mathrm{Pro}_{\mathbf{v}} \rangle$, $\langle \mathrm{VQ}_{\mathbf{c}}, \mathrm{Conf}_{\mathbf{c}}, \mathbf{v} \rangle$ and $\langle \mathrm{VQ}_{\mathbf{c}}, \mathrm{Conf}_{\mathbf{c}}, \mathrm{Pro}_{\mathbf{v}} \rangle$.

Now, no more methods can be further added without reducing the global success rate; that is, no quartet, quintet or sextet of methods satisfies the condition for its members to be mutually pairwise complementary. Letting apart $\langle \mathrm{VQ}_{\Delta\mathbf{c}} \rangle$ the worst individual method, and considering that $\langle \mathrm{Pro}_{\mathbf{v}} \rangle$ is more efficient than $\langle \mathbf{v} \rangle$, we finally retain as winning triplet $\langle \mathrm{Pro}_{\mathbf{v}} \rangle$, $\langle \mathrm{Conf}_{\mathbf{c}} \rangle$ (d_w) and $\langle \mathrm{VQ}_{\mathbf{c}} \rangle$ (d_w). These three methods happen to show also eventually the best individual results on our database.

There are several ways to select weights for combining the methods together. In the process of building Table 6, we wanted a uniform representation of the success of the combined methods in order to compare them without bias. Hence, we used weights which equalise the variance, this choice giving the same importance to each method in the pair. Now, we are in a position where a given triplet has already been selected, and our goal is no more comparison but efficiency. The way to give a weight in each method in the triplet is then to consider the inverse of its error rate, still multiplied with the variance normalisation factor. Accordingly, a good method retains relatively more significance than a worse one. With this choice of combination weights, with a priori thresholds and in an open-test methodology, the global verification false-acceptance and false-reject mean error rate is 4.0%.

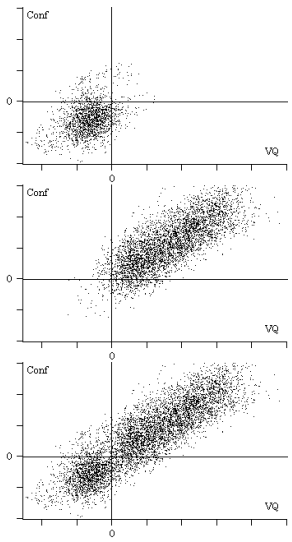

Figure 9: *Scatter diagrams* $\langle \mathrm{VQ}_{\mathbf{c}} \rangle$–$\langle \mathrm{Conf}_{\mathbf{c}} \rangle$. *The first diagram shows the intra-speaker domain and the second diagram shows the inter-speaker domain. In the third and last diagram, the size of the intra-speaker dots is made larger in order to permit the comparison of the two domains. The distances are translated by the subtraction of the associated thresholds, which explains the appearance of negative distances.*

Figures 9, 10 and 11 show the scatter diagrams associated with the considered pairs of methods. In these diagrams, the individual ζ_e threshold estimations for each representative are first subtracted from the distance values, and then the results are globally weighted in such a way that the average of intra- and inter-speaker square root variances take the same value whatever the SV method is. This procedure explains the presence of negative distances and eases the comprehension of the diagrams, since no visual artefact arises from scaling or clipping (all data points are

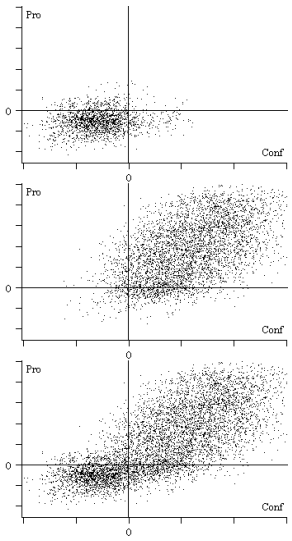

Figure 10: *Scatter diagrams* $\langle \mathrm{Conf}_{\mathbf{c}} \rangle$–$\langle \mathrm{Pro}_{\mathbf{v}} \rangle$. *The disposition is the same as for the preceding figure. Here, the domains are definitely* not *stripe-like, which tends to confirm a good independence between the two considered SV methods.*

