# THE USE OF HARMONIC FEATURES IN SPEAKER RECOGNITION

*Bojan Imperl and Zdravko Kačič and Bogomir Horvat*

Laboratory for Digital Signal Processing,
University of Maribor, Faculty of Electr. Eng. and Comp. Sci.,
Smetanova 17, 2000 Maribor, Slovenia

## ABSTRACT

In this paper the *Harmonic* features based on the harmonic decomposition of the Hildebrand – Prony line spectrum are introduced. A Hildebrand – Prony method of spectral analysis was applied because of its high resolution and accuracy. Comparative tests with the LP and LP – cepstral features were made with 50 speakers from the Slovene database SNABI (isolated words corpus) and 50 speakers of the German database BAS Siemens 100 (utterances of sentences). With both databases the advantages of the Harmonic features were noticed especially for the speaker identification while for the speaker verification the Harmonic features have performed better on the SNABI database and as good as the LP cepstral features on the BAS Siemens 100 database.

## 1. INTRODUCTION

Different features have been investigated in speaker recognition systems. The cepstral or Linear Prediction - cepstral features (LPCC) have usually been reported to have yielded good performance [13, 12, 1, 8]. Comparative tests with other features such as Line Spectrum Pairs (LSP) have shown better results [11, 2]. In the LSP representation the variations in the glottis and the vocal tract that determine the speaker's voice are transferred into frequency domain. In this paper we are introducing another type of features that also detect the differences between different speakers in the frequency domain – the Harmonic features.

The Harmonic features are produced by the harmonic decomposition of high resolution spectral line estimate. Feature vector consists of the fundamental frequency followed by amplitudes of several harmonic components. The Harmonic features can be produced only on segments of speech that contain the harmonic structure, i. e. the voiced sounds. Moreover, the long vowels and nasals were also found to be the most speaker specific in numerous studies on speaker discriminating properties of the phonemes [4, 7, 3].

Providing an accurate spectral line estimate is the crucial task in Harmonic feature extraction process. Several spectral line estimation methods can be used for this purpose. In this case the Hildebrand – Prony spectral line estimation method, which has been reported to be very accurate when applied on short data records with high signal to noise ratio [6, 14, 10], was used.

To evaluate the performance of the Harmonic features the comparative tests with the LPCC features were performed on the phoneme based speaker identification / verification system.

## 2. SPECTRAL ANALYSIS AND FEATURE EXTRACTION

Compared to the Discrete Fourier Transformation the Hildebrand – Prony method has several performance advantages. The major difference between the Hildebrand – Prony method and the Discrete Fourier Transformation (DFT) is that the DFT may be regarded as a least square fit of sines and cosines of predetermined frequencies to the input data while in the Hildebrand – Prony method the frequencies are calculated from the input data. Moreover, the DFT problems such as leakage [10] are not met with the Hildebrand – Prony method. The result of the Hildebrand – Prony analysis is hence much more accurate. Figure 1 shows the result of both analyses on the segment of vowel <aː>.

There are several problems when applying the Hildebrand – Prony method of which the problem of order definition [10] is the essential. The order can be determined by monitoring the residual squared error:

$$E = \sum_{k=0}^{N-1} |x_k - \tilde{x}_k|^2 \,, \qquad (1)$$

where $x_k$ is the sample and the $\tilde{x}_k$ its $x_k$ approximation produced by the Hildebrand – Prony model. In our tests on speech signals (long vowel segments) the order 80 was usually found to yield the smallest residual squared error.

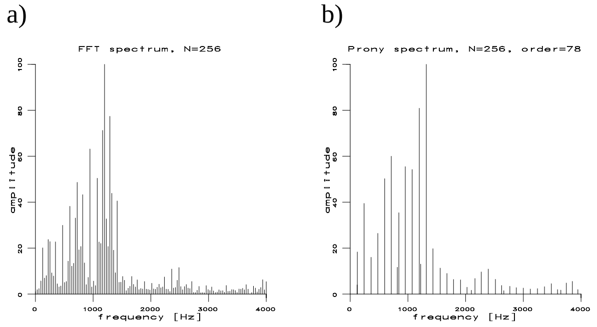

Figure 1: The result of the Fourier (a) and Hildebrand-Prony (b) analysis of speech signal (vowel <aː>).

