Pattern Recognition ▮▮▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Investigation on LP-residual representations for speaker identification

M. Chetouani[a,*], M. Faundez-Zanuy[b], B. Gas[a], J.L. Zarader[a]

[a]Université Pierre et Marie Curie (UPMC), 4 Place Jussieu, 75252 Paris Cedex 05, France
[b]Escola Universitària Politècnica de Mataró, Barcelona, Spain

ARTICLE INFO

Article history:
Received 9 February 2007
Received in revised form 23 May 2008
Accepted 5 August 2008

Keywords:
Feature extraction
Speaker identification
LP-residue
Non-linear speech processing

ABSTRACT

Feature extraction is an essential and important step for speaker recognition systems. In this paper, we propose to improve these systems by exploiting both conventional features such as mel frequency cepstral coding (MFCC), linear predictive cepstral coding (LPCC) and non-conventional ones. The method exploits information present in the linear predictive (LP) residual signal. The features extracted from the LP-residue are then combined to the MFCC or the LPCC. We investigate two approaches termed as temporal and frequential representations. The first one consists of an auto-regressive (AR) modelling of the signal followed by a cepstral transformation in a similar way to the LPC–LPCC transformation. In order to take into account the non-linear nature of the speech signals we used two estimation methods based on second and third-order statistics. They are, respectively, termed as R-SOS-LPCC (residual plus second-order statistic based estimation of the AR model plus cepstral transformation) and R-HOS-LPCC (higher order). Concerning the frequential approach, we exploit a filter bank method called the power difference of spectra in sub-band (PDSS) which measures the spectral flatness over the sub-bands. The resulting features are named R-PDSS. The analysis of these proposed schemes are done over a speaker identification problem with two different databases. The first one is the Gaudi database and contains 49 speakers. The main interest lies in the controlled acquisition conditions: mismatch between the microphones and the interval sessions. The second database is the well-known NTIMIT corpus with 630 speakers. The performances of the features are confirmed over this larger corpus. In addition, we propose to compare traditional features and residual ones by the fusion of recognizers (feature extractor + classifier). The results show that residual features carry speaker-dependent features and the combination with the LPCC or the MFCC shows global improvements in terms of robustness under different mismatches. A comparison between the residual features under the opinion fusion framework gives us useful information about the potential of both temporal and frequential representations.

1. Introduction

During the last decades, significant efforts have been made for the design of efficient features for the improvement of speaker recognition systems. As a result, several features have been proposed. For instance, Jang et al. [1] proposed an approach based on speech signal decomposition by using the independent component analysis (ICA). It mainly consists of an optimisation of basis functions for statistical independent feature extraction. The resulting features, similar to Gabor wavelets, increase the speaker identification rate by 7.7% compared to the discrete cosine transform (DCT) for a subset of TIMIT. Following the speech production model (i.e. source–filter model), some authors attempt to extract features known as speaker-dependent such as glottal information [2]. Mary et al. [3] used the potential of auto-associative neural networks for capturing short-segment (10–30 ms) and sub-segmental (1–5 ms) features extracted from linear predictive (LP) analysis. This leads to the modelling of not only traditional spectral features but also source and phase modelling. The results on speaker identification show good performances in case of combination of these features. Despite these investigations, state-of-art systems are mostly based on the mel cepstral frequency coding (MFCC) or the linear predictive cepstral coding (LPCC). Indeed, these short-term features have proven their efficiency in terms of performances and are adapted for the Gaussian mixture models (GMMs).

In this contribution, we propose to use additional features with the traditional ones (MFCC and LPCC) for the improvement of recognition rates. These features are based on the LP-residual signal. The paper investigates different representations for the design of a useful framework for conventional speaker recognition systems. Indeed, in the case of LPCC based systems, the extraction of LP-residual features does not need too much computation.

* Corresponding author.
E-mail address: mohamed.chetouani@upmc.fr (M. Chetouani).

0031-3203/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2008.08.008

Related works on LP-residual analysis are reported in Section 2. Section 3 presents two different representations based on temporal and frequential models, respectively. The proposed representations are tested on two different databases described in Section 4. The first one is the Gaudi database [4] which allows to control the performances under different conditions: interval between the sessions and the microphones mismatch. The second one is the well-known NTIMIT corpus which has been intensely used in speaker recognition even if there is no mismatch between the sessions. Both databases are used for speaker identification. The results of the experiments are discussed in Section 5. Finally, we give conclusions and future plans for the proposed work.

2. Related works and problem

Concerning the speech production, it is generally assumed that the signals are the result of the excitation of the vocal tract. Under the framework of the LP analysis, the vocal tract is associated to the filter (linear predictive coding, LPC) and the excitation to the residual signal. The LP analysis consists in the estimation of LPC coefficients by minimising the prediction error. The predicted sample $\widehat{s}$ results from a linear combination of the p past samples [5]:

$$\widehat{s}(n) = -\sum_{k=1}^{p} a_k s(n-k) \tag{1}$$

The LPC coefficients a_k are related to the vocal tract and may also partly capture speaker-dependent information. Indeed, derived features from these coefficients, namely the LPCC, are intensely used in speaker recognition tasks. The parameter p (filter order) plays a major role in speech recognition tasks and the best scores are obtained with 12th order whereas in speaker recognition the most used order is 16.