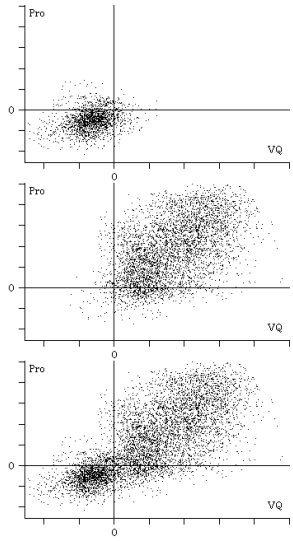

Figure 11: *Scatter diagrams* $\langle \mathrm{VQ}_{\mathbf{c}} \rangle$–$\langle \mathrm{Pro}_{\mathbf{v}} \rangle$. *The disposition is the same as for the two preceding figures. Here, the independence shows up perceptually even better, although it is not confirmed by Table 6 or Table 7.*

present in each diagram). As the shape of the clusters of points clearly differ from just a stripe, the selected methods are truly complementary and the intra- and inter-speaker domains are well separated.

In Figure 9, where the two methods at hand examine a feature dependent both on LPC-H, we see that the shape of the two regions in the scatter diagram is rather elongated, most markedly in the inter-speaker domain. This happens even though $\langle \mathrm{VQ_c} \rangle$ and $\langle \mathrm{Conf_c} \rangle$ are supposed to address different kinds of information within the data. By contrast, the next two figures do not exhibit the same behaviour; it is much easier to find in them points where one method yields a short distance and the other a big one at the same time, which tends to confirm our hypothesis stating that a distance computed by using a feature based on LPC-H is independent of a distance computed through LPC-R.

8 Summary, discussion and conclusions

We motivated our approach and argued that the residue, in the role of a feature complementary to the synthesis filter, might be useful for speaker recognition. In order to verify this claim, we conducted a large series of experiments using an open-test methodology in a multi-session database consisting of non-constrained speech from 22 speakers. We chose to implement a text-independent verification task, which allows a robust estimation of the full-scale performance of the examined methods and the relevance of the involved features.

We first compared the separate verification performances of several methods exploiting linear prediction based features, some from the synthesis filter (LPC-H) and some others from the residue (LPC-R). The best results were observed when using LPC-H in conjunction with vector quantization (VQ). However, methods based on the use of LPC-R showed also good results, being better than some other LPC-H methods. Hence, although LPC-H produced the best available results, we observed that methods based on LPC-R are still useful for speaker verification, even when used alone.

We then combined pairwise the previous methods and determined their joint performance. We conducted these experiments in order to compare combinations of mixed LPC-H and LPC-R methods with combinations of methods based on the same features. Our results pointed out that methods based on LPC-H and methods based on LPC-R were in general complementary, while methods based on features of a same kind were not, which confirms our hypothesis. In particular, the best available LPC-H method and the best available LPC-R method could be combined with good success, resulting in an overall error rate reduction from 5.7% (best separate method) to 4.0% (best combination of three methods). Therefore, the residue is not only a useful feature for speaker recognition on its own, but above all combines favourably with methods based on synthesis filter, as expected.

There are however several topics which should be addressed in future work. Some concern practical matters; for example, although our database is rather large in comparison to other authors, especially with respect to the number of verification tests ($12240 \mathrm{method}^{-1}$), the number of speakers involved still stays small (here: 22). Further, the inter-session stability of LPC-R (here: 3 consecutive days) or, from an application point of view, the robustness of the LPC-R features with respect to noise in the transmission channels (for example: telephone) and in the background (for example: music or other speakers) is not addressed in this paper.

