Use of The Harmonic Phase in Speaker Recognition

Inma Hernáez, Ibon Saratxaga, Jon Sanchez, Eva Navas, Iker Luengo

Aholab-Signal Processing Laboratory – Faculty of Engineering
University of the Basque Country
Urkijo zum. z/g 48013. Bilbao-Spain
{inma, ibon, ion, eva, iker}@aholab.ehu.es

Abstract

In this paper a novel set of features with a promising ability to identify speakers is presented. These features are based on the harmonic phase of the speech signal and have been previously used successfully in an ASR task. Using the SI-284 subset of the WSJ database, a GMM has been trained for each of the 283 speakers and several speaker identification experiments have been performed, with a high level of success. The feature extraction method and the performed experiments are described. The results show that the features present excellent identification performance, very close to the performance of the MFCC parameters.

Index Terms: Speech analysis, Phase analysis, Speaker recognition, Speaker modeling, Harmonic modeling.

1. Introduction

Conventional speaker recognition (SR) methods use short time spectral information to model the speaker specific vocal tract parameters. From this spectral information, typically only the magnitude spectrum is used and the phase component is ignored. Magnitude information is directly related with spectral power density, and therefore with the formant structure and intelligibility, and it can be parameterized in a relatively easy way. However, the use of other kinds of features has also proved useful in a speaker recognition problem. In [1] parameters extracted from the glottal flow derivative are used to identify speakers. Parameters derived from the residual phase extracted using linear prediction analysis were used in [2] and [3], both in isolation and in combination with MFCC features, offering significant improvements. The Modified Group Delay function derived from the phase spectrum has also been used and many recent works report significant improvements of the recognition rate when used in combination with MFCC parameters [4][5][6]. In [7] and [8] a phase information extraction method is proposed and great improvements in the recognition rate are obtained when this information is used in combination with MFCC parameters.

In this paper we present a new set of parameters also derived from the phase of the speech signal. The used parameters are derived from the so called Relative Phase Shift [9], which was first successfully used in a speech polarity detection problem [10]. The parameterization of the RPS using the Discrete Cosine Transform (DCT) was proposed in [12] to improve the performance of an ASR system. This parameterization was also used to detect synthetic signals in a problem of imposture in a speaker verification framework [11]. In this work we present the results of a set of experiments that demonstrate the ability of this new representation of the phase to capture speaker specific features and thus to identify the speaker. The results show that the parameters used are almost as powerful as the MFCC in our experimental set-up.

2. The DCT-mel-RPS Representation

The Relative Phase Shift is a representation for the harmonic phase information and was first described in [9]. In this section we present the process followed to obtain the parameters used to train the speakers' models.

2.1. Definition of the Relative Phase Shift

Harmonic analysis models each frame of a signal by means of a sum of sinusoids harmonically related to the pitch or fundamental frequency:

$$h(t) = \sum_{k=1}^{N} A_k \cos\left(\varphi_k(t)\right) \qquad \varphi_k(t) = 2\pi k f_o t + \theta_k \tag{1}$$

where N is the number of bands, A_k and $\varphi_k(t)$ are the amplitudes and instantaneous phases of the harmonics, f_0 is the pitch or fundamental frequency, and θ_k is the initial phase shifts of the k-th sinusoid.

Usually the term "phase" is applied to the whole instantaneous phase of every sinusoid, $\varphi_k(t)$, instead of the initial phase shift θ_k. This instantaneous phase changes depending on the analysis instant as well as on the frequency of the harmonic, due to the linear phase term $2\pi k f_o t$. On the contrary, the initial phase shift θ_k is constant while the waveform shape is stable under the assumption of local stationarity, regardless of the time instant chosen for the analysis.

The initial phase shift determines the waveform shape of the signal. For a given set of harmonic sinusoids the resulting waveform shape depends only on the differences between the initial phase shifts θ_k of the components, which we call Relative Phase Shifts (RPS). These RPSs are also constant as long as the initial phase shifts are so. Thus, they can be calculated at any analysis point wherever local stationarity conditions can be assumed, avoiding the necessity of determining any special point for the analysis. Being relative, the RPSs are computed using a common reference (f_0).

