Autoassociative Neural Network Models for Language Identification

Leena Mary and B. Yegnanarayana
Speech and Vision Laboratory
Department of Computer Science and Engineering
Indian Institute of Technology Madras, Chennai - 600 036 India
{leena,yegna}@cs.iitm.ernet.in

Abstract

The objective of this paper is to demonstrate the feasibility of automatic language identification (LID) systems, using spectral features. The powerful features of autoassociative neural network models are exploited for capturing the language specific features for developing the language identification system. The nonlinear models capture the complex distribution of spectral vectors in the feature space for developing system parameters. The LID system can be easily extended for more number of languages without any additional higher level linguistic information. Effectiveness of the proposed method is demonstrated for identification of speech utterances from four Indian languages.

1. INTRODUCTION

Automatic language identification (LID) is the task of identifying the language of a digitized speech utterance by a computer. There are numerous applications for LID which fall in two main categories: Pre-processing for machines and pre-processing for human listeners. A multi-lingual voice controlled information retrieval system is an example of the first category. Language identification system used to route an incoming telephone call to a human switchboard operator fluent in the corresponding language is an example of the second category [1].

Multi-lingual interoperability is an important issue for many applications of modern speech technology. In a multi-lingual country like India, automatic LID systems has special significance. The need for development of multi-lingual speech recognizer and spoken dialog systems are very important in Indian scenario. An LID system can be connected as an excellent front-end device for multi-lingual speech recognizers or language translation systems [2].

Human beings and machines use a wide variety of cues to distinguish one language from another. Language differs in phonology, morphology, syntax and prosody. Based on these characteristics the following are some of the approaches for performing LID [1].

1. Spectral similarity approaches
2. Prosody based approaches
3. Phoneme based approaches
4. Word level approaches
5. Continuous speech recognition approaches

It has been observed that human beings often can identify the language of an utterance even when they have no strong linguistic knowledge of that language. This suggest that they are able to learn and recognize language-specific patterns directly from the signal [3]. In the absence of higher level knowledge of a language, a listener presumably relies on lower level constraints such as acoustic-phonetics, syntactics and prosody. Automatic LID can make use of any of the above information or combinations.

Autoassociative neural network (AANN) is a special class of feedforward neural network architecture having some interesting properties which can be exploited for some pattern recognition tasks [4]. It can be used to capture the nonlinear distribution of feature vectors which characterizes each language. As a first step toward language identification, we have built an LID system using autoassociative neural networks to model each language.

This paper is organized as follows: Speech data used in the study and its representation is explained in Section 2. This section also describes the theory behind the selection of feature vectors. In Section 3, we discuss the design and implementation of LID system using AANN models for identification of unknown language. We present our studies on language identification in Section 4. Final section gives the conclusion from this study.

2. SPEECH DATA AND REPRESENTATION

The database consists of speech segments excised from continuous speech in broadcast TV news bulletins for Indian languages. It contains data of four different languages namely, Hindi, Kannada, Tamil and Telugu. Training data for each language were obtained by concatenating speech data of different male and female speakers to make the model speaker-independent. For each language speech data of duration 120 sec is used for training the models. During testing 40

0-7803-8243-9/04/$17.00 © 2004 IEEE

utterances of varying durations are used for each language.

The spectral similarity approach for language identification concentrates on the differences in spectral content among languages. This is for exploiting the fact that speech spoken in different languages contain different phonemes and phones. The training and testing spectra could be used directly as feature vectors, or they could be used instead to compute spectral or cepstral feature vectors.

Linear prediction (LP) analysis of speech signal predicts a given speech sample at an instant n as a linear weighted sum of p previous samples [5]. Equation (1) gives the value of predicted sample at an instant n and (2) represents the difference between the actual sample $s(n)$ and predicted sample value $\hat{s}(n)$.

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

$$e(n) = s(n) - \hat{s}(n) \tag{2}$$

The cepstral coefficients which are the coefficients of the Fourier transform representation of the log-magnitude spectrum have been shown to be a more reliable feature set for speech recognition than LP coefficients [6]. They are obtained from the linear prediction coefficients (LPC) using the following set of recursive relations:

$$c_o = ln(\sigma^2)$$

$$c_m = a_m + \sum_{k=1}^{m-1} (k/m) c_k a_{(m-k)} \qquad 1 \leq m \leq p$$

$$c_m = \sum_{k=1}^{m-1} (k/m) c_k a_{(m-k)} \qquad m > p \tag{3}$$

where σ^2 is the gain term in the LPC model, c_m is the m^{th} cepstral coefficient and a_m is the m^{th} LP coefficient.
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 (4 to 8) 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 (> 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 speaker independent LID system lower order analysis is preferable.

