ISCA Archive
http://www.isca-speech.org/archive

ODYSSEY04 - The Speaker and
Language Recognition Workshop
Toledo, Spain
May 31 - June 3, 2004

Features for Speaker and Language Identification

Leena Mary, K. Sri Rama Murty, S.R. Mahadeva Prasanna and B. Yegnanarayana

Speech and Vision Laboratory
Department of Computer Science and Engineering
Indian Institute of Technology Madras, Chennai-600036, India
{leena,sriram,prasanna,yegna}@cs.iitm.ernet.in

Abstract

In this paper we examine several features derived from the speech signal for the purpose of identification of speaker or language from the speech signal. Most of the current systems for speaker and language identification use spectral features from short segments of speech. There are additional features which can be derived from the residual of the speech signal, which correspond to the excitation source of speech signal. These features at the subsegmental (less than a pitch period) level correspond to the glottal vibration in each cycle, and at the suprasegmental (several pitch periods) level the features correspond to intonation and duration characteristics of speech. At the subsegmental level features can be extracted from the residual signal and also from the phase of the residual signal. The characteristics of speaker or language can be captured from the spectral or subsegmental features using Autoassociative Neural Network (AANN) models. We demonstrate that these features indeed contain speaker-specific and language-specific information. Since these features are more or less from independent sources, it is likely that they provide complementary information, which when combined suitably will increase the effectiveness of speaker and language identification systems.

1. Introduction

Speech signal contains information about the message, speaker characteristics and language characteristics, besides emotional state of the speaker and the environment in which the signal is collected. One of the main challenges in speech processing is to extract features relevant for each application, such as speech recognition, speaker recognition and language identification. Short-time (10-30 ms) spectrum analysis is performed to extract the time-varying spectral envelope characteristics, attributing them to the shape of the vocal tract system. Generally the residual of the speech signal, obtained after removing the spectral envelope information, is considered not useful for many speech applications. But the residual signal contains information, both at the subsegmental (less than a pitch period) level and at the suprasegmental (>100 ms containing several pitch periods in voiced segments) level. The information at the subsegmental level mostly corresponds to the excitation, mainly due to glottal vibration. The information at the suprasegmental level consists of intonation and duration knowledge. Since it is difficult to extract and represent the information at the subsegmental and suprasegmental levels for speech applications, the information present at these two levels are generally ignored. In this paper we show that it is indeed possible to extract the information present at the subsegmental level, and represent it in a manner useful for applications in speaker and language identification.

The main source of of excitation for production of speech is the glottal vibration. In each glottal cycle, the instant of glottal closure is the instant at which significant excitation of vocal tract takes place. Hence a small region (1-5ms) around the instant of glottal closure contains significant informations about the speaker and language, which may be exploited for developing speaker and language identification systems

In Section 2 we describe briefly the tasks of speaker and language identification, and discuss the importance of relevant features for these tasks. In Section 3 the features at the short segment (10-30 ms) level and at the subsegmental (1-5 ms) level are described in the context of Linear Prediction (LP) analysis of speech [1]. In Section 4 the Autoassociative Neural Network (AANN) models used to capture the speaker-specific and language-specific features at the short segment and subsegmental levels are described. In Sections 5 and 6 the speaker and language identification tasks are discussed using these features on a small dataset for each task. In particular, the evidence obtained from the three set of features, namely, spectral, LP residual and the phase of the LP residual are likely to be independent and hence may provide complementary information. All these features are used for speaker and language identification. Section 7 gives conclusions of this study and issues to be addressed to exploit the potential of the features present at different levels for various speech applications.

2. Speaker and Language Identification

Speaker/Language identification is the task of identifying the speaker/language from a set of given speakers/languages using the speaker-specific/language-specific information extracted from the speech signal [2, 3]. Speaker and language identification tasks mainly involve three stages namely, feature extraction, training and testing. Feature extraction deals with extracting speaker-specific and language-specific features from the speech signal. The process of building models from the features is termed as training. The models are tested with the features from the test utterances for identifying the speaker or the language. Performance of the speaker and language identification systems are influenced by all the three stages, namely, feature extraction, model building and testing strategies. The present study focuses on exploring different features for speaker and language identification tasks.