Under the traditional LP analysis, the residual is obtained by the error between the current and the predicted samples:

$$r(n) = s(n) - \widehat{s}(n) \tag{2}$$

Theoretically, the residual is uncorrelated to the speech signal and it is related to the excitation which is speaker-dependent. These features are known as source features. However, recent works on non-linear speech processing have shown that the source-filter model is not suitable for the speech production modelling [6,7]. Different phenomena that occur during the production are non-linear and chaotic. From these investigations on non-linear processing, one can assume that there is a dependency between the speech signal and the residual.

Several investigations have been carried out to use this residual for the improvement of speaker recognition systems [3,8–12]. Thevenaz and Hügli [8] exploit the theoretical orthogonality between the filter (i.e. the LPC coefficients) model and the residue model. Their results confirm the complement nature of these representations for speaker verification. As we mentioned previously, neural networks have also been tested for the characterisation of the LP residual [3]. In Ref. [11], auto-associative neural networks are used for the characterisation of the linear residue. They show that speaker recognition systems can attain efficient rates by using only residual features.

For an efficient design, the methods should take into account the nature of the residual. In the case of an original speech signal, several investigations have been carried out [6,7,13–15]. The different phenomena (turbulence, chaos, etc.) [13] occurring during the production, mainly due to physiological reasons, cause the presence of non-linearities in the speech signals. These non-linearities have been characterised by statistical tests such as higher-order statistics and signal distribution confirm the non-linear and non-gaussian assumptions [7,16]. Consequently, several representations attempting to model the speech signals have been investigated (for more details see Ref. [17]). As far as the temporal models are concerned, we previously proposed to extend the auto-regressive (AR) model used in the LPC analysis (cf. Eq. (1)) by predictive neural networks [18,19].

Given the predicted samples $\widehat{s}(n)$, the residual r is obtained by subtracting the original signal s to the predicted one (cf. Eq. (2)). The residual should contain all the information that is not modelled by the filter (cf. Eq. (1)). The filter coefficients estimation is based on second-order analysis (i.e. covariance, auto-correlation) which cannot model non-gaussian processes. One can postulate that the residual has not only to be modelled by higher-order statistics but also by second-order statistics due to the lack of efficiency of the estimation (p order, algorithm, noise, etc.). From these considerations, several ways can be followed to model the residual. Non-linear modelling is one of the solutions used in several applications [11,20,21] due to the non-linear nature of the residual [18,22]. The results show the potential and confirm the presence of non-linearities. For instance, an interesting work done by Thyssen et al. [21] suggest the presence of non-linearities in the residual since several series of LPC analysis are required to remove all linear information from the residual. However one has to be careful with this approach because, it has been noticed by Kubin [7], adaptive methods can lead to nearly Gaussian residual signals. Other solutions can be used such as wavelet transform as in wavelet octave coefficients of residues (WOCOR) features [12].

In this contribution, we propose to exploit the fact that the residue conveys all information that are not modelled by the LPC filter (cf. Eq. (1)). Unlike to previously proposed methods mainly based on machine learning [10,20] or signal processing [12], the approach employed in this paper is based on the combination of temporal (second and higher-order statistics for AR models) and frequential (filter banks) models. These investigations aim to show the potential of residual speech signal processing for speaker recognition tasks. The features extracted from the residual can be used as complementary ones with the LPCC or even with the MFCC.

3. Proposed representations for the LP-residue

The previous sections have shown the importance of residual signals for speaker recognition tasks. The efficiency of this additional feature is totally related to a suitable representation. In this contribution, we investigate two different approaches termed as temporal and frequential ones.

3.1. Temporal approach

The temporal approach is based on an AR model of the LP-residue:

$$\widehat{r}(n) = -\sum_{k=1}^{\rho} \alpha_k r(n-k) \tag{3}$$

where r and ρ, respectively, represent the LP-residue and the filter order, respectively. To be efficient for speech applications, cepstral derived features have to be computed. The α_k coefficients are transformed into cepstral ones γ_k in a similar manner as the LPC–LPCC transformation.

For the feature extraction process, two methods are investigated: second and higher-order statistics. The first one basically consists of a LPC analysis of the residual r followed by a cepstral derivation, resulting in LPCC equivalent features. The features obtained are noted R-SOS-LPCC features in order to make a difference from the well-known LPCC features. LP cepstral models of the residue have been already tested on speaker recognition [23] leading to some improvements. In contrast to what is done in [23] where LP analysis of the residual is combined to the MFCC by a linear discriminant analysis, the residual models are considered as additional features (as the Δ

Please cite this article as: M. Chetouani, et al., Investigation on LP-residual representations for speaker identification, Pattern Recognition (2008), doi: 10.1016/j.patcog.2008.08.008

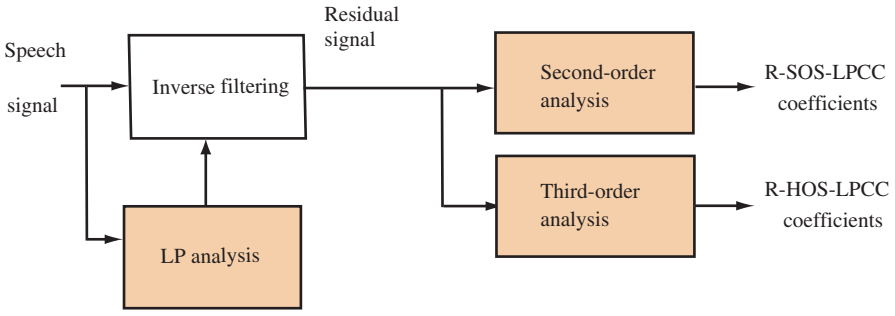

Fig. 1. Temporal processing applied to the residual signal r.