We have developed an expression to obtain the relative differences of the initial phase shifts from the measured instantaneous phases. Let us consider two sinusoids:

$$x_1(t) = \cos\left(2\pi f_1 t + \theta_1\right) \qquad x_k(t) = \cos\left(2\pi f_k t + \theta_k\right) \tag{2}$$

where $x_1(t)$ will be a reference sinusoid with frequency f_1 and $x_k(t)$ another sinusoid with frequency $f_k > f_1$. For the sake of simplicity we will consider θ_1=0, which implies setting the time origin at the point where $x_1(t)$ has instantaneous phase zero. For any arbitrary analysis point (t_a) the instantaneous phases are:

$$\varphi_1(t_a) = 2\pi f_1 t_a \qquad \varphi_k(t_a) = 2\pi f_k t_a + \theta_k \tag{3}$$

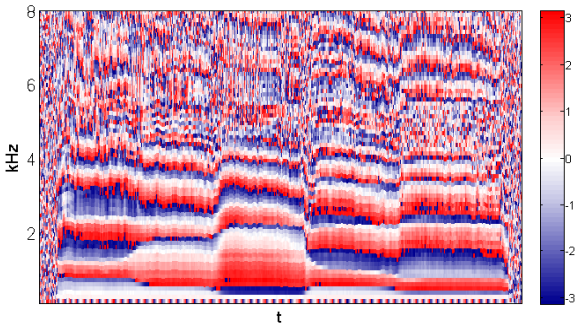

Figure 1: RPS phasegram of a voiced speech signal /aeiou/.

In the case of harmonic analysis, f_1 will be the fundamental frequency f_0 and the frequencies of the two sinusoids will be harmonically related, so $f_k=kf_1$. Applying this condition, we get the RPS expression:

$$\theta_k = \varphi_k(t_a) - k\varphi_1(t_a) \quad (4)$$

Finally the RPS is wrapped to values in the [-π, π] interval.

Among other interesting properties of the RPS a major feature is that it reveals a structured pattern in the phase information of the voiced segments. This can be noticed in Figure 1 which shows a "RPS phasegram" that, as its magnitude counterpart the spectrogram, shows the evolution along time of the RPS for each harmonic. Figure 1 shows a phasegram of the voiced speech segment of five sustained vowels */aeiou/*, where the stable pattern of every vowel can be clearly distinguished.

Despite the defined and clear looking of the RPS patterns, they can not be directly employed as parameters in a GMM based SR system. There are several points that need to be addressed, and they are discussed next in this section.

2.2. Unvoiced speech frames

There is no meaningful RPS for unvoiced speech. In the unvoiced segments of the speech there is no valid reference (f_0) to base the RPS calculations. Furthermore, in unvoiced segments the excitation signal is supposed to be a random signal, whose phases will be random, and this will mask the phase structure of the vocal tract. This lack of meaningful information in the unvoiced segments is not only a problem of the RPS transformation, but also an inherent problem of the random phase signals which hinder the effect of the phase response of the filter. Because no temporal information will be used in the speaker recognition experiments described next, unvoiced frames will be eliminated, and only voiced frames will be considered.

2.3. RPS envelope features

In order to parameterize the RPS data by frame it is necessary to analyze the usual shape of the RPS function across the frequencies in different segments of different speakers.

First the problem of wrapping arises: the RPS values are wrapped values given in the range [-π,π] which produces discontinuities in the frequency axis. Unwrapping is performed in order to get a smooth function that can be parameterized by a reduced number of values. But unwrapping is an ambiguous operation that can produce very different results for similar data. This suggests the use of differentiated unwrapped RPSs as source information for the parameterization.

Figures 2a and 2b show the unwrapped RPS envelope for several consecutive frames of the same vowel uttered by two male speakers. Figures 2c and 2d show the differentiated RPS envelopes, where common features between both speakers can be more easily appreciated, in particular the phase jumps at the formant frequencies. However, it is also clear from Figure 2a and 2b that each speaker presents a very different slope of the RPS envelope, which is seen as an offset in the differentiated RPS. This property can help in the characterization of the speakers. Therefore we explicitly add the value of the average slope (the average value of the differentiated RPS) as a parameter. In order to avoid redundant information in the parameterization, and to make it easier to compare the shape of the RPS envelopes, this offset is subtracted in the differentiated RPS envelope.