The speaker and language information present in the speech signal may be attributed to the vocal tract dimension, excitation source characteristics and learning habits of the speaker. Apart from the knowledge about the vocal tract, information from the excitation source and learning habits of the speaker are known to be exploited by the humans for identifying speakers and languages. Also it has been observed that humans often can identify the language from a speech signal even when they do not have a good knowledge of that language. This suggest that the listener is able to learn and recognize language-specific patterns derived from the speech signal [4].

The higher level knowledge like vocabulary, lexical structure and phonetic or syllabic statistics generally give information about the language [3]. Features of the vocal tract system and excitation source present in the speech signal may also help in developing a language identification system. Developing such a system is a better choice for practical applications, using it as a front-end for multilingual speech recognizer [5].

3. Features from Speech Signal

Speech is produced as a result of excitation of time-varying vocal tract system with time-varying excitation. The information corresponding to the vocal tract system and the excitation source may be separated approximately from the speech signal using LP analysis [1]. In the LP analysis each sample is predicted as a linear combination of the past p samples, where p is the order of the prediction. The predicted sample of $s(n)$ is given by

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

In Eq.(1), $\{a_k\}$ are termed as LP coefficients, and they are obtained by minimizing the squared error between the actual and predicted samples. This leads to solving a set of normal equations given by

$$\sum_{k=1}^{p} a_k R(n-k) = -R(n), \qquad n = 1, 2, \ldots, p \tag{2}$$

$$R(k) = \sum_{n=0}^{N-(p-1)} s(n)s(n-k), \qquad k = 1, 2, \ldots, p \tag{3}$$

The cepstral coefficients may be derived from the LPCs using the following relations [6]:

$$c_0 = ln\sigma^2 \tag{4}$$

$$c_n = a_n + \sum_{k=1}^{n-1} \left(\frac{k}{n}\right) c_k a_{n-k}, \quad 1 \leq n \leq p \tag{5}$$

$$c_n = \sum_{k=1}^{n-1} \left(\frac{k}{n}\right) c_k a_{n-k}, \quad n > p \tag{6}$$

where σ^2 is the gain term in the LP analysis. The Weighted Linear Prediction Cepstral Coefficients (WLPCC) given by nc_n, are used for representing the spectral features for the speaker and language identification tasks.

The choice of prediction order p is very important from identification point of view. Fig. 1 compares LP log-spectra for different orders of prediction along with short-time speech spectrum. It is clear that low order (p=6) LP analysis captures the gross features of the envelope of speech spectrum. Speaker information may be lost in such a representation, but linguistic information may be preserved. In contrast, a higher order (p= 12) LP analysis captures both the gross and finer details of the envelope of the spectrum, thus preserving both linguistic and speaker-specific information. Hence for implementing a system based on spectral features, lower order LP analysis is preferable for language identification and higher order analysis for speaker identification.

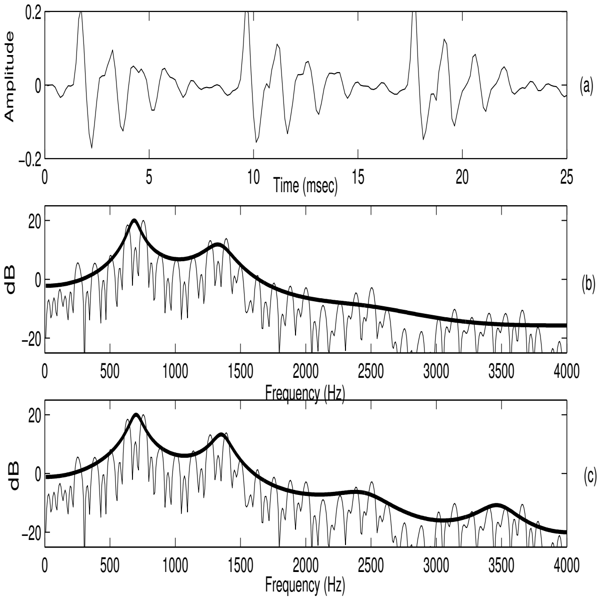

Figure 1: *Comparison of linear prediction analysis with different order of prediction. (a) A voiced region of speech. (b) Short-time spectrum and LP log-spectrum for p=6. (c) Short-time spectrum and LP log-spectrum for p=12.*