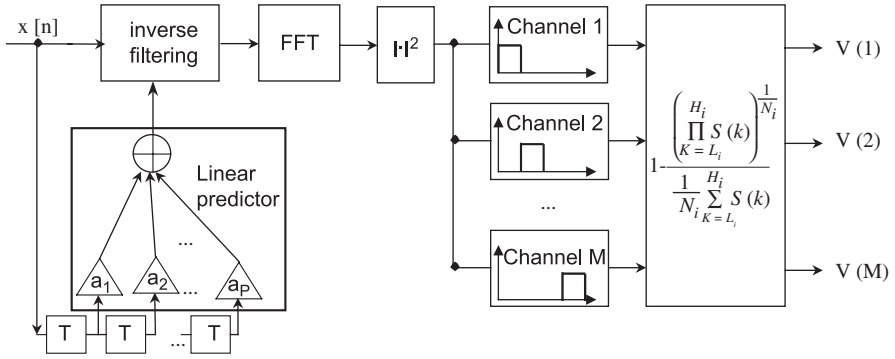

Fig. 2. Principle of the power difference of spectra in sub-band (PDSS) applied to the residual signal r.

coefficients). The next method is also based on an AR model (Eq. (3)) but with the estimation of higher-order statistics.

The traditional LPC analysis is based on second-order statistics [5] such as the covariance and auto-correlation methods. The LPC coefficients (Eq. (1)) are obtained by the resolution of the Yulke–Walker equations [5] defined as a function of the coefficients a_k and the auto-correlation R (i.e. second-order statistic). A natural extension to this procedure consists of the definition of equivalent Yulke–Walker equations but with higher-statistics such as third-order or fourth-order moments. In speech recognition, Paliwal et al. [24] applied similar ideas for the estimation of an AR model. They used a constrained third-order cumulant approach, noted C, resulting in equivalent Yulke–Walker equations:

$$\sum_{k=0}^{p} a_k C_k(i,j) = 0 \tag{4}$$

with $1 \leqslant i \leqslant p, 0 \leqslant j \leqslant i$.

The third-order cumulant of signal s is defined as

$$C_k(i,j) = \sum_{m=p+1}^{M} s_{m-k} s_{m-i} s_{m-j} \tag{5}$$

where M is the analysis window size which is equivalent to the one used for the auto-correlation in the LPC computation.

Following this formulation, a traditional recursion algorithm is used for the estimation of AR coefficients [24]. Derived cepstral features, similar procedure as the LPC–LPCC transformation, are applied to noisy speech recognition. The results show that at low SNR (20 dB) the cumulant estimation outperforms the auto-correlation one but it is not the case for higher SNRs.

In this contribution, we use similar models but, rather than applying them to the signal s (Eq. (1)), we apply them to the residual r (Eq. (3)). They are named R-HOS-LPCC. Fig. 1 represents the temporal analyses compared in this paper.

3.2. Frequential approach

Unlike the previous approach, in this section, we describe frequential processing of the residual signal r (Eq. (2)). This approach was originally proposed by Hayakawa et al. [25] and was called the power difference of spectra in sub-band (PDSS). They tested it on a speaker identification problem. The R-PDSS features gave a rate of 66.9% and the combination with LPCC features gave 99% (99.8% for the LPCC alone).

The R-PDSS features are obtained by the following steps (cf. Fig. 2):

- Calculate the LP-residual r.
- Fast Fourier transform of the residual using zero padding in order to increase the frequency resolution: $S = |fft(residue)|^2$.
- Group the power spectrum into M sub-bands.
- Calculate the ratio of the geometric to the arithmetic mean of the power spectrum of the ith sub-band and subtract from 1:

$$R - PDSS(i) = 1 - \frac{(\prod_{k=L_i}^{H_i} S(k))^{1/N_i}}{1/N_i \sum_{k=L_i}^{H_i} S(k)} \tag{6}$$

where $N_i = H_i - L_i + 1$ is the number of frequency samples in the ith sub-band. L_i and H_i are, respectively, the lower and upper frequency limits of the ith sub-band. The same bandwidth is used for all the sub-bands.

Cepstrum analysis of the residual has been also investigated in speech recognition [26]: filter bank analysis of the one-sided auto-correlation of the residual r plus a cepstral transformation. The features obtained named as residual cepstrum (RCEP) present some

Please cite this article as: M. Chetouani, et al., Investigation on LP-residual representations for speaker identification, Pattern Recognition (2008), doi: 10.1016/j.patcog.2008.08.008

linguistic information and in combination to the LPCC, improves the recognition rates. This result and the previous arguments (cf. Section 2) concerning the source-filter model are interesting because they prove that linguistic and speaker information are present in both the features: LPCC and residual. The rest of this contribution is dedicated to the experiments and the discussion on the proposed features.

4. Experimental conditions

This section is dedicated to the description of the used corpus and the different tasks that we addressed for the evaluation of the proposed feature extraction schemes. These features are obviously compared to the most used methods such as the MFCC and the LPCC. The dimension of feature vectors is set to 16 for both the traditional and residual ones (cf. Section 3, $\rho = M = 16$).