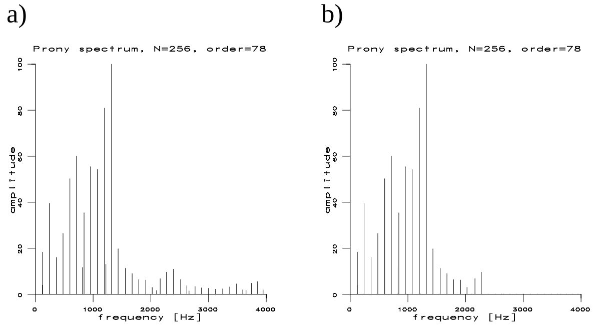

Figure 2: The Hildebrand-Prony spectrum of a vowel <aː> (order of 78): a) the complete spectrum, b) first 19 extracted harmonic components.

Having the Hildebrand – Prony line spectrum the Harmonic features were generated by its harmonic decomposition, i.e., the fundamental frequency and the amplitudes of higher harmonics were extracted. Figure 2 shows the result of the Hildebrand – Prony analysis of vowel <aː> and the result of the harmonic decomposition. The fundamental frequency and the amplitudes of the first 19 harmonic components were used to form the Harmonic feature vector.

## 3. ARCHITECTURE AND SYSTEM OPERATION

In order to evaluate the Harmonic features the phoneme based speaker recognition system that uses either Harmonic features or 20-th order LPC cepstrum features (LPCC) was designed. Evaluation of Harmonic features was done by comparing its performance to the LPCC features efficiency.

The long vowels and nasals exhibit the highest speaker specificy [4, 7, 3] when cepstral features are concerned. Experiments with Harmonic features have shown that not all long vowels and nasals are equally speaker specific. Long vowel <aː> was found to be the most characteristic for each individual speaker, therefore the speaker recognition system was based on the analysis of this vowel in the utterances. The vowel detection problem was not considered in this system – all vowels were manually extracted from the test and train data. In each such segment feature vectors were extracted on five places uniformly distributed over the vowel steady state (inner 60% of the vowel). The number of feature vectors was limited to five because of the complexity of the Hildebrand – Prony spectral analysis method. The edge - sections of vowel (40%) were not included in the analysis since they are assumed to be too affected by the articulatory effect.

The system is designed to operate in either speaker identification or speaker verification mode. In both cases the same reference database consisting of extracted reference feature vectors was used. Two separate classification procedures were implemented. For the speaker identification using the Harmonic features, the feature vectors derived from the test utterances were simply matched to the reference feature vectors where the following distance measure was used:

Harmonic features:

$$\begin{aligned} \Delta_{ij} &= (f_{i_0} - f_{j_0})^2 + \sum_{n=1}^{19} w \mid f_{i_n} - f_{j_n} \mid \qquad (2) \\ w &= \begin{cases} 6\,, & 1 \le n \le 4 \\ 4\,, & 5 \le n \le 8 \\ 2\,, & 9 \le n \le 14 \\ 1\,, & 15 \le n \le 19 \end{cases} \end{aligned}$$

other:

$$\begin{aligned} \Delta_{ij} &= \sum_{n=0}^{19} w \mid f_{i_n} - f_{j_n} \mid \qquad (3) \\ w &= 1\,, \qquad 0 \le n \le 19 \end{aligned}$$

where:
$\Delta_{ij}$ . . . is the difference between i-th and j-th feature vector,
$f_{i_n}$ . . . is the n-th element of the i-th (reference) feature vector,
$f_{j_n}$ . . . is the n-th element of the j-th (test) feature vector,
$w$ . . . weight.

For speaker verification the same distance measures were used for each type of features as for speaker identification. The decision whether the speaker is accepted or not is made according to the speaker verification threshold. The threshold is speaker dependent and is defined according to the intra-speaker variability:

$$\sigma_k = W \, \bar{\Delta}_{k_{ref}} \;, \tag{4}$$

where the $\sigma_k$ is the threshold for speaker $k$, $W$ is experimentally derived scaling factor (values are given with experimental results) and the $\bar{\Delta}_{k_{ref}}$ is the average distance between all reference feature vectors of speaker $k$. Tests with threshold equal for all speakers have yielded higher error rates.