The error between the actual and predicted samples is termed as LP residual, and is obtained as

$$r(n) = s(n) - \hat{s}(n) = s(n) + \sum_{k=1}^{p} a_k s(n-k) \tag{7}$$

Since in the LP residual the vocal tract system features are removed, it contains mostly the information about the excitation source [7, 8]. Hence the LP residual is used as a representation of the excitation source of speech production for speaker and language identification studies.

Intuitively we believe that the phase of the analytic signal derived from the Linear Prediction (LP) residual of speech may also contain information about the speaker and the language. In this work we show experimentally that the speaker-specific and language-specific information are present even in the phase of the LP residual of speech. But, extraction of the phase information is a difficult task due to phase warping problem. Also the phase of the speech signal may be degraded due to various factors like background noise and channel effects. The phase information extracted from the LP residual signal is represented as $sin\theta(n)$, and is used to represent the speaker-specific or language-specific information. The knowledge of the Hilbert envelope of the LP residual is used to derive the phase information.

The phase information can be obtained from the LP residual using the knowledge of the Hilbert transform ($r_h(n)$), which is a 90^o phase shifted version of $r(n)$. The Hilbert envelope of the LP residual is computed from $r(n)$ and $r_h(n)$ as [9]

$$h_e(n) = \sqrt{r^2(n) + r_h^2(n)} \tag{8}$$

and $r_h(n)$ is given by

$$e_h(n) = \left\{ \begin{array}{rl} IDFT[-jE(\omega)], & 0 < \omega < \pi \\ IDFT[jE(\omega)], & 0 > \omega > -\pi \\ 0, & \omega = 0, \pi \end{array} \right\} \quad (9)$$

where IDFT is the inverse discrete Fourier transform and $R(\omega)$ is the discrete Fourier transform of r(n). Since Hilbert envelope $h_e(n)$ represents the magnitude information of the LP residual signal, we can obtain the sine of the phase from $r(n)$ by dividing it with $h_e(n)$. Therefore the phase information $(\theta(n))$ is given by

$$sin\theta(n) = r(n)/h_e(n) \quad (10)$$

In this work $sin\theta(n)$ is used to represent the phase information for speaker recognition studies. A segment of voiced speech, the corresponding LP residual, the Hilbert envelope of the LP residual and the $sin\theta(n)$ of the LP residual are shown in Figure 2. Sequence of $sin\theta(n)$ extracted for segments of the vowel /a/ for three different speakers are shown in Figure 3. It is difficult to see any speaker-specific and language-specific features from the phase information . Since in the LP analysis the second order statistical features are extracted, the LP residual and the phase of the LP residual do not contain any significant second order correlations. We conjecture that the speaker-specific and language-specific information may be present in some higher order relations among the samples of the LP residual, and among the samples of $sin\theta(n)$ of the LP residual.