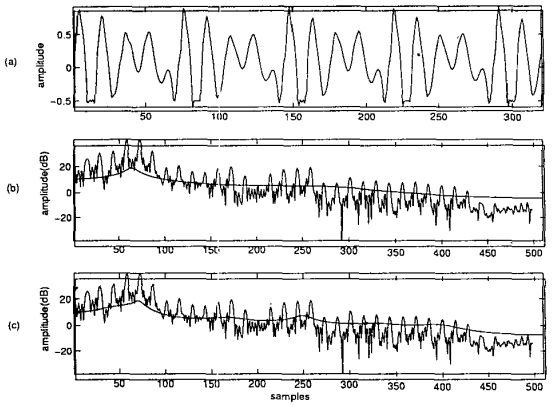

Figure 1. Comparison of linear prediction analysis with different order of prediction. (a) A voiced region of speech. (b) Short-time spectrum and LP log-spectrum for p=6. (c) Short-time spectrum and LP log-spectrum for p=14.

3. AANN MODELS FOR LANGUAGE IDENTIFICATION

AANN models are feedforward neural networks, performing an identity mapping of the input space [7] [8]. From a different prospective the AANN models can be used to capture the distribution of input data [4] [9]. Separate AANN models are used to capture the distribution of feature vectors of each language. A five layer AANN model is shown in Fig. 2. The structure of the AANN model used in the present studies is 12L 38N 4N 38N 12L, where L denotes linear units and N denotes nonlinear units. The activation function of the nonlinear unit is a hyperbolic tangent function. The network is trained using error backpropagation learning algorithm for 60 epochs [7]. The number of epochs was chosen using cross-validation for verification, to obtain the best performance for this data.

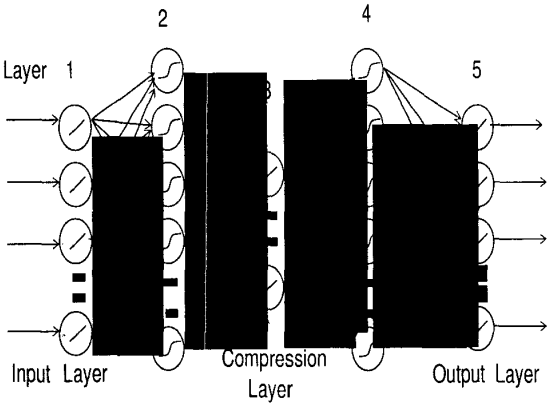

Figure 2. Five layer AANN model.

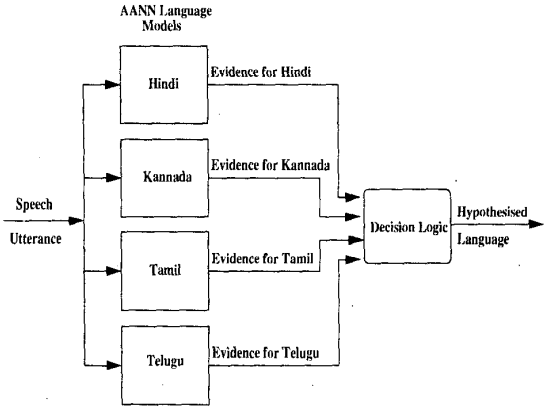

Figure 3. Block diagram of the proposed LID system based on high evidence from different models.

The block diagram of the proposed system for language identification is shown in Fig. 3. It is seen that the system consists of AANN models for each of the languages. The differenced speech signal is segmented into frames of 20 msec using a Hamming window with a shift of 5 msec. The silence frames are removed using an amplitude threshold. An 8^{th} order LP analysis is used to capture the properties of the signal spectrum [5]. The recursive relation between the predictor coefficients and cepstral coefficients is used to convert the 8 LP coefficients into 12 cepstral coefficients as per (1) to (3). The cepstral coefficients for each frame are linearly weighted. The weighted linear prediction cepstral coefficients (WLPCC) feature vectors extracted from the training data of a language are used to train AANN models using backpropagation learning algorithm in the pattern mode [7] [8]. The learning algorithm adjusts weights of the network to minimize the mean squared error obtained for each feature vectors. Once the AANN model is trained, it is used as a language model.