Other topics concern more fundamental matters. For example, we feel important to investigate other representations of the residue, or to determine the optimal LPC analysis order p with respect to SV *when the residue is taken into account*, or to discover the most efficient distance (given our proposed LPC-R representation), or to look for the best LPC-R representation (given a simple Euclidean distance). Since it is well known that the speaker identity is spread on virtually all aspects of the speech signal, from acoustic cues to semantic ones [23], we feel also important not to ignore the last output of an LPC

analysis, namely the gain factor LPC-G. So far, the investigation presented in this paper and the performed experiments confirm the expected significance of the residue for automatic speaker verification.

Annex: Comparison by prominence

Formally, if the superscripts (i) and (k) relate respectively to a test speech sample and to a reference speech sample, and if T is the associated length, then we may compute the reference clipping value of the $\langle \mathrm{Pro}_{\mathbf{v}} \rangle$ method as

$$\left\langle v_n^{(k)} \right\rangle = \frac{1}{T^{(k)}} \cdot \sum_{t=1}^{T^{(k)}} v_n^{(k)}(t) \quad \forall n \in [0, N/2] \tag{7}$$

Let $\varepsilon(x)$ be the Heaviside function

$$\varepsilon(x) = \begin{cases} 0 & x \leq 0 \\ 1 & x > 0 \end{cases} \quad \forall x \tag{8}$$

Let $d_n^{(i,k)}(t)$ be the positive difference between a test residue (i) at time t and the averaged representative (k)

$$d_n^{(i,k)}(t) = \varepsilon\left(v_n^{(i)}(t) - \left\langle v_n^{(k)} \right\rangle\right) \cdot \left(v_n^{(i)}(t) - \left\langle v_n^{(k)} \right\rangle\right)$$

$$\forall n \in [0, N/2] \ \forall t \in [1, T^{(i)}] \tag{9}$$

Let $w_n^{(i,k)}$ be the ratio between the speech sample length and the number of strictly positive occurrences of $d_n^{(i,k)}$

$$w_n^{(i,k)} = \frac{1}{T^{(i)}} \cdot \sum_{t=1}^{T^{(i)}} \varepsilon\left(v_n^{(i)} - \left\langle v_n^{(k)} \right\rangle\right) \quad \forall n \in [0, N/2] \tag{10}$$

Let $\sigma_n^{(k)}$ be the square root variance of the strictly positive differences within the reference itself

$$\sigma_n^{(k)} = \sqrt{\frac{\sum_{t=1}^{T^{(k)}} \left(d_n^{(k,k)}\right)^2 - \frac{1}{T^{(k)} w_n^{(k,k)}} \cdot \left(\sum_{t=1}^{T^{(k)}} d_n^{(k,k)}\right)^2}{T^{(k)} w_n^{(k,k)} - 1}}$$

$$\forall n \in [0, N/2] \tag{11}$$

Normalising the strictly positive difference between a test and a reference by the square root variance previously computed, and lowering the importance of great values by a well chosen non-linearity, one gets

$$p_n^{(i,k)} = \frac{1}{T^{(i)} w_n^{(i,k)}} \cdot \sum_{t=1}^{T^{(i)}} \ln\left(1 + \frac{d_n^{(i,k)}}{\sigma_n^{(k)}}\right) \quad \forall n \in [0, N/2] \tag{12}$$

Finally, the distance between a pre-computed reference representative $(\langle \mathbf{v}^{(k)} \rangle, \sigma^{(k)})$ and a test speech sample is given by

$$\delta\left((i),(k)\right) = \sqrt{\sum_{n=0}^{N/2} \left(\frac{w_n^{(i,k)} p_n^{(i,k)} - w_n^{(k,k)} p_n^{(k,k)}}{w_n^{(i,k)} + w_n^{(k,k)}}\right)^2} \tag{13}$$

References

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

[2] B. S. Atal (1986), "High Quality Speech at Low Bit Rate: Multi-Pulse and Stochastically Excited Linear Predictive Coders," *Proc. ICASSP* (Tokyo), pp. 1681–1684.