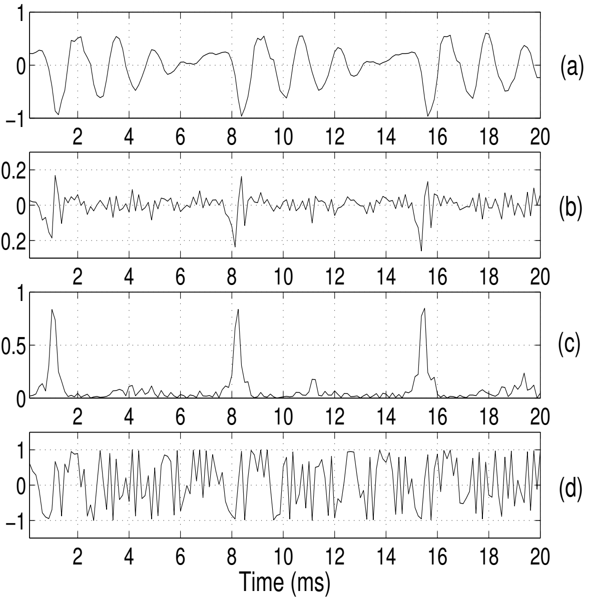

Figure 2: *(a) A segment of voiced speech and its (b) LP residual, (c) Hilbert envelope of the LP residual, and (d) $sin\theta(n)$ derived from the LP residual.*

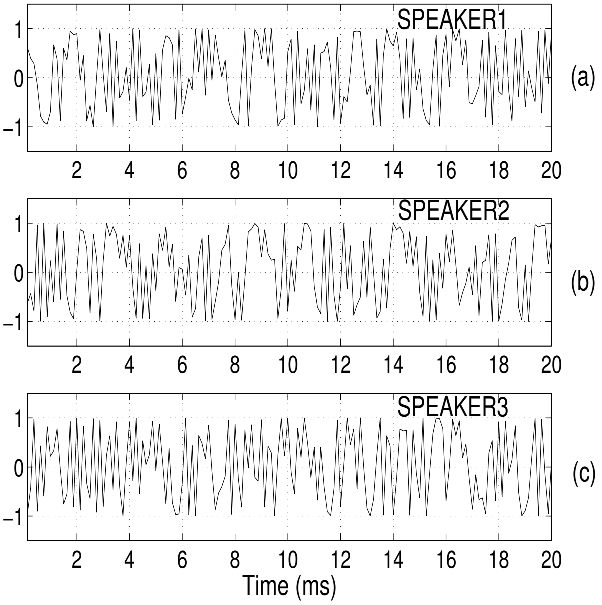

Figure 3: Phase $(\sin\theta(n))$ information from the LP residual for segments of the vowel /a/ for three different speakers.

4. AANN Models for Speaker and Language Identification

AANN is a feedforward neural network which tries to map an input vector onto itself, and hence the name autoassociation or identity mapping [10]. It consists of an input layer, an output layer and one or more hidden layers. A typical structure of a five layer AANN is shown in Figure 4. The number of units in the input and output layers are equal to the size of the input vectors. The number of units in the middle hidden layer is less than the number of units in the input and output layers, and this layer is called the dimension compression hidden layer. The activation function of the units in the input and output layers are linear, where as the activation function of the units in the hidden layers can be either linear or nonlinear.

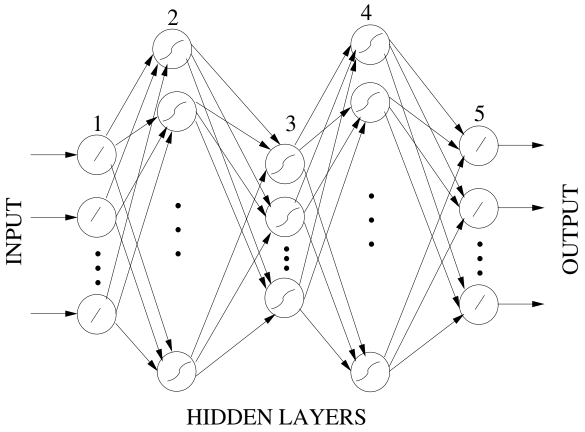

Figure 4: *Structure of five layer AANN model.*

The AANN has been used in the literature in two ways [11]: (a) As a model to capture the distribution of the feature vectors, and (b) as a model to learn the features implicitly from the raw data. However, the way of presenting the input data is different in each of these cases. To capture the distribution of the input feature vectors in the feature space, the feature vectors are extracted from the signal, and are presented in a random order to the AANN. To learn the information in the sequence of samples, the raw samples (normalized) from the signal are presented in a sequential order using blocks of certain size with a shift of one sample between successive blocks.