4.1. Databases

4.1.1. Gaudi

The Gaudi database [4,27] was originally designed in order to measure the performances under different controlled conditions: language, interval session and microphone. The corpus is composed of:

- 49 speakers;
- four sessions with different tasks: isolated numbers, connected numbers, text reading, conversational speech, etc.);
- for each session, the utterances have been acquired in two languages (Catalan and Spanish) and simultaneously with different microphones as described in Table 1.

In this contribution, the training protocol consists of using one text reading of an average duration of 1 min using session 1 and MIC1. Consequently, the training session is always done with M1. Concerning the tests, we use nine phonologically balanced utterances (Spanish) identical for all the speakers through the sessions (3–5 s): M1–M6. We focus on the first three sessions with different microphones (cf. Table 2). The number of tests is $49 \times 9 = 441$ for each session and the average score is estimated on $49 \times 9 \times 6 = 2646$ tests.

The speech signal has been down-sampled to 8 kHz (producing a telephonic bandwidth), pre-emphasised by a first-order filter whose transfer function is $H(z)=1-0.95z^{-1}$ and normalised between $-1, +1$ (for cumulant estimation). A 30 ms Hamming window is used, and the overlapping between adjacent frames is $\frac{2}{3}$. A parameterised vector of 16th order was computed for each feature extraction method.

Table 1
The microphones used for the Gaudi database

MIC1	SONY ECM 66B	Lapel unidirectional electret (≈ 10 cm from the speaker)
MIC2	AKG D40S	Dynamic cardoid (≈ 30 cm from the speaker)
MIC3	AKG C420	Head-mounted (low-cost microphone)

Table 2
Different sessions and microphones

Ref.	Session	Microphone
M1	1	MIC1
M2	1	MIC2
M3	2	MIC1
M4	2	MIC2
M5	3	MIC1
M6	3	MIC3

4.1.2. NTIMIT

The NTIMIT database [28] is a telephonic version of the TIMIT corpus including local and long distance calls. The database contains 630 speakers (438 male and 192 female) and each of them have uttered 10 sentences:

- Two different sentences SA1 and SA2. They are the same across the 630 speakers and they have an average duration of 2.9 s.
- Eight sentences different across the speakers: three SI (average duration of 2.9 s) and five SX sentences (average duration of 3.2 s).

Contrary to the Gaudi database (cf. Section 4.1.1), NTIMIT contains only single session recordings with a fixed handset. However this database has been largely used for speaker recognition applications [29–31]. In spite of these successful applications, results on this database are useful because they can be compared easily. Lot of training and test protocols have been defined for NTIMIT [29–31]. In this article, we use the protocol called "long training–short test" initially proposed by Bimbot et al. [29] which consists of:

- "long training": the five SX sentences are concatenated as a single reference pattern for each speaker. As a result, the "long training" pattern average duration is 14.4 s.
- "short test": SA and SI sentences are tested separately resulting in $630 \times 5 = 3150$ tests (with an average duration of 3.2 s).

The training duration is less than the one used for the Gaudi protocol but with more utterances for test and consequently the obtained results have a higher statistical significance [29].

The speech signal is recorded through a high quality microphone and is sampled at 16 kHz but with a bandwidth of 300–3400 Hz (telephone bandwidth). A 31.5 ms Hamming window is used at a frame rate of 10 ms.

4.2. Speaker identification method

For the evaluation of feature schemes, we test them on the speaker identification problem using both the databases: Gaudi (cf. Section 4.1.1) and NTIMIT (cf. Section 4.1.2).

The speaker models have been designed by a simple second-order statistic method. A covariance matrix (C) is computed for each speaker and the arithmetic-harmonic sphericity measure [32] is applied for comparison:

$$\mu(\mathbf{C}_j, \mathbf{C}_{test}) = \log(\mathrm{tr}(\mathbf{C}_{test}\mathbf{C}_j^{-1})\,\mathrm{tr}(\mathbf{C}_j\mathbf{C}_{test}^{-1})) - 2\,\log(P) \quad (7)$$

where tr is the trace of the matrix, P is the dimension of feature vector ($P = 16$). The number of parameters for each speaker model is $P^2 + P/2$ (the covariance matrix is symmetric).

5. Results and discussions

5.1. Mismatch identification

Mismatch conditions due to acquisition or interval sessions seriously decrease the recognition rates of speaker recognition systems. As previously described in the experimental section (cf. Section 4.1.1), the Gaudi database is used for speaker identification under controlled conditions.

Table 3 presents the speaker identification rates for the different conditions. Baseline results are represented by both the MFCC and the LPCC features. For no mismatch, training and test on M1, best results are achieved by the MFCC. However, if we add to the LPCC features residual information (R-SOS-LPCC, R-HOS-LPCC and R-PDSS), improvements are obtained but the number of features is

Please cite this article as: M. Chetouani, et al., Investigation on LP-residual representations for speaker identification, Pattern Recognition (2008), doi: 10.1016/j.patcog.2008.08.008

Table 3
Correct speaker identification rates for mismatch training (with M1) and test for temporal, frequential and mixed methods

	Feature extraction	M1	M2	M3	M4	M5	M6	Average
Temporal	LPCC	94.78	73.7	74.60	66.213	55.33	52.15	69.46
	R-SOS-LPCC	87.98	63.72	60.32	59.18	44.45	43.99	59.94
	R-HOS-LPCC	83.45	55.33	57.14	50.79	42.40	33.10	53.70
	LPCC + R-SOS-LPCC	97.5	81.86	79.82	71.43	56.92	62.81	75.05
	LPCC + R-HOS-LPCC	97.96	80.04	80.04	70.521	58.05	59.64	74.37
Frequential	MFCC	97.50	76.64	78.23	72.34	57.59	62.36	74.11
	R-PDSS	82.09	59.86	62.36	60.99	45.35	42.18	58.80
Mixed	LPCC + R-PDSS	99.77	82.54	85.26	83.22	66.43	67.35	80.76

also increased from $P(16)$ to $2 \times P(32)$ resulting in more computation. Looking to the performances of the residual information, temporal and frequential representations (cf. Section 3) alone give non negligible results: more than 80% of correct speaker identification.