## 4. DATABASES

Two sets of comparative closed-set tests were performed. First, the system was tested with 50 speakers (19 female, 31 male) of Slovene Speech Database SNABI [9]. 14 different utterances per speaker containing vowel <aː> were retrieved from the isolated word corpus. Next, the system was tested with 50 speakers (22 female, 28 male) from the German database BAS Siemens 100 (SI100). 14 various sentences containing vowel <aː> per speaker were used. For the results reported in this paper first half of the utterances were used for training and other half for testing (7 training and 7 test utterances). Speech material in both databases was of studio quality, sampled at 16 kHz by 16 bit A/D conversion.

## 5. EXPERIMENTAL RESULTS

Table 1 shows the results of speaker identification for 50 speakers of the SNABI database, using different number of test utterances.

| SNABI – identification | | |
|---|---|---|
| features | UT | $E_I$ |
| Harmonic | 1 | 11.0 |
| | 3 | 4.6 |
| | 5 | 1.5 |
| | 7 | 0.0 |
| LPCC | 1 | 29.7 |
| | 3 | 14.3 |
| | 5 | 8.4 |
| | 7 | 6.0 |

UT - utterances for testing
$E_I$ - identification error rate [%]

Table 1: The identification error rates for different number of test utterances with SNABI database.

The identification error rate was the highest when the identification of each speaker was done using only one out of seven available test utterances. Table 2 shows the results of speaker verification for the same set of speakers where the error rates $E_{FA}$, $E_{FR}$, $E_V$ and $E_T$ were defined as in [5].

| SNABI – verification | | | | | | |
|---|---|---|---|---|---|---|
| features | $W$ | UT | $E_{FA}$ | $E_{FR}$ | $E_V$ | $E_T$ |
| Harmonic | 2.6 | 1 | 11.0 | 11.0 | 11.0 | 10.7 |
| | | 3 | 8.1 | 7.5 | 7.8 | 8.1 |
| | | 5 | 5.5 | 3.4 | 4.5 | 5.4 |
| | | 7 | 3.9 | 2.0 | 2.9 | 3.8 |
| LPCC | 2.1 | 1 | 14.1 | 19.0 | 16.6 | 14.2 |
| | | 3 | 12.3 | 15.8 | 14.1 | 12.4 |
| | | 5 | 11.8 | 13.7 | 12.8 | 11.8 |
| | | 7 | 11.5 | 12.0 | 11.8 | 11.5 |

$W$ - threshold weight (Equation 4)
UT - utterances for testing
$E_{FA}$ - false acceptance error rate [%]
$E_{FR}$ - false rejection error rate [%]
$E_V$ - verification error rate [%]
$E_T$ - total error rate [%]

Table 2: The results of verification for different number of test utterances with SNABI database.

In the speaker identification experiment significantly higher accuracy was achieved with Harmonic features at different number of used test utterances. Similar result can be observed also in speaker verification, however, the advantages of the Harmonic features seems less obvious when small amount of test data is used (1 test utterance; UT = 1)

The threshold weight $W$ was reduced for LPC features. This seems to indicate that the Harmonic features are more influenced by the intra speaker variability than the LPC features.

Tests with the BAS Siemens 100 database included the segments extracted from the sentence utterances. The identification and the verification rates were reduced due to incorporation of the prosody. Tables 3 and 4 presents the speaker identification and verification results for 50 speakers of the SI100 database.

In tests with SI100 database the difference between the accuracy of both types of features is smaller than in tests with isolated words (SNABI database). In the speaker identification the LPC features were still outperformed by the Harmonic features while in the verification tests the results with both types of features were similar.

| SI100 – identification | | |
|---|---|---|
| features | UT | $E_I$ |
| Harmonic | 1 | 49.1 |
| | 3 | 23.1 |
| | 5 | 12.0 |
| | 7 | 6.0 |
| LPCC | 1 | 35.7 |
| | 3 | 21.6 |
| | 5 | 13.2 |
| | 7 | 10.0 |

UT - utterances for testing
$E_I$ - identification error rate [%]

Table 3: The identification error rates for different number of test utterances with SI100 database.