In the present work the AANN model trained using

WLPCCs are expected to capture the distribution of spectral feature vectors which is unique for a language or a speaker. For the LP residual and phase of the LP residual, the AANN model is used as a nonlinear model to capture the implicit speaker-specific or language-specific information. The LP residual obtained after the removal of spectral features as shown in Figure 2(b). contains informations about the strength of excitation and sequence evidences. The model trained using LP residual will be dominated by the strength of excitation. This strength of excitation is removed in the residual phase as in Figure 2(d) and hence the model built using residual phase as input will capture the sequence information . Thus the three different features used in this work are capable of providing nearly complementary evidences.

5. Speaker Identification Studies

5.1. Database and feature extraction

We have used a subset of speech data taken from the NIST 2003 (one speaker detection) evaluation database. The speech data was collected over telephone and cell phone. Among the total 149 male speakers, 40 speakers are randomly chosen to form two sets MSET1 and MSET2, each of 20 speakers. The training data consists of speech of about two minutes. For each speaker, two genuine trial utterances are randomly chosen as the test data for this study, and the duration of the test data varied from 5-30 seconds. Thus we have four sets of test utterances namely, TSET1 and TSET2 for MSET1, and TSET3 and TSET4 for MSET2. The speech signal sampled at 8 kHz is processed using a 12^{th} order LP analysis, with frames of size 20 ms and a frame shift of 5 ms, to extract the LPCs and the LP residual. Nineteen WLPCCs are computed from the LPCs for each frame. The phase information from the LP residual is obtained using Eq.(10).

5.2. Speaker-specific spectral features

The structure of the AANN model used for capturing the distribution of the speaker-specific spectral features is *19L 38N 4N 38N 19L*, where L represents linear activation function, N represent nonlinear activation function, and the numerals represent the number of units in the layers. One AANN model is trained for each speaker using 200 epochs. During testing, for each WLPCC vector the error between the output of the AANN and the input vector is noted and is converted into confidence value using $C_i = exp(-E_i)$where E_i is the squared error for i^{th} frame. The average confidence is computed as $C = \frac{1}{N}\sum_{i=1}^{N} C_i$, where C_i is frame confidence and N is number of frames in the test utterance. The performance of the speaker recognition system measured in terms of percentage of test utterances identified with first rank and identified with first two ranks for the MSET1 and MSET2 sets are given in Table 1.

5.3. Speaker-specific excitation source information

The structure of the AANN model used is *40L 48N 12N 48N 40L*. Blocks of 40 samples from the high voiced regions are used as input to the AANN. Successive blocks are formed with a shift of one sample. Each block is normalized to the range -1 to 1. One AANN model is trained per speaker using 500 epochs. Since the block size is less than a pitch period, only the characteristics of the excitation source within each glottal pulse are captured. For testing, blocks of 40 samples of the LP residual from the high voiced regions of speech are given as input to the AANN models. The output of each model is compared with its input to compute the squared error for each block. The average confidence value is computed, and the performance for both the MSET1 and MSET2 sets is given in Table 1.

5.4. Speaker-specific phase information

The structure of the AANN model used is *40L 48N 12N 48N 40L*. The phase sequence from the high voiced speech region, and in particular from the regions around the instants of glottal closure, are used for this study. This is because the phase information in the high voiced regions is likely to be more representative of a speaker compared to other regions. Blocks of 40 samples of $sin\theta(n)$ values are given as input values, to train the AANN model. One model is trained for each speaker. Testing is performed as in the case of the LP residual. The performance for the MSET1 and MSET2 sets is given in Table 1. The phase of the LP residual also gives good performance as in the case of the LP residual. We conjecture that in the case of LP residual the strength of excitation dominates the learning process, whereas in the phase of the LP residual, the information in the sequence dominates.