2.4. Variable number of parameters and dimensionality reduction

The number of RPS values varies from frame to frame as it is dependent on the number of pitch harmonic components that fit in the analyzed spectral bandwidth, which varies with the pitch value. For usual pitch values, the number of harmonic components is too high, forcing high dimensionality models if used directly.

To cope with this situation, the differentiated RPS data are filtered using a normalized Mel filter bank, where the number of filters can be fixed according to the sampling frequency. For 8 kHz sampling frequency, 32 filters have been used. This non-uniform averaging of the harmonic phase values produced better results than other linear approaches in an ASR task [12].

The DCT has been frequently employed in the literature to parameterize both magnitude and phase data to be used in ASR models. This transformation decorrelates the elements of the feature vector, making it suitable for diagonal covariance matrix statistical models. At the same time, the DCT successfully reduces the number of parameters needed to model speech. We also use the DCT even though the spiky shape of the differentiated unwrapped RPS worsens the DCT modeling capability.

3. Experiments

This section presents the experiments carried out in order to assess the effectiveness of the proposed parameterization in SR tasks performed. First the database and the classifier used are described. Then the performed experiments and results are commented.

3.1. Evaluation setup

To evaluate the behaviour of the new set of parameters the Wall Street Journal database [13] has been selected. The use of this database has been considered appropriate for our purposes, mainly because it is a clean speech database, microphonic, it contains a large number of speakers and a sufficient amount of recordings from each speaker. In this way, problems related to data scarcity or noisy conditions are avoided, and the focus is centred on the performance of the parameters.

From this database the part known as SI-284 has been selected. It contains utterances read by 283 speakers with a Sennheiser microphone sampled at 16 kHz. In the experiments described here, the signals were downsampled to 8 kHz. The amount of speech available for each speaker is variable, ranging from 100 to 150 utterances. The utterances are also of different lengths, with durations ranging from 5 to 8 s each.

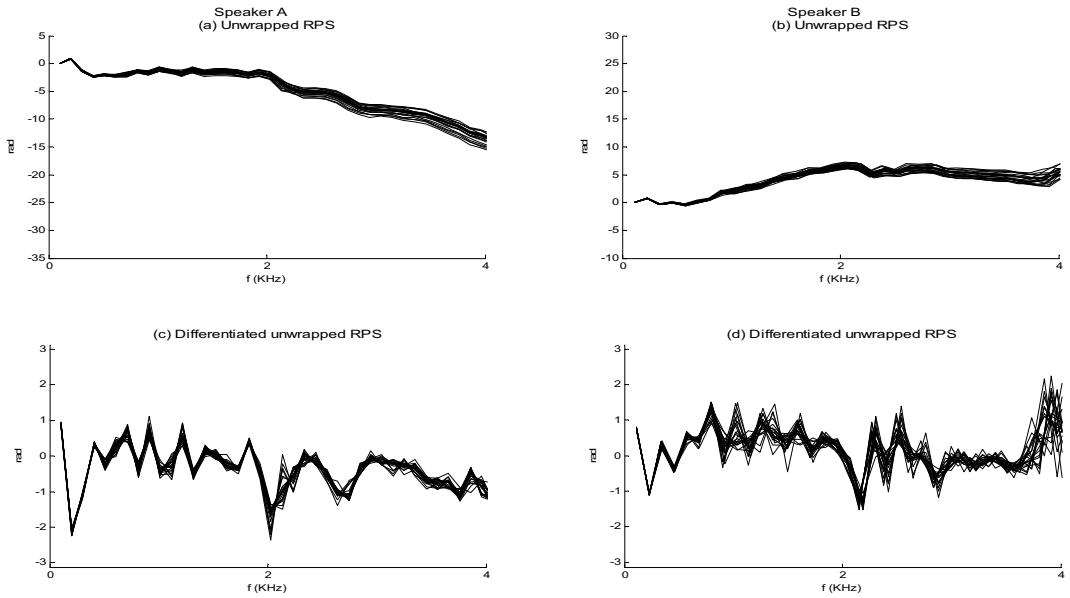

Figure 2: *a-b Unwrapped RPS envelope of several consecutive frames for two speakers. c-d differentiated unwrapped RPS envelope for two speakers.*

The database has been randomly divided into a training set (60% of the utterances) and a testing set (40%).

