ON THE USEFULNESS OF LINEAR AND NONLINEAR PREDICTION RESIDUAL SIGNALS FOR SPEAKER RECOGNITION[1]

Marcos Faundez-Zanuy

Escola Universitaria Politécnica de Mataró, UPC Barcelona, SPAIN

ABSTRACT

This paper compares the identification rates of a speaker recognition system using several parameterizations, with special emphasis on the residual signal obtained from linear and nonlinear predictive analysis. It is found that the residual signal is still useful even when using a high dimensional linear predictive analysis. On the other hand, it is shown that the residual signal of a nonlinear analysis contains less useful information, even for a prediction order of 10, than the linear residual signal. This shows the inability of the linear models to cope with nonlinear dependences present in speech signals, which are useful for recognition purposes.

Index Terms— Neural networks, speaker recognition, nonlinearities, prediction methods

1. INTRODUCTION

Several parameterization techniques exist for speech [17] and speaker [15] recognition, cepstral analysis and its related parameterizations such as Delta-Cepstral features, Cepstral Mean Subtraction, etc. being the most popular.

There are two main ways to compute the cepstral coefficients and one important drawback in both cases: relevant information is discarded, as follows.

1. LP-derived cepstral coefficients. The linear prediction analysis produces two main components, the prediction coefficients (synthesis filter) and the residue of the predictive analysis. This latter signal is usually discarded. However, experiments exist [9] where it is shown that human beings are able to recognize the identity of the speaker listening to residual signals of LP analysis. Based on this fact several authors have evaluated the usefulness of the LPC-residue and have found that although the identification rates using this kind of information alone does not perform as well as the LP-derived cepstral coefficients, a combination of both can improve the results [20,12,14,22,11].
2. Fourier Transform derived cepstral coefficients. Instead of working out a set of Linear prediction coefficients, are based on the power spectrum information, where phase information has been discarded. [19] proposed the use of new acoustic features based on the short-term Fourier phase spectrum. The results are similar to the LP-derived cepstral coefficients. Although these (phase spectrum) features cannot outperform the classical cepstral parameterization, the results are improved using a combination of both features.

In this paper we will focus on the first kind of parameterization, because they are a clear alternative to the nonlinear predictive models, which have shown an improvement over the classical linear techniques in several fields (for a recent overview about these techniques [7]).

In [4,6] we proposed a new set of features and models based on these types of nonlinear models and an improvement was also found when this information was combined with the traditional cepstral analysis, but so far, the relevance of the residual signals from linear and nonlinear predictive analysis has not been studied and compared.

In this paper we will study if the relevance of the residual signal is due to an insufficient linear predictive analysis order or because of the incapability of the linear analysis to model nonlinearities present in speech and demonstrate is usefulness for speaker recognition purposes. This important question has not been solved in previous papers that focus on a typical 8 to 16 prediction order.

2. EXPERIMENT SET UP

2.1. Database

For our experiments we have used the Gaudi database [16]. We have used one subcorpora of 49 speakers acquired with a simultaneous stereo recording with two different microphones. The speech is in wav format with a sampling frequency (f_s) = 16 kHz, 16 bit/sample and the bandwidth is 8 kHz.

From this database we have generated narrow-band signals using the potsband routine that can be downloaded from [21]. This function meets the specifications of G.151 for any sampling frequency. Thus, our study has been performed on telephone bandwidth.

2.2. Identification algorithm

In this study, we are only interested in the relative performance between linear and nonlinear analyses. Thus, we have chosen a simple algorithm for speaker recognition.

In the training phase, we compute, for each speaker, empirical covariance matrices based on feature vectors extracted from overlapped short time segments of the speech signals. As features representing short time spectra we use both linear prediction cepstral coefficients (LPCC) and mel-frequency cepstral coefficients melceps [3]. In the speaker-recognition system, the trained covariance matrices for each speaker are compared with an estimate of the covariance matrix obtained from a test sequence from a speaker. An arithmetic-harmonic sphericity measure is used in order to compare the matrices [1]:

[1] This work has been supported by FEDER and MEC TIC-2003-08382-C05-02, TEC2006-13141-C03-02/TCM

$d = \log\left(\mathrm{tr}(C_{test} C_j^{-1})\ \mathrm{tr}(C_j C_{test}^{-1})\right) - 2\log(l)$, where $\mathrm{tr}(\cdot)$ denotes the trace operator, *l* is the dimension of the feature vector, C_{test} and C_j is the covariance estimate from the test speaker and speaker model *j*, respectively.