While testing, features extracted from the test utterance are given as input to all the AANN models. The output of each of the model is computed with its input to calculate the squared error for each frame. The error E_i for i^{th} frame is transformed into confidence value by using $C_i = exp(-E_i)$. A given test utterance is passed through each of the language models to obtain the confidence value $C = 1/N \sum_{i=1}^{N} C_i$ for each model, where N is the total number of frames in the test utterance. The model which gives the highest confidence value is hypothesized as the language of test utterance.

Table 1. Performance of AANN based LID system for four languages. Speech data of duration 200 sec is used for training. Testing is done for varying durations. The entries from columns 2 to 4 represent the percentage of language identification.

Language	Duration (sec)		
	10	5	1
Hindi	100	100	95
Kannada	80	77.5	57.5
Tamil	95	90	70
Telugu	100	97.5	90

4. PERFORMANCE EVALUATION OF LID SYSTEM

The performance of the AANN based LID system for is given in Table 1. It is seen that the LID system gives better performance when the duration of the test utterance is sufficiently large. The better recognition of all four languages is primarily due to clean database used where there is no channel or enviornment variability.

The performance of Kannada language is comparatively poor due to the varying recording levels in the database for that language compared to the other languages. Fig. 4 shows the frame-wise confidence score in a case, in which Hindi utterance is tested against Hindi, Kannada, Tamil and Telugu models. Hindi model obviously has the highest score for majority of the frames clearly indicating that it will yield the highest average confidence score than all other language models.

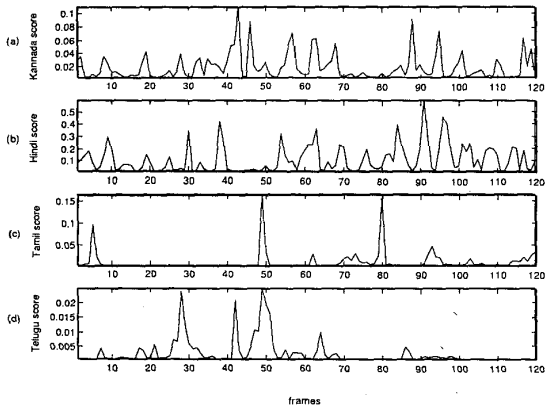

Figure 4. (a)-(d) Frame-wise confidence score (shown only for continuous 120 frames) obtained from Kannada, Hindi, Tamil and Telugu models respectively, when tested against an utterance of Hindi language.

5. SUMMARY AND CONCLUSIONS

In this paper we have shown that the autoassociative neural networks (AANN) can be used for capturing language specific features for developing language identification system. As the AANN models are capable of capturing the underlying distribution of the feature vectors, a language identification system was developed by exploiting this property. For each of the four languages, language models are built by training the AANN models with feature vectors extracted from the speech signal. Effectiveness of the proposed approach is demonstrated for identification of four languages.

Identifying the language spoken within a few seconds of the speech will enable the speech recognizer to adapt the appropriate recognizer for processing that utterance. Literature review shows that improved performance of majority of the LID systems are due to their use of higher level linguistic information and this is achieved by using large corpora of transcribed training speech which may not be available in all languages required by a specific application. In this work, we have shown that it is possible to implement an LID system with reasonably good performance which requires only digitized speech samples in languages to be recognized. Results presented in this paper are promising. The results can be improved by optimizing the model structures. The performance of the LID system can be further enhanced by the use of additional models to capture prosodic features which are yet to be explored.

REFERENCES

[1] M. A. Zissman and K. M. Berkling, "Automatic language identification," Speech Communication, vol. 35, pp. 115–124, 2001.

[2] Alex Waibel, "Multilinguality in speech and spoken systems," Proc. IEEE, vol. 88, no. 8, pp. 1297–1313, Aug. 2000.

[3] Kung-Pu Li, "Automatic language identification using syllabic spectral features," in Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing, Apr. 1994, vol. 1, pp. 297–300.

[4] B. Yegnanarayana and S. P. Kishore, "AANN-An alternative to GMM for pattern recognition," Neural Networks, vol. 15, no. 3, pp. 459–469, Apr. 2002.

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

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

[7] B. Yegnanarayana, Artificial Neural Networks, Prentice-Hall of India, New Delhi, 1999.

[8] S. Haykin, Neural Networks: A Comprehensive Foundation, Prentice-Hall International, New Jersey, 1999.

[9] S. P. Kishore, Speaker Verification using Autoassociative Neural Network Models, MS thesis, Department of Computer Science and Engineering, Indian Institute of Technology Madras, Dec. 2000.