A classic classifier based on GMM has been used [14]. One GMM with 32 mixtures and diagonal covariance matrices has been trained for each speaker, using varying training time lengths and different number of parameters. The number of Gaussians was experimentally determined. In all the experiments presented here, the whole testing set has been used. One important consideration is that in the experiments that use the phase parameters, only voiced frames are used both for training and testing. On the contrary, the MFCC baseline experiments use all the available non-silent frames.

Speaker identification is performed based on maximum likelihood:

$$\arg\max_j \left\{\log p(\mathbf{X}_i / \lambda_j)\right\} \quad (5)$$

where $\mathbf{X}_i$ is the feature vector sequence $\{\mathbf{x}_n\}_i$ for the test speaker i and λ_j is the GMM trained for speaker j. The log-probability is calculated for N test feature vectors as:

$$\log p(\mathbf{X} / \lambda) = \frac{1}{N}\sum_{n=1}^{N} \log p(x_n / \lambda) \quad (6)$$

3.2. Experiments and results

As this was the first time that this set of parameters was used for this task, we focused on the evaluation and optimization of the feature vector rather than on the development of the classifier.

On account of the great variability of the sentences lengths and to obtain a more solid score, the process of training the models with a reduced number of sentences N consisted on making as many groups of N sentences as possible with the training sets, and then averaging the results.

The phase feature vector is formed by the average slope obtained as explained in section 2, together with a number of coefficients from the DCT of the differentiated unwrapped RPS envelope. The first set of experiments was directed to estimate the number of DCT parameters needed, and for that purpose a series of experiments was performed using all the available training frames. As a first approximation, the number of DCT coefficients was truncated to 20, using a criterion based on the reconstruction error. After some tests, it was observed that if the training time was not limited, a very accurate identification was possible. As shown in Table 1, 100% success is obtained with only 4 DCT parameters.

However, limiting the training time degrades significantly the performance of the system. It has to be considered that only voiced frames are used both for training and testing. This implies that approximately 60% of the non-silent frames are used (i.e. 60% of the number of frames used by the baseline MFCC system).

As can be seen in Table 1, using the full set of parameters 100% accuracy is achieved when training with only 10 utterances (50 to 80 s depending on the speaker and the set of sentences). A further reduction of the training time to 5 utterances causes the accuracy falling down to 98.2%.

An interesting result is the performance of the system using only one parameter: the slope of the unwrapped RPS envelope (represented as S in Table 1). It is clear that important information about the speaker identity is kept in this slope: when used in isolation and with a sufficient number of training frames an accuracy of 93.6 is achieved. However, the contribution of the DCT coefficients is also very relevant as show the results for 4 DCT set in Table 1.

Finally, the results of a fusion experiment of spectral envelope and phase parameters (MFCC+S+20) are shown in the last row of Table 1. This experiment also uses 32 mixtures for the models. The use of the additional information is not able to improve the performance of the MFCC parameters in isolation.

Further experiments developed using still less number of training sentences revealed as expected a fast degradation on the performance of the phase based system. The MFCC baseline system on the contrary remains quite robust up to as few as 2 sentences (95% identification rate). At this point we have to remember again that the phase parameters are based on the harmonic phase, which is not meaningful for unvoiced

Table 1. *Speaker identification rate (%) for different number of training utterances (N) using different parameter sets (S: average slope of the RPS envelope)*

Parameters	*N=all*	*N=15*	*N=10*	*N=5*
MFCC	100	100	100	99.8
S+20 DCT	100	100	100	98.2
S+4 DCT	100	98.6	97.5	88.3
4 DCT	100	98.2	96.8	78.8
S	93.6	52.3	41.7	25.8
MFCC+S+20DCT	100	100	100	99.8

frames, which means that for so small amount of time an insufficient number of frames are used for training.

The fact that the baseline experiment produces so good results reveals that this task is an easy one, mainly because it is a very clean database. However, up to now, no other set of parameters based on the phase of the signals, and with no use of the envelope has been described that presents a performance comparable to that of MFCC. Other phase representations described in the literature achieve important improvements over the baseline, only when used in combination with it. The novelty of our system is that it shows a comparable performance when used by itself, even with the handicap that only approximately 60% of the time is useful.