2.3. Parameterizations

We have used the following parameterizations

1. LP-derived cepstral coefficients (LPCC)
2. Fourier transform derived cepstral coefficients (melceps)
3. LP- residue coefficients

The first two first parameterizations can be found, for instance, in [17,15,3], while the third is proposed in [11] and will be described in more detail next.

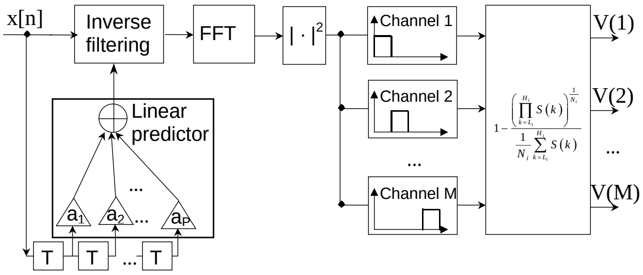

Figure 1. LP residual signal parameterization

Feature extraction from the LP-residual signal

We will use the Power Difference of Spectrum in Subband (PDSS) obtained as follows [11]:

1. Calculate the LP-residual signal using the Pth-order linear prediction coefficients.
2. Calculate the Fast Fourier Transform (fft) of the LP-residual signal using zero padding in order to increase the frequency resolution: $S = \left| fft\left(residue\right)\right|^2$
3. Group power spectrum into P subbands.
4. Calculate the ratio of the geometric to the arithmetic mean of the power spectrum in the ith subband, and subtract it from 1:

$$PDSS(i) = 1 - \frac{\left(\prod_{k=L_i}^{H_i} S(k)\right)^{\frac{1}{N_i}}}{\frac{1}{N_i}\sum_{k=L_i}^{H_i} S(k)},$$ where $N_i = H_i - L_i + 1$ is the sample number of frequency points in the ith subband and L_i, H_i is the lower and upper limit of frequency in ith subband respectively. We have used the same bandwidth for all the bands.

PDSS can be interpreted as the subband version of spectral flatness measure for quantifying the flatness of the signal spectrum. Figure 1 summarizes the procedure.

3. NEW POSSIBILITIES USING NON-LINEAR PREDICTIVE ANALYSIS

Although the relevance of residual NL-predictive analysis for speaker recognition has not been studied previously, nonlinear predictive analysis has been widely studied in the context of speech coding. For instance, [5] revealed that a forward ADPCM scheme with nonlinear prediction can achieve the same Segmental Signal to Noise Ratio (SEGSNR) as the equivalent linear predictive system (same prediction order) with one less quantization bit.

We propose an analogous scheme replacing the linear predictor with a nonlinear predictor. Figure 2 shows the scheme.

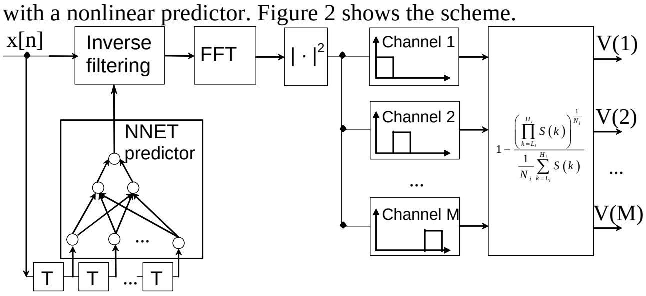

Figure 2: Block diagram used to calculate PDSS parameters from NL-prediction residual signal.

We have used a Multi-Layer Perceptron (MLP). The structure of the neural net has 10 inputs, 2 neurons in the hidden layer, and one output. The selected training algorithm was the Levenberg-Marquardt [10]. The number of epochs has been set up to 6. First layer and hidden layer transfer functions are tansig, while the output layer is linear.