Concerning the mismatch conditions, as it can be expected, for all the features the identification rates decrease. However, the loss of performances differs according to the mismatch: interval session and/or microphone (cf. Section 4.1.1). The impact of the acquisition is more important than the interval session impact. When the microphone changes for the same session, for instance M2, the performances are degraded and the rates are more or less equivalent to the interval session mismatch with the same microphone M3 (cf. Table 3). For conventional features, MFCC features give the best results for these different conditions. The speaker-dependent information contained in the residual are also non negligible even if the conditions differ seriously. Moreover, when the residual features are added to the LPCC as complementary features, the robustness under the different mismatches is clearly improved resulting in better identification rates than the LPCC or the MFCC alone.

The tests carried for M5 and M6 mismatches (long interval, microphone, cf. Section 4.1.1) are mostly equivalent for all the features (cf. Table 3). These tests are interesting because they give information about the robustness of the features for real applications. However, for all the features except MFCC, the performance slightly decreases. Once again, the robustness is improved by using residual information and the conventional LPCC resulting in at least equivalent MFCC results or even better.

In order to compare the performances of these features under the different mismatches, we compute the average speaker identification rate for each feature (cf. Table 3) through the conditions M1–M6 and they are presented in Table 3. For conventional features, the best results are achieved by the MFCC. For the residual information, the performances of the R-SOS-LPCC and R-PDSS are mostly equivalent and are better than the higher-order statistic based features namely the R-HOS-LPCC. As it has been previously mentioned (cf. Section 2), after an LPC analysis, linear information are still present in the LP-residue.

Concerning the additional features, the LPCC plus the residual information improve the recognition rates. The average performances show that temporal methods are mostly equivalent. This means that linear and non-linear information, respectively, modelled by R-SOS-LPCC and R-HOS-LPCC, carry speaker-dependent information and are complementary to the LPCC. It can be explained by two main remarks:

- Due to the imperfect LPC analysis, the LP-residue still carries Gaussian information modelled by the R-SOS-LPCC and features (cf. Section 2).
- The R-HOS-LPCC model allows to model non-gaussian distributions but it is limited by the fact that it is only a third-order based model (cf. Section 3).

Table 4
Correct speaker identification rates for the NTIMIT database

MFCC	27.3
LPCC	24.6
R-SOS-LPCC	8.22
R-HOS-LPCC	5.08
R-PDSS	8.73

In order to overcome these limitations, non-linear models have been directly applied to the speech signal such as predictive neural networks [9,19] resulting in improvements of the speaker identification rates. Those models are inspired by the LPC analysis since they are a direct extension of them. For instance in the neural predictive coding (NPC) scheme [19], the neural weights are used as features. Furthermore, this model can be initialised by the LPC analysis.

5.2. Large database

The previous Gaudi database shows that residual information carries speaker-dependent information and it is true for all types of models (temporal or frequential). In this section, we propose to confirm these results by doing training and test on a larger database such as the NTIMIT (cf. Section 4.1.2).

Table 4 presents the speaker identification rates for the whole NTIMIT database (630 speakers) with respect to the feature extraction methods. Baseline results (MFCC and LPCC) are the best ones and are more-or-less equivalent, for the same "long training–short test", to the results obtained in Ref. [29]. One can notice that with a different protocol or classifier (i.e. GMMs, support vector machines), better results can be expected as noted in Refs. [29–31].

The results of the residual models for the whole NTMIT database confirm the presence of speaker-dependent information but as it can be expected that they are worse than the traditional features (MFCC and LPCC). Concerning the temporal models, the linear model R-SOS-LPCC gives the best results. We previously noticed similar behaviour which can be justified by the lack of efficiency of the LPCC analysis and the used non-gaussian model based on third-order statistics (cf. Section 5.1). The speaker identification rates given by the R-PDSS method are higher than for the temporal representations.

For the Gaudi database (cf. Section 5.1), we show that residual models can be used as complementary features for a global improvement of the recognition rates. Rather than doing that, we propose, in the next section, the fusion of these features in order to evaluate this complementarity.