[3] Y. Bennani, P. Gallinari (1994), "Connectionist Approaches for Automatic Speaker Recognition," *ESCA Workshop on Automatic Speaker Recognition, Identification and Verification* (Martigny), pp. 95–102.

[4] F. Bimbot, L. Mathan (1994), "Second-Order Statistical Measures for Text-Independent Speaker Identification," *ESCA Workshop on Automatic Speaker Recognition, Identification and Verification* (Martigny), pp. 51–54.

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

[6] M. J. Carey, E. S. Parris, J. S. Bridle (1991), "A Speaker Verification System Using Alpha-Nets," *Proc. ICASSP* (Toronto), pp. 397–400.

[7] M.-S. Chen, P.-H. Lin, H.-S. Wang (1993), "Speaker Identification Based on a Matrix Quantization Method," *IEEE Trans. ASSP*, Vol. 41, No. 1, pp. 398–403.

[8] D. G. Childers, K. Wu (1990), "Quality of Speech Produced by Analysis-Synthesis," *Speech Comm.*, Vol. 9, pp. 97–117.

[9] D. G. Childers, K. Wu (1990), "Gender Recognition from Speech. Part II: Fine Analysis," *J. Acoust. Soc. America*, Vol. 4, Part 1, pp. 1841–1856.

[10] D. G. Childers, C. K. Lee (1991), "Vocal Quality Factors: Analysis, Synthesis and Perception," *J. Acoust. Soc. America*, Vol. 90, No. 5, pp. 2394–2410.

[11] D. G. Childers, H. T. Hu (1994), "Speech Synthesis by Glottal Excited Linear Prediction," *J. Acoust. Soc. America*, Vol. 96, No. 4, pp. 2026–2036.

[12] P. Corsi (1981), "Speaker Recognition: A Survey," *Proc. Second NATO Advanced Study Institute on Speech Processing* (Bonas), pp. 277–308.

[13] G. R. Doddington (1985), "Speaker Recognition — Identifying People by Their Voices," *Proc. IEEE*, Vol. 73, No. 11, pp. 1651–1664.

[14] V. Dubreucq, C. Vlœberghs (1994), "The Use of the Pitch to Improve an HMM-Based Speaker Recognition Method," *ESCA Workshop on Automatic Speaker Recognition, Identification and Verification* (Martigny), pp. 15–17.

[15] L. Eskenazi, D. G. Childers (1990), "Acoustic Correlates of Vocal Quality," *J. Speech and Hearing Research*, Vol. 33, pp. 298–306.

[16] G. Fant (1993), "Some Problems in Voice Source Analysis," *Speech Communication*, Vol. 13, pp. 7–22.

[17] T. C. Feustel, G. A. Velius, R. J. Logan (1989), "Human and Machine Performance on Speaker Identity Verification," *Speech Tech'89*, pp. 169–170.

[18] K. Fukunaga (1972), *Introduction to Statistical Pattern Recognition* (Academic Press), 369 p.

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

[20] S. Furui (1986), "Research on Individuality Features in Speech Waves and Automatic Speaker Recognition Techniques," *Speech Comm.*, Vol. 5, No. 2, pp. 183–187.

[21] S. Furui (1990), "Speaker-Dependent Feature Extraction, Recognition and Processing Techniques," *ESCA Proc. Speaker Characterisation in Speech Technology* (Edinburgh), pp. 10–27.

[22] S. Furui (1994), "An Overview of Speaker Recognition Technology," *ESCA Workshop on Automatic Speaker Recognition, Identification and Verification* (Martigny), pp. 1–9.

[23] A. Giannini, M. Pettorino, U. Cinque (1989), "Speaker's Identification by Voice," *Eurospeech* (Paris), Vol. 1, pp. 283–286.

[24] H. Gish, M. Schmidt (1994), "Text-Independent Speaker Identification," *IEEE Signal Processing Magazine*, Vol. 11, No. 4, pp. 18–32.

[25] P. Jesorsky (1978), "Principles of Automatic Speaker Recognition," in *Speech Communication with Computers* (Carl Hanser Verlag), pp. 93–137.