4. EXPERIMENTAL RESULTS

Obviously one important question when dealing with residual LP signals is: Is the information contained in this residual signal coming from an insufficient predictive analysis order? That is, what happens when the prediction analysis order is so high that it is not possible to extract more relevant information using a linear analysis?

The experimental approach used to solve this question is to use a number of LP coefficients higher than usual. Two possible results can be obtained:

1. When the analysis order is increased, the discriminative power of the residual signal is reduced to simple chance results. This means that there is potential for speaker recognition rate improvements through extraction of the LP coefficients in a more efficient manner, probably by increasing the number of coefficients.
2. When the analysis order is increased, the residual signal still contains useful information. This means that a linear analysis is unable to extract this information, and there is room for improvement combining parameterizations defined on the LP coefficients and the residual signal. In order to obtain the optimal results, both signals should be extracted and optimized jointly.

Figure 3 shows the results obtained with the following parameterizations: Melcepstrum, LPC –P residue, LPCC, LPC-80 residue, MLP 10x2x1 and several combinations between them.

LPC-P residue is the parameterization obtained from the residual P-analysis order.

It is interesting to observe the following:

- The residual signal of an LPC-80 analysis can produce a recognition rate higher than 80% for a 15 dimensional vector extraction. Thus, it was found that the residual signal of a LP analysis contains relevant information, and this is due to the inability to extract this information using a linear analysis (80th order analysis is enough to model short term and long term dependencies between samples, but if the analysis is linear, it is limited to linear dependencies).

- The residual signal of a nonlinear predictive analysis, as expected, produces the lower recognition rates, because the relevant information has been retained in the predictor coefficients. However, a maximum of 70% recognition rate is possible.

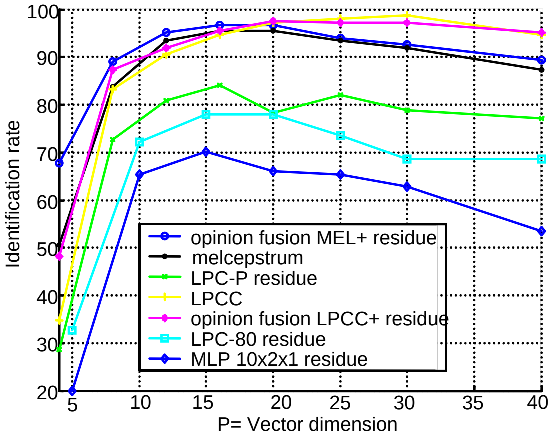

Figure 3: Identification for several parameterization algorithms.

4.1. Opinion fusion

One way to improve the results is by means of a combination of different classifiers opinion [13,8]. In our case, we will use the same classifier scheme, but different parameterizations. In order to study the complementarity of the parameterizations studied, we have computed the correlation coefficient and scatter diagrams.

Table 1 shows the correlation coefficients between distances of several parameterizations. The higher the correlation, the smaller the complementarity of both measures. Figure 2 shows a scatter diagram, which represents points on a two-dimensional space. The coordinates correspond to the obtained distance measures, which correspond to each parameterization (one in each axis). Looking at the diagram we observe that the points diverge from a strip. Thus, they have complementary information and can be combined in order to improve the results.

Table 1: Correlation coefficients between obtained distance values for P=20

	LPCC	Melceps	LP-20 resid	LP-80 resid	MLP 10x2x1
LPCC		0,79	0,68	0,52	0,55
melceps	0,79		0,69	0,56	0,62
LP-20 resid	0,68	0,69		0,78	0,64
LP-80 resid	0,52	0,56	0,78		0,60
MLP 10x2x1	0,55	0,62	0,64	0,60	

When combining different measures, special care must be taken for the range of the values. If they are not commensurate, some kind of normalization must be applied. We have tested the following, based on a sigmoid function [18], $o_i' = \dfrac{1}{1+e^{-k_i}}$

where: $k_i = \dfrac{o_i - (m_i - 2\sigma_i)}{2\sigma_i}$, $o_i' \in [0,1]$, and o_i is the initial opinion of the i^{th} classifier. m_i, σ_i are the mean and standard deviation of the opinions of the i classifier, obtained with data from the authentic speakers (intra-model distances).