5.3. Opinion fusion

Information fusion is an important and effective stage for global improvements of the recognition rates. In this subsection, our purpose is to evaluate and to compare the features. We combine the

Please cite this article as: M. Chetouani, et al., Investigation on LP-residual representations for speaker identification, Pattern Recognition (2008), doi: 10.1016/j.patcog.2008.08.008

Table 5
Experimental results for different combinations (temporal)

Feature extraction		Temporal		Frequential	
		R-SOS-LPCC	R-HOS-LPCC	R-PDSS	MFCC
Temporal	LPCC	28.06	26.19	28.09	31.75
	R-SOS-LPCC		12.54	14.92	34.98
	R-HOS-LPCC			12.06	33.33

Table 6
Experimental results for different combinations (frequential)

Feature extraction		Temporal			Frequential
		LPCC	R-SOS-LPCC	R-HOS-LPCC	R-PDSS
Frequential	MFCC	31.75	34.98	33.33	33.33
	R-PDSS	28.09	14.92	12.06	

Table 7
Selected combination factor α for the results shown in Tables 5 and 6

	LPCC	R-SOS-LPCC	R-HOS-LPCC	R-PDSS
MFCC	0.91	0.83	0.72	0.66
LPCC		0.57	0.58	0.51
R-SOS-LPCC			0.28	0.29
R-HOS-LPCC				0.46

The indicated factors give the best scores (following Eq. (9)).

output of the recognizers (i.e. covariance matrix cf. Section 4.2) for all the features (i.e. conventional and non-conventional ones). This scheme is known as opinion fusion [33,34].

The opinion fusion procedure mainly consists in the following steps:

(1) Distance normalisation [35]:

$$o_i' = \frac{1}{1+e^{-k_i}} \tag{8}$$

with $k = o_i - (m_i - 2\sigma_i)/2\sigma_i$. o_i is the opinion of the classifier i. $o_i' \in [0,1]$ is the normalised opinion, m_i, σ_i are the mean and the standard deviation of the opinions of classifier i using the genuine speakers (intra-distances).

(2) Weighted sum combination with trained rule [34,35]:

$$O = \alpha o_1 + (1-\alpha) o_2 \tag{9}$$

where o_1, o_2 are scores (distances) provided by each classifier. α is a weighting or combination factor. A high value of α implies a high importance of recognizer 1 (feature extractor plus classifier).

The fusion scores with the different features are presented in Tables 5 and 6 and the scores without fusion have been reported in Table 4. One can expect that the fusion of the best scores such as the MFCC and LPCC should give the best results. But, in Tables 5 and 6, the best scores are obtained by the MFCC/R-SOS-LPCC couple and moreover, the fusion of the MFCC and all the residual features are better than the MFCC–LPCC fusion. This result shows that the combination of MFCC and residual features is efficient for a global improvement. The combination factor α gives useful information about the contribution of each method (cf. Table 7). Even if the R-SOS-LPCC gives better scores, the MFCC contribution ($\alpha = 0.83$) is higher than the other ones R-HOS-LPCC ($\alpha = 0.72$) and R-PDSS ($\alpha = 0.66$). One can also notice that the robustness of the LPCC (cf. Table 4) is clearly improved by the proposed schemes (cf. Tables 5 and 6). Regarding the combination factors (cf. Table 7), the contributions of both LPCC and residual features are mostly of the same orders.

Concerning the fusion of residual features between them, it allows improvements but the attained scores are clearly less than the MFCC and LPCC alone (cf. Tables 4–6). However, these experiments have also been carried out in order to compare the residual models between them. R-SOS-LPCC/R-HOS-LPCC fusion is interesting because it compares two predictive models based on second and third-order statistics, respectively (cf. Section 3). For a combination factor of $\alpha = 0.28$, it seems that the second-order statistic based model (i.e R-SOS-LPCC) carries less speaker-dependent information than the third-order one (R-HOS-LPCC) which seems to be in contradiction with the results obtained in Table 4. However, R-SOS-LPCC/R-PDSS fusion gives better results with a same behaviour which means that the speaker-dependent information is not present in the similar way in all the features. These results show that the exploitation of the complementarity between the features can be improved by suitable representations.

Finally for temporal/frequential fusion, the best scores are obtained with a small contribution of the R-SOS-LPCC. A more important contribution by the R-HOS-LPCC is needed (cf. Table 7) but a worse score is obtained for the temporal/frequential fusion.

The results obtained using fusion show that the performances and the robustness of the traditional features (MFCC and LPCC) are improved by the residual ones. And, as one can expect, the contribution of the conventional features are higher than the residual ones. Concerning the combination of residual features, the best scores are obtained by the fusion of the second-order model (R-SOS-LPCC) and the frequential one (R-PDSS).

6. Conclusions

In this paper, we proposed to extract features from the LP-residual for the improvement of speaker identification systems. Several models have been investigated based on temporal and frequential approaches. The temporal models are based on an auto-regressive (AR) filter and the coefficients of this model are estimated by second (SOS) or higher-order (HOSs) statistics. The SOS based model is obtained by the application of a traditional LPC analysis to the residue followed by a cepstral transformation of the LPC coefficients. The resulting features are termed R-SOS-LPCC features. Following the same scheme and the recent works on non-linear speech processing, we proposed to use higher-order statistics for the improvement of the modelling resulting in features called R-HOS-LPCC features. Concerning the frequential approach, a filter bank is investigated termed as the power difference of spectra in sub-band (PDSS) which can be interpreted

Please cite this article as: M. Chetouani, et al., Investigation on LP-residual representations for speaker identification, Pattern Recognition (2008), doi: 10.1016/j.patcog.2008.08.008

as a sub-band version of the spectral flatness measure. The key idea is to extract frequential information from the LP-residue.