[26] D. H. Klatt, L. C. Klatt (1990), "Analysis, Synthesis, and Perception of Voice Quality Variations Among Female and Male Talkers," in *J. Acoust. Soc. America*, Vol. 87, No. 2, pp. 820–857.

[27] Y. Koike, J. Markel (1975), "Application of Inverse Filtering for Detecting Laryngeal Pathology," in *Annals of Otology, Rhinology and Laryngology*, Vol. 84, pp. 117–124.

[28] J. Makhoul (1975), "Linear Prediction: A Tutorial Review," *Proc. IEEE*, Vol. 63, No. 4, pp. 561–580.

[29] A. I. C. Monaghan, D. R. Ladd (1990), "Speaker-Dependent and Speaker-Independent Parameters in Intonation," *ESCA Proc. Speaker Characterisation in Speech Technology* (Edinburgh), pp. 167–174.

[30] J. Naik (1994), "Speaker Verification over the Telephone Network: Databases, Algorithms and Performance Assessment," *ESCA Workshop on Automatic Speaker Recognition, Identification and Verification* (Martigny), pp. 31–38.

[31] K. Osawa, T. Araseki (1986), "Low Bit Rate Multi-Pulse Speech Coder with Natural Speech Quality," *Proc. ICASSP* (Tokyo), pp. 457–460.

[32] M. R. Pakravan (1992), "A New Way for Implementing a Speaker Identification System," *Int. Conf. Signal Processing Applications and Technology* (Boston), pp. 1035–1041.

[33] N. B. Pinto, D. G. Childers, A. L. Lalwani (1989), "Formant Speech Synthesis: Improving Production Quality," *IEEE Trans. ASSP*, Vol. 37, No. 12, pp. 1870–1887.

[34] R. A. Prosek, A. A. Montgomery, B. E. Walden, D. B. Hawkins (1987), "An Evaluation of Residue Features as Correlates of Voice Disorders," *J. Commun. Disord.*, Vol. 20, pp. 105–117.

[35] D. A. Reynolds, R. C. Rose (1992), "An Integrated Speech-Background Model for Robust Speaker Identification," *Proc. ICASSP* (San Francisco), pp. II.185–II.188.

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

[37] A. E. Rosenberg, C.-H. Lee, S. Gokcen (1991), "Connected Word Talker Verification Using Whole Word Hidden Markov Model," *Proc. ICASSP* (Toronto), pp. 381–384.

[38] D. O'Shaughnessy (1987), *Speech Communication (Human and Machine)* (Addison-Wesley), 568 p.

[39] M. Shridar, N. Mohankrishnan (1982), "Text-Independent Speaker Recognition: A Review and Some New Results," *Speech Comm.* (San Francisco), Vol. 1, No. 3–4, pp. 257–267.

[40] J. Skvarc, M. Miletic (1990), "Speaker Sex Estimation," *ESCA Proc. Speaker Characterisation in Speech Technology* (Edinburgh), pp. 181–186.

[41] F. K. Soong, A. E. Rosenberg (1986), "On the Use of Instantaneous and Transitional Spectral Information in Speaker Recognition," *Proc. ICASSP* (Tokyo), pp. 877–880.

[42] P. Thévenaz, H. Hügli (1994), "Conformity, a New Method for Text-Independent Speaker Recognition," *ESCA Workshop on Automatic Speaker Recognition, Identification and Verification* (Martigny), pp. 63–66.

[43] J. J. Wolf (1972), "Efficient Acoustic Parameters for Speaker Recognition," *J. Acoust. Soc. America*, Vol. 51, No. 6, Part 2, pp. 2044–2056.

[44] X. Zhu, Y. Gao, S. Ran, F. Chen, I. MacLeod, B. Millar, M. Wagner (1994), "Text-Independent Speaker Recognition using VQ, mixture Gaussian VQ and Ergodic HMMs," *ESCA Workshop on Automatic Speaker Recognition, Identification and Verification* (Martigny), pp. 55–58.