Proceedings of
ICISIP - 2005

Language Identification in Noisy Environments using Throat Microphone Signals

A. Shahina, B. Yegnanarayana
Speech and Vision Laboratory
Department of Computer Science and Engineering
Indian Institute of Technology Madras, Chennai - 600 036, India.
email: {shahina, yegna}@cs.iitm.ernet.in

Abstract

Automatic identification of a language in a noisy environment is a challenging task. Performance of a language identification system depends on the quality of the input speech signal. In the presence of high levels of background noise, speech signals recorded using a close-speaking microphone are degraded. In contrast, a throat microphone picks up high quality speech unaffected by the surrounding noise. This paper explores the possibility of using throat microphone speech signals for text-independent language identification in noisy conditions. Languages are modelled using autoassociative neural networks based on the vocal tract system features and excitation source features derived from the throat speech signal. The results of this study show that the throat microphone speech-based language identification system performs well in noisy environments.

1. Introduction

Automatic language identification (LID) is the task of identifying the language of a spoken utterance. Language identification systems have many applications like helping telephone companies in handling foreign language calls, serving as a front-end device for a multi-lingual speech recognizer, multi-language speech translation systems among others [1] [2]. Generally, language identification studies use a database comprising of uncorrupted close-speaking microphone speech or telephone speech. In certain applications, such as handling calls from non-native speakers in noisy environments, the LID systems may not perform as desired. This is because the input speech to the system is corrupted by noise. Under noisy conditions, satisfactory performance of LID systems could be achieved if clean speech signals were used.

A throat microphone is a device that records clean speech even in the presence of high background noise. It is a transducer that is placed in contact with the skin surrounding the larynx. It picks up speech signals transmitted through the skin. The intelligibility of such a speech is high, almost similar to normal speech. In noisy environments, a close-speaking microphone speech is degraded since it picks up the background noise along with the speech. A throat mirophone does not pick vibrations in the air and hence the throat microphone speech is relatively unaffected by environmental distortions.

When clean throat microphone speech signals are used for LID applications in noisy environments, the systems are expected to perform much better than systems with noisy input speech signals.

This paper is organised as follows: Section 2 briefly mentions the acoustic analysis of the throat microphone speech. Section 3 details the language identification study conducted. The database and features used and the training of language models is discussed there. The performance of the LID systems in clean and noisy suroundings are discussed in Section 4.

2. The Throat Microphone Speech Signal

Perceptually, the throat microphone speech is intelligible. The waveforms of a speech segment recorded simultaneously using a throat microphone and a close-speaking microphone is shown in fig.1. The gross envelop of the two waveforms are similar. However, a study of the throat microphone speech signal reveals that some of the higher frequencies are missing. This is observed in the case of vowels, semi-vowels and fricatives. We also observe presence of formant structures associated with the articulation of voiced consonants, unlike in the case of close-speaking microphone signal where only faint voicing striations are seen near the baseline. The formant structure associated with the articulation of nasal consonants are seen as dark bands in the case of throat microphone speech while they are fainter in the case of close-speaking microphone speech, though the locations of the formants are the same. In fricatives, the energy distribution in the higher frequencies is absent in the case of throat microphone speech signal [3].

From the above discussion, we know that some of the spectral components are missing in the throat microphone speech signal. However, since most of the information is preserved in the signal, the acoustic cues pertaining to the language would also be preserved in the signal. The goal is to model language-distinguishing information from the throat microphone signal. The differences between languages could be at various levels like at frame-level, phoneme-level, syllable-level, word-level and sentence-level. The differencies could be due to phonetic inventories, frequency of occurence

0-7803-8840-2/05/$20.00 ©2005 IEEE

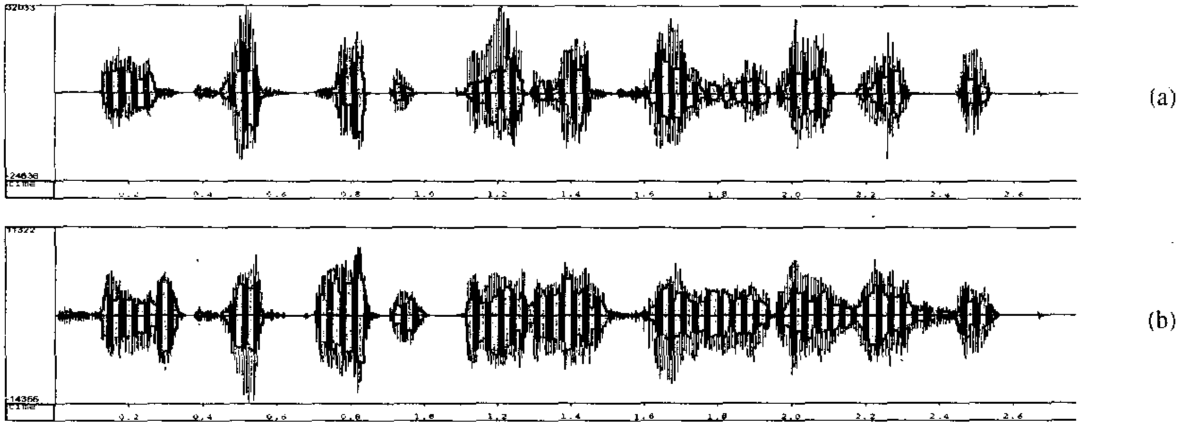

Fig. 1: The waveforms of a speech signal from a male speaker recorded simultaneously using a close speaking microphone and a throat microphone ((a) and (b) respectively) for the sentence *don't ask me to carry an oily rag like that.*

of different phones, phonotactics and duration and intonation of words in the language [4]. In this paper, we investigate the presence of language information in the system and source features derived from the throat microphone speech signal.