5.5. Combining evidence from different systems

The evidence about the speaker from the three systems may be combined in several ways to achieve better performance. One simple approach is to combine the evidences at the rank level. For instance, in the combined system the speaker is accepted if he has the first rank in at least one system. The result of this combination is given in Table 1. The performance of the combined system is better than the individual systems. Results of the study by considering first two ranks is also given in Table 1. The improvement in the performance by considering first two ranks instead of only first rank indicates that the performance of the system may be improved further by optimization of various parameters of the individual systems.

6. Language Identification Studies

6.1. Database and feature extraction

The database used in this study consists of speech segments excised from continuous speech in broadcast TV news bulletins for four Indian languages namely, Hindi, Kannada, Tamil and Telugu. The training data for each language is of about 200 seconds, obtained by concatenating speech from three male and three female speakers to make the system speaker independent and the test utterances are of 10 seconds duration. The low order LP analysis captures the gross features of the envelope of the speech spectrum. Speaker-specific information may be lost in such a representation, but linguistic information may be present. Hence speech signal sampled at 16 kHz is processed using an 8^{th} order LP analysis with a frame size of 10 ms and a frame shift of 2.5 ms, to extract the LPCs. The 12 WLPCCs are derived from the LPCs for each frame.

6.2. Language-specific spectral features

Distribution of the spectral feature vectors in the feature space is considered to be unique for the speech of a given language. The structure of AANN model used is $12L$ $38N$ $4N$ $38N$ $12L$. The model is trained using the WLPCC vectors for 200 epochs. One model is trained for each language.

Table 1: *Performance of the speaker identification system. The table shows identification accuracy in percentage, identified with first rank and first two ranks.*

Speakers	Feature	# Percentage of tests accurately identified with first rank			# Percentage of tests identified with first two ranks		
		TSET1(20)	TSET2(20)	Total(40)	TSET1(20)	TSET2(20)	Total(40)
MSET1	Spectral	70	75	72.5	70	75	72.5
	Source	55	55	55	65	60	62.5
	Phase	65	55	62.5	70	70	70
	Combined	80	85	82.5	80	95	87.5
MSET2	Spectral	55	40	47.5	65	45	55
	Source	50	30	40	65	45	55
	Phase	30	40	35	55	55	55
	Combined	70	55	62.5	80	70	75

Table 2: *Performance of language identification system for four languages. The entries from columns 2 to 5 represent the percentage of language identification for 40 test utterances from each language.*

Language	Spectral	Source	Phase	Combined
Hindi	100	65	90	100
Kannada	80	52.5	52.5	87.5
Tamil	77.5	77.5	75	80
Telugu	95	72.5	100	100

6.3. Language-specific source features

The excitation sequence may be unique for each sound unit and hence may contain language-specific information. The structure of the AANN used is $40L$ $48N$ $12N$ $48N$ $40L$. Frames of 40 samples of the LP residual (normalized) are used as input to the AANN. Successive blocks are formed with one sample shift. One model is trained for each language using 500 epochs.

6.4. Language-specific phase features

The structure of the AANN model used is *40L 48N 12N 48N 40L*. Blocks of 40 samples of $sin\theta(n)$ values are given as input to train the AANN model. One model is trained for each language using 500 epochs.

6.5. Combining evidence from different systems

During training, three AANN models per language are created, one based on spectral features, one based on source features and another based on residual phase features. While testing, source, system and phase features extracted from the test utterance are given as input to all the AANN models as shown in Figure 5. The average confidence value is computed for the given test utterance for all the languages. The scores of the spectral, source, and phase models of each language are added to get the combined evidence. The language of the model which gives the highest evidence is hypothesized as the language of the test utterance. The performance of different features is given in Table 2. The performance is better when the scores of all the three models are (linearly) combined.