These temporal and frequential approaches are evaluated in a speaker identification task. Firstly, we evaluated the robustness of the features (R-SOS-LPCC, R-HOS-LPCC and R-PDSS) with controlled conditions: interval between the sessions, microphones. The obtained results show that residual information improve the speaker identification scores (at least 7% better than the LPCC alone). The R-HOS-LPCC features give worse results than the R-SOS-LPCC and it has been partly justified by the presence of linear information in the LP-residue and the modelling limitation of the R-HOS-LPCC (third-order based). The best speaker identification rates have been attained by the combination of the LPCC and the R-PDSS features. Secondly, the different features have been tested on the well-known NTIMIT database following the "long training–short test" protocol. The results on this larger corpus confirm that the LP-residue carries speaker-dependent information. In order to evaluate the potential of the residual features for the global improvement of speaker recognition systems, we proposed to compare the recognizers (feature extractor + classifier) by the opinion fusion framework. Once again the robustness of the LPCC is clearly improved by the combination with residual features. And we can notice that the residual features can also be used with the MFCC, which initially gives best scores alone, for a global improvement. We also focused on the fusion of the residual features between them in order to evaluate their respective performances showing that temporal (R-SOS-LPCC and R-HOS-LPCC) and frequential (R-PDSS) features convey complementary information due to the different extraction schemes: AR model and bank filter.

This investigation on LP-residue gives us useful information about the properties of the signal. Clearly, speaker-dependent information are present and they have to be used with conventional features such as the MFCC or the LPCC. Moreover, the robustness over the recognition conditions (interval sessions, microphones and telephone) is improved. However, one can notice that this last point can be significantly improved by the use of robust methods such as cepstral mean subtraction (CMS). Concerning the future works, the limitation of the R-HOS-LPCC model mainly due to its estimation (third-order statistic) should be investigated. It can be done by the use of more higher-orders (i.e. fourth) or an association of them. It can also be done by non-linear models such as neural networks such as the NPC scheme [36]. Furthermore, in this contribution, we used the LP-residue but other strategies can be followed as the analysis of the NLP-residue (non-linear) as done in Ref. [9].

Acknowledgements

A part of this research was carried out during a visit at the Escola Universitària Politècnica de Mataró, Barcelona, Spain, and was funded by the European COST action. This work has been supported by FEDER and MEC, TEC2006-13141-C03-02/TCM.