| SI100 – verification | | | | | | |
|---|---|---|---|---|---|---|
| features | $W$ | UT | $E_{FA}$ | $E_{FR}$ | $E_V$ | $E_T$ |
| Harmonic | 2.9 | 1 | 19.0 | 25.0 | 22.0 | 19.1 |
| | | 3 | 17.6 | 20.4 | 19.0 | 17.6 |
| | | 5 | 16.7 | 16.5 | 16.6 | 16.7 |
| | | 7 | 16.0 | 14.0 | 15.0 | 16.0 |
| LPCC | 2.2 | 1 | 15.0 | 21.4 | 18.2 | 15.1 |
| | | 3 | 14.6 | 19.1 | 16.8 | 14.7 |
| | | 5 | 14.3 | 17.4 | 15.8 | 14.4 |
| | | 7 | 14.2 | 16.0 | 15.1 | 14.2 |

$W$ - threshold weight (Equation 4)
UT - utterances for testing
$E_{FA}$ - false acceptance error rate [%]
$E_{FR}$ - false rejection error rate [%]
$E_V$ - verification error rate [%]
$E_T$ - total error rate [%]

Table 4: The results of verification for different number of test utterances with SI100 database.

## 6. CONCLUSION

The Harmonic features were found to be effective when applied in a speaker recognition systems. Its performance is found to be superior to the LPCC features especially in case of speaker identification. The Harmonic features have shown better ability to distinguish different speakers than LPCC features, which is an essential requirement in speaker recognition tasks with larger number of speakers. Tests with the SI100 database (utterances of sentences) have shown significant increase of identification and verification error rate for the Harmonic features while the error rates for the LPCC features were not that critically increased. This shows that the Harmonic features are more influenced by the prosodic features than the LP – based features. One possible solution to this problem may be in increasing the size of the training data.

## 7. REFERENCES

[1] K. T. Assaleh and R. J. Mammone. Robust cepstral features for speaker identification. In *ICASSP*, pages 129 – 132, Adelaide, 1994.

[2] J L. Bonifas, I. Hernandez Rioja, B. Etxebarria Gonzalez, and S. Saoudi. Text – dependent speaker verification using dynamic time warping and vector quantization of lsf. In *EUROSPEECH*, pages 359 – 362, Madrid, 1995.

[3] I. Magrin Chagnolleau, J. F. Bonastre, and F. Bimbot. Effect of utterance duration and phonetic content on speaker identification using second – order statistical methods. In *EUROSPEECH*, pages 337 – 340, Madrid, 1995.

[4] J. P. Eatock and J. S. Mason. A quantitative assessment of the relative speaker discriminating properties of the phonemes. In *ICASSP*, pages 133 – 136, Adelaide, 1994.

[5] S. Furui. Cepstral analysis technique for automatic speaker verification. *IEEE Trans. on ASSP*, 29(2):254 – 272, 1981.

[6] F. B. Hildebrand. *Introduction to Numerical Analysis*. McGraw-Hill, New York, 1956.

[7] M.-J. Caraty J.-L. Le Floch, C. Montcié. Investigations on speaker characterization from orphée system technics. In *ICASSP*, pages 149 – 152, 1994.

[8] J. S. Mason J. Thompson. Within class optimization of cepstra for speaker recogniton. In *EUROSPEECH*, pages 165 – 168, 1993.

[9] Z. Kačič, B. Horvat, and R. Derlič. Zasnova baze izgovarjav slovenskega jezika snabi. In *ERK'94*, pages 327 – 330, Portorož, 1994. Slovene Section of IEEE.

[10] S. M. Kay and S. L. Marple. Spectrum analysis – a modern perspective. *Proceedings of the IEEE*, 69(2):1380 – 1419, november 1981.

[11] Chi Shi Liu, Wern Jun Wang, Min Tau Lin, and Hsiao Chuan Wang. Study of line spectrum pair frequencies for speaker recognition. In *ICASSP*, pages 277 – 280, 1990.

[12] M.Savic and J.Sorensen. Phoneme based speaker verification. In *Proceed. IEEE International Conference on Acoustics, Speech and Signal Processing*, pages II–165 – II–168, 1992.

[13] J. P. Openshaw, Z. P. Sun, and J. S. Mason. A comparison of composite features under degraded speech in speaker recognition. In *ICASSP*, pages 371 – 374, 1993.

[14] T. M. Sullivan and O. L. Frost and J. R. Treichler. High resolution signal estimation. Technical report, ARGO Systems, Internal report, June, 1978.