We have limited the combinations to the outputs of two different classifiers, and the sum and product combination rules [13].

Table 2. Identification rates (combinations with sum rule)

P Param.	5	10	15	20	25	30	40
LPCC	46.9	90.6	93.5	97.1	98.0	98.8	94.7
Melceps	65.7	89.8	92.7	95.5	93.5	91.8	87.4
LP-P resid	44.1	75.9	84.1	78.4	82.0	78.8	77.1
LP-80 resid	32.7	72.2	78.0	78.0	73.5	68.6	68.6
MLP resid	20.0	65.3	70.2	66.1	65.3	62.9	53.5
LPCC+LP-P	64.9	89.8	94.7	97.6	97.1	97.1	95.1
LPCC+MLP	51.0	91.4	95.1	97.1	97.96	98.4	95.1

We have experimentally observed that slightly better results are obtained without normalization. Looking at figure 4 it can be seen that the distance values obtained with the residual signal parameterization have less amplitude (about 2 to 3 times). Thus, if the normalization is not done, it is equivalent to a weighted combination where the LPCC distances have more influence over the combined result than the residual signal.

Figure 3 and table 2 summarize the identification rates for several vector dimensions (*P*) and different combined parameters.

5. CONCLUSIONS

So far several papers have established that a combination between classical parameters (LPCC, melceps) with some kind of parameterization computed over the residual analysis signal can yield improvements in recognition rates. In our experiments we have found that this is only true when the analysis order ranges from 8 to 16. These values have been selected mainly because a spectral envelope can be sufficiently fitted with this amount of data, so there was no reason to increase the number of parameters. Although we consider that this is true for speech analysis, synthesis and coding, it is interesting to observe that the parameterization step for a speaker recognition system is twofold:

1. We make a dimensionality reduction, so it is easier to compute models, distances between vectors, etc.
2. We make a transformation from one space to another one. In this new domain, it can be easier to discriminate between speakers, and some parameterizations are better than others.

Thus, we are not looking for good quality representation of the speech signal (or a compromise between good representation with the smallest number of parameters). We are just looking for good discrimination capability.

In our experiments we have found that for parameter vectors of high order, although the residual signal has a significant discriminative power among speakers, this signal seems to be redundant with LPCC or melceps, and it is not useful.

If instead of using the residual signal of a linear analysis a nonlinear analysis is used, both combined signals are more uncorrelated and although the discriminative power of the NL residual signal is lower, the combined scheme outperforms the linear one for several analysis orders.

The results show that there is just a marginal improvement on the results when increasing the number of parameters (the identification rate plot saturates), but the residual signal is whiter when increasing the prediction order, especially for the nonlinear analysis. This is a promising result, because although a good parameterization based on nonlinear analysis has not yet been established, this paper reveals that the NL analysis can extract more relevant information with the same prediction order as a linear analysis. Thus, it opens a new way for investigation that has started to provide successful results [2].

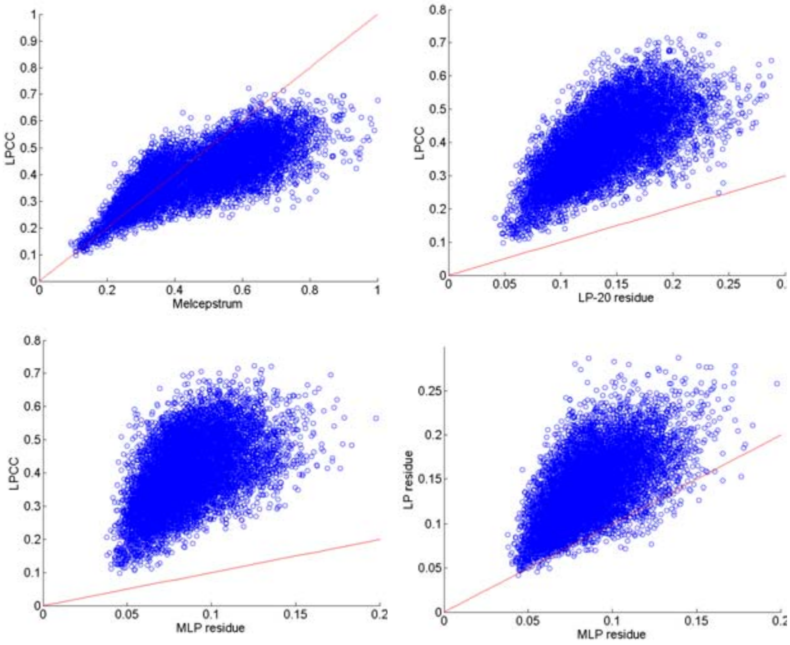

Figure 4: Scatter diagram of distances for observing the correlation between parameters.