References

[1] G.J. Jang, T.L. Lee, Y.H. Oh, Learning statistically efficient features for speaker recognition, Neurocomputing 49 (2002) 329–348.
[2] R.E. Slyh, E.G. Hansen, T.R. Anderson, Glottal modeling and closed-phase analysis for speaker recognition, in: Proceedings of the ISCA Tutorial and Research Workshop on Speaker and Language Recognition (Odyssey'04), 2004, pp. 315–322.
[3] L. Mary, K. Sri Rama Murty, S.R. Mahadeva Prasanna, B. Yegnanaraya, Features for speaker and language identification, in: Proceedings of the ISCA Tutorial and Research Workshop on Speaker and Language Recognition (Odyssey'04), 2004, pp. 323–328.
[4] J. Ortega, et al., Ahumada: a large speech corpus in Spanish for speaker identification and verification, in: Proceedings of the IEEE ICASSP'98, vol. 2, 1998, pp. 773–775.
[5] B.S. Atal, S.L. Hanauer, Speech analysis and synthesis by linear prediction of speech wave, J. Acoust. Soc. Am. 50 (1971) 637–655.
[6] M. Faundez-Zanuy, G. Kubin, W.B. Kleijn, P. Maragos, S. McLaughlin, A. Esposito, A. Hussain, J. Schoentgen, Nonlinear speech processing: overview and applications, Control Intelligent Syst. 30 (1) (2002) 1–10.
[7] G. Kubin, Nonlinear processing of speech, in: W.B. Kleijn, K.K. Paliwal (Eds.), Speech Coding and Synthesis, 1995, pp. 557–610.
[8] P. Thevenaz, H. Hügli, Usefulness of the LPC-residue in text-independent speaker verification, Speech Commun. 17 (1–2) (1995) 145–157.
[9] M. Faundez, D. Rodriguez, Speaker recognition using residual signal of linear and nonlinear prediction models, ICSLP 2 (1998) 121–124.
[10] B. Yegnanaraya, K.S. Reddy, S.P. Kishore, Source and system features for speaker recognition using AANN models, in: Proceedings of the IEEE ICASSP, 2001, pp. 409–412.
[11] S.R. Mahadeva Prasanna, C.S. Gupta, B. Yegnanaraya, Extraction of speaker-specific excitation from linear prediction residual of speech, Speech Commun. 48 (2006) 1243–1261.
[12] N. Zheng, T. Lee, P.C. Ching, Integration of complementary acoustic features for speaker recognition, IEEE Signal Process. Lett., 2006.
[13] A. Esposito, M. Marinaro, Some notes on nonlinearities of speech, in: G. Chollet, et al. (Eds.), Nonlinear Speech Modeling, Lecture Notes in Artificial Intelligence, vol. 3445, 2005, pp. 1–4.
[14] S. McLaughlin, S. Hovell, A. Lowry, Identification of nonlinearities in vowel generation, in: Proceedings of the EUSIPCO, 1988, pp. 1133–1136.
[15] H. Teager, S. Teager, Evidence for nonlinear sound production mechanisms in the vocal tract, in: Proceedings of the NATO ASI on Speech Production and Speech Modeling, vol. II, 1989, pp. 241–261.
[16] S. Gazor, W. Zhang, Speech probability distribution, IEEE Signal Process. Lett. 10 (7) (2003) 204–207.
[17] G. Chollet, A. Esposito, M. Faundez-Zanuy, M. Marinaro, Nonlinear speech modeling and applications, in: Lecture Notes in Artificial Intelligence, vol. 3445, 2005.
[18] M. Faundez, D. Rodriguez, Speaker recognition by means of a combination of linear and nonlinear predictive models, in: Proceedings of the IEEE ICASSP'99, 1999.
[19] M. Chetouani, M. Faundez-Zanuy, B. Gas, J.L. Zarader, A new nonlinear speaker parameterization algorithm for speaker identification, in: Proceedings of the ISCA Tutorial and Research Workshop on Speaker and Language Recognition (Odyssey'04), 2004, pp. 309–314.
[20] E. Rank, G. Kubin, Nonlinear synthesis of vowels in the LP residual domain with a regularized RBF network, in: Proceedings of the IWANN, vol. 2085(II), 2001, pp. 746–753.
[21] J. Thyssen, H. Nielsen, S.D. Hansen, Non-linearities short-term prediction in speech coding, in: Proceedings of the IEEE ICASSP'94, vol. 1, 1994, pp. 185–188.
[22] C. Tao, J. Mu, X. Xu, G. Du, Chaotic characteristics of speech signal and its LPC residual, Acoust. Sci. Technol. 25 (1) (2004) 50–53.
[23] S.H. Chen, H.C. Wang, Improvement of speaker recognition by combining residual and prosodic features with acoustic features, in: Proceedings of the IEEE ICASSP'04, vol. 1, 2004, pp. 93–96.
[24] K.K. Paliwal, M.M. Sondhi, Recognition of noisy speech using cumulant-based linear prediction analysis, in: Proceedings of the IEEE ICASSP'91, vol. 1, 1991, pp. 429–432.
[25] S. Hayakawa, K. Takeda, F. Itakura, Speaker identification using harmonic structure of LP-residual spectrum, in: Audio Video Biometric Personal Authentification, Lecture Notes in Computer Science, vol. 1206, Springer, Berlin, 1997, pp. 253–260.
[26] J. He, L. Liu, G. Palm, On the use of residual cepstrum in speech recognition, in: Proceedings of the IEEE ICASSP'96, vol. 1, 1991, pp. 5–8.
[27] A. Satue-Villar, M. Faundez-Zanuy, On the relevance of language in speaker recognition, in: Proceedings of the EUROSPEECH'99, vol. 3, 1999, pp. 1231–1234.
[28] C. Jankowski, A. Kalyanswamy, S. Basson, J. Spitz, NTIMIT: a phonetically balanced, continuous speech, telephone bandwidth speech database, in: Proceedings of the IEEE ICASSP, vol. 1, 1990, pp. 109–112.
[29] F. Bimbot, I. Magrin-Chagnolleau, L. Mathan, Second-order statistical measures for text-independent speaker identification, Speech Commun. 17 (1995) 177–192.
[30] D.A. Reynolds, Speaker identification and verification using Gaussian mixture speaker models, Speech Commun. 17 (1995) 91–108.
[31] L. Besacier, J.F. Bonastre, Subband architecture for automatic speaker recognition, Signal Process. 80 (2000) 1245–1259.
[32] F. Bimbot, L. Mathan, Text-free speaker recognition using an arithmetic-harmonic sphericity measure, in: Proceedings of the EUROSPEECH'91, 1999, pp. 169–172.
[33] J. Kittler, M. Hatef, R.P.W. Duin, J. Matas, On combining classifiers, IEEE Trans. Pattern Anal. Mach. Intell. 20 (3) (1998) 226–239.
[34] M. Faundez-Zanuy, Data fusion in biometrics, IEEE Aerosp. Electron. Syst. Mag. 20 (1) (2005) 34–38.
[35] C. Sanderson, Information fusion and person verification using speech and face information, IDIAP Research Report 02-33, 1–37, September 2002.
[36] M. Chetouani, M. Faundez-Zanuy, B. Gas, J.L. Zarader, Non-linear speech feature extraction for phoneme classification and speaker recognition, in: G. Chollet et al. (Eds.), Nonlinear Speech Modeling, Lecture Notes in Artificial Intelligence, vol. 3445, 2005, pp. 344–350.

Please cite this article as: M. Chetouani, et al., Investigation on LP-residual representations for speaker identification, Pattern Recognition (2008), doi: 10.1016/j.patcog.2008.08.008

About the Author—M. CHETOUANI received the M.S. degree in Robotics and Intelligent Systems from the University Pierre and Marie Curie (UPMC), Paris, 2001. He received the Ph.D. degree in Speech Signal Processing from the same university in 2004. In 2005, he was an invited Visiting Research Fellow at the Department of Computer Science and Mathematics of the University of Stirling, UK. He was also an invited researcher at the Signal Processing Group of Escola Universitaria Politecnica de Mataro, Barcelona, Spain. He is currently an Associate Professor in Signal Processing and Pattern Recognition at the UPMC. His research activities, carried out at the Institute of Intelligent Systems and Robotics, cover the areas of non-linear speech processing, feature extraction and pattern classification for speech, speaker and language recognition. He is a member of different scientific societies (ISCA, AFCP and ISIS). He has also served as chairman, reviewer and member of scientific committees of several journals, conferences and workshops.

Please cite this article as: M. Chetouani, et al., Investigation on LP-residual representations for speaker identification, Pattern Recognition (2008), doi: 10.1016/j.patcog.2008.08.008