4. Conclusions

A novel parameterization method capable of discriminating speakers' identity has been presented. The experiments and results presented in this paper are preliminary and do not pretend the implementation of a realistic Speaker Identification or Verification System: a very simple detection method has been used and the signals tested fulfilled ideal conditions. The phase spectrum has already been used in previous works cited above. However, our proposal deals with the phase in a completely novel way, leading to 100% accuracy under ideal conditions, with a performance very close to the MFCC static parameters.

Many open questions remain: how will the performance of the parameters under noisy conditions be? Will the parameters be robust to channel distortions? Is there a better transformation other that the one proposed? Communication channels nowadays introduce coding stages that distort the phase of the signal, and it is unknown to us yet up to what point the speaker identification features remain in the transcoded signal. On the other hand, the parameter set should be evaluated in a more standard verification task following NIST procedures and databases. Most of these databases are telephonic databases, where the first harmonic f0 is missing. Therefore a new reference frequency must be used. The properties of the phase representation in that case pose new challenges and have not been investigated yet. In any case, an interesting research space has been opened in the speaker characterization field.

5. Acknowledgements

This work has been partially supported by the Spanish Ministry of Science and Innovation (Buceador Project, TEC2009-14094-C04-02) and The Basque Government (Berbatek, IE09-262, MV20090225).

6. References

[1] M.D. Plumpe, T.F. Quatieri and D.A. Reynolds, "Modeling of the glottal flow derivative waveform with application to speaker identification", IEEE Trans. Speech and Audio Processing, Vol. 7, No. 5, pp. 569–586, 1999.

[2] K.S.R. Murty and B. Yegnanarayana, "Combining evidence from residual phase and MFCC features for speaker verification", IEEE Signal Processing Letters, Vol.13, No. 1, pp. 52–55, 2006.

[3] N. Zheng, T. Lee and P.C. Ching, "Integration of complementary acoustic features for speaker recognition", IEEE Signal Processing Letters, Vol. 14, No. 3, pp. 181–184, 2007.

[4] R. Padmanabhan, S. Parthasarathi, H. Murthy, "Robustness of phase based features for speaker recognition", Proc. INTERSPEECH, pp. 2355–2358, 2009.

[5] J. Kua, J. Epps, E. Ambikairajah, E. Choi, "LS regularization of group delay features for speaker recognition", Proc. INTERSPEECH, pp. 2887–2890, 2009.

[6] R. M. Hegde, H. A. Murthy, G. V. Ramana Rao, "Application of the modified group delay function to speaker identification and discrimination". Proc. ICASSP 2004, pp. I-517-I-520, 2004.

[7] L. Wang, S. Ohtsuka, and S. Nakagawa, "High improvement of speaker identification and verification by combining MFCC and phase information" Proc. ICASSP 2009, pp. 4529-4532, 2009.

[8] L. Wang, K. Minami, K. Yamamoto, and S. Nakagawa, "Speaker identification by combining MFCC and phase information in noisy environments", Proc. ICASSP 2010, pp. 4502-4505, 2010.

[9] I. Saratxaga, I. Hernáez, D. Erro, E. Navas, J. Sánchez, "Simple representation of signal phase for harmonic speech models", Electronics Letters, 45: 381-383, 2009.

[10] I. Saratxaga, D. Erro, I. Hernáez, I. Sainz, E. Navas, "Use of harmonic phase information for polarity detection in speech signals" Proc. INTERSPEECH 2009, 1075-1078, 2009.

[11] P. de Leon, I. Hernaez, I. Saratxaga, M. Pucher, J. Yamagishi, "Detection of synthetic speech for the problema of imposture" Proc. ICASSP 2011 (*to appear*), 2011.

[12] I. Saratxaga, I. Hernaez, I. Odriozola, E. Navas, I. Luengo, D. Erro, "Using harmonic phase information to improve ASR rate" in Proc. INTERSPEECH 2010, pp. 1185-1188, 2010.

[13] D.B. Paul and J.M. Baker, "The design for the wall street journal-based CSR corpus," in Proc. of the workshop on Speech and Natural Language, Harriman, New York, pp. 357–362, 1992.

[14] M. Brooks, "VOICEBOX: Speech Processing Toolbox for MATLAB", Online: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html, accessed on 31st March 2011