6. REFERENCES

[1] F. Bimbot, L. Mathan "Text-free speaker recognition using an arithmetic-harmonic sphericity measure." pp.169-172, Eurospeech 1993.

[2] M. Chetouani, M. Faundez, B. Gas, J. L. Zarader “A New Nonlinear speaker parameterization algorithm for speaker identification”. ISCA Speaker Odyssey Workshop·, 2004.

[3] J. Deller et al. “Discrete-Time Processing of Speech Signals,” Prentice-Hall, 1993.

[4] M. Faundez and D. Rodriguez "Speaker recognition using residual signal of linear and nonlinear prediction models". Vol.2 pp.121-124. ICSLP'98, Sidney.

[5] M. Faundez, F. Vallverdú, E. Monte, "Nonlinear prediction with neural nets in adpcm" IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 1998, Vol I, pp.345-348.Seattle.

[6] M. Faundez "Speaker recognition by means of a combination of linear and nonlinear predictive models" EUROSPEECH'99, Budapest. Vol. 2 pp. 763-766.

[7] M. Faundez et al.. “nonlinear speech processing: overview and applications”. Control and intelligent systems, Vol. 30 Nº 1, pp.1-10, 2002, ACTA Press

[8] M. Faundez “Data fusion in Biometrics”. IEEE Aerosp. Electron. Syst. Mag. Vol.20 nº 1, pp.34-38, January 2005

[9] T. C. Feustel & G. A. Velius “Human and machine performance on speaker identity verification”. Speech Tech 1989, pp.169-170.

[10] F. D. Foresee and M. T. Hagan, “Gauss-Newton approximation to Bayesian regulariza-tion”, proceedings of the 1997 International Joint Conference on Neural Networks, pp.1930-1935, 1997

[11] S. Hayakawa, K. Takeda & F. Itakura “Speaker identification using harmonic structure of LP-Residual spectrum” Audio Video Biometric personal autentification 1997, pp. 253-260 LNCS-1206.

[12] J. He, L. Liu & G. Palm “On the use of features from prediction residual signals in speaker identification”. EUROSPEECH’1995 pp.313-316.

[13] J. Kittler, M. Hatef, R. P. W. Duin & J. Matas “On combining classifiers”. IEEE Trans. On PAMI, Vol. 20, Nº 3, pp. 226-239, 1998

[14] L. Liu et al. “Signal modelling for speaker identification”. Proceedings of the IEEE ICASSP 1996, Vol.2, pp. 665 - 668

[15] R. J. Mammone, X. Zhang & R. Ramachandran “Robust speaker recognition” IEEE Sig. Proc. magazine, 1996, pp.58-70.

[16] J. Ortega et al. “Ahumada: a large speech corpus in Spanish for speaker identification and verification”. ICASSP 1998 Seattle, Vol. 2, pp. 773-776.

[17] J. W. Picone “Signal Modeling techniques in speech recognition” Proceedings of the IEEE, Vol. 79, Nº 4, April 1991, pp.1215-1247

[18] C. Sanderson “Information fusion and person verification using speech & face information”. IDIAP Research Report 02-33, pp. 1-37. September 2002

[19] R. Schlüter, H. Ney “Using phase spectrum information for improved speech recognition performance” Proceedings of the IEEE ICASSP 2001, Vol.1, pp.133-136

[20] P. Thévenaz, H. Hügli “Usefulness of the LPC-residue in text-independent speaker verification” Speech Communication 17 (1995) pp. 145-157

[21] http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[22] B. Yegnanarayana et al. “Source and system features for speaker recognition using AANN models” IEEE ICASSP 2001, Vol.1, pp.409-412