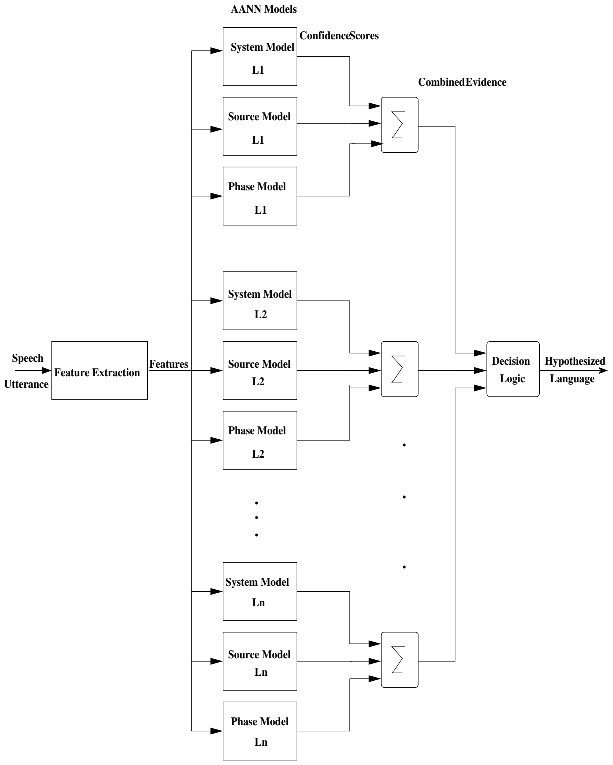

Figure 5: *Block diagram of language identification system based on spectral, source and phase features.*

7. Summary and Conclusions

The goal of this study was to explore the usefulness of features extracted from the segmental and subsegmental levels of the speech signal for speaker and language identification tasks. The presence of speaker-specific and language-specific information in the LP residual and in the phase of the LP residual at subsegmental level was demonstrated using AANN models. We have demonstrated using experimental studies on a small data set that, apart from the spectral features extracted at the segmental level, features extracted at the subsegmental level from the LP residual also contains significant speaker-specific and language-specific information. It is interesting to note that even the phase of the analytic signal derived from the LP residual contains significant speaker-specific and language-specific information.

The effectiveness of these features needs to be examined on a large database. Methods for extracting the speaker-specific and language-specific features at the suprasegmental level of the LP residual, and using them for speaker and language identification tasks also needs to be explored.

8. References

[1] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, pp. 561–580, Apr. 1975.

[2] D. O'Shaughnessy, "Speaker recognition," *IEEE ASSP Magazine*, vol. 3, pp. 4–17, Oct. 1986.

[3] J. Navratil, "Spoken language recognition - a step toward multilinguality in speech processing," *IEEE Trans. Speech Audio Processing*, vol. 9, pp. 678–685, Sept. 2001.

[4] K.-P. Li, "Automatic language identification using syllabic spectral features," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, pp. 297–300, Apr. 1994.

[5] A. Waibel, P. Geutner, L. M. Tomokiyo, T. Schultz, and M. Woszczyna, "Multilinguality in speech and spoken language systems," *Proc. IEEE*, vol. 88, pp. 1297–1313, Aug. 2000.

[6] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, New Jersy: Prentice-Hall, 1993.

[7] P. Thevenaz and H. Hugli, "Usefulness of the LPC-residue in text-independent speaker verification," *Speech Communication*, vol. 17, pp. 145–157, 1995.

[8] F.-Z. M. and D. Rodriguez-Porcheron, "Speaker recognition using residual signal of linear and nonlinear prediction models," in *Proc. Int. Conf. Spoken Language Processing*, Paper No.SL981102, Nov-Dec 1998.

[9] T. V. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction from linear prediction residual for identification of closed glottis interval," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 309–319, Aug. 1979.

[10] S. Haykin, *Neural networks: A comprehensive foundation*. New Jersey: Prentice-Hall Inc., 1999.

[11] B. Yegnanarayana, K. S. Reddy, and S. P. Kishore, "Source and system features for speaker recognition using AANN models," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, (Salt Lake City, Utah, USA), pp. 409–412, May 2001.