The vocal tract system and the excitation source features are extracted from the speech signal using Linear Prediction (LP) analysis [5]. In LP analysis, the n^{th} speech sample is predicted as a linear weighted sum of the past p samples, and is given by:

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

where $\{a_k\}$ are the LP coefficients and p is the order of prediction.

A lower order LP analysis (6^{th} or 8^{th} order) captures the gross features of the envelope of the speech spectrum, whereas a higher order LP analysis (12^{th} or 14^{th}) captures the finer details of the envelope. The shape of the spectral envelope is due to the formants which reflect the resonances of the vocal tract. The formant locations and bandwidths vary for different speakers, even for the same category of sound unit. This is due to the variations in vocal tract shapes and lengths among different speakers. These variations are more pronounced in the finer fluctuations of the spectral envelope, as compared to the gross spectral envelope. In contrast, the gross spectral envel is similar for different speakers due to the underlying sound unit. Language-specific information may be better represented while speaker-specific information may be poorly represented in lower order LP analysis [6]. We use a lower order LP analysis in our study, since we are interested in language-specific information independent of the speaker. The LP coefficients (representing the vocal tract characteristics) are converted to 12 linearly weighted LP cepstral coefficients (WLPCC) using the following relations [7] :

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

$$c_m = -a_m - \sum_{k=1}^{m-1} \frac{k}{m} c_k a_{m-k} \quad 1 \leq \mathrm{m} \leq \mathrm{p} \qquad (3)$$

$$c_m = -\sum_{k=1}^{m-1} \frac{k}{m} c_k a_{m-k} \quad \mathrm{m} > \mathrm{p} \qquad (4)$$

where σ^2 is the gain term in the LPC model, $\{c_m\}$ and $\{a_m\}$ are the LPCC and LP coefficients respectively.

The difference between the actual value $s(n)$ and the predicted value $\hat{s}(n)$ of the n^{th} speech sample is the prediction error, known as the LP residual and given by:

$$e(n) = s(n) - \hat{s}(n) \qquad (5)$$

The 12 linearly weighted cepstral coefficients (WLPCCs) corresponding to each speech frame are used as the system features. The LP residual obtained after removing the vocal tract system features mostly contains information about the excitation source. Hence, the LP residual is used for extracting language-specific source features for this study.

3. Identification Approach

A. Multi-language speech corpus

The corpus comprises of recordings from 80 native speakers in four Indian languages, namely, Hindi, Kannada, Telugu and Tamil. All the languages belong to the same family of languages and share a common set of phonemes. The confusability among them is likely to be high. The recordings were carried out in the laboratory under clean and simulated noisy conditions. The noisy conditions were simulated using radio static. Speech from the volunteers was collected simultaneously using an unidirectional close-speaking microphone and a throat microphone. The signal-to-noise ratio (SNR) of the noisy close-speaking microphone speech signal is 20dB. The SNR of the throat microphone speech is similar both in clean and noisy conditions. Text-independent speech was used in the study. The speech was sampled at 8 kHz. Features are extracted using an 8^{th} order LP analysis performed on overlapping Hamming windowed speech frames of 20 msec duration taken with a frame shift of 5 msec duration [8].

B. Models for language classification

This approach models an entire language by a single neural network model. The potential of artificial neural networks as nonlinear models is explored to capture the characteristics of the vectors unique to a language from the given training data [9][10]. Specifically, autoassociative neural network models, which are feedforward neural networks that perform the task of autoassociation are used [11]. It has been shown that AANN models capture the distribution of the feature vectors in the high dimensional space. The AANN consists of an input layer, an output layer and one or more hidden layers. Typical structure of a five layer AANN used in this study is shown in Figure 2.

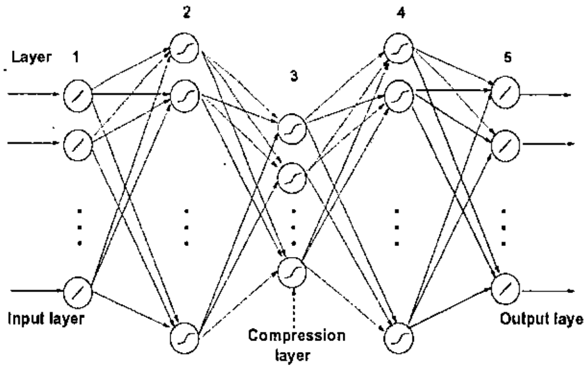

Fig. 2: Five layer AANN model.

The number of units in the input and output layers is equal to the size of the input vector. The hidden layer has lesser number of units and is called the dimension compression layer. The activation functions of the input and output layers are linear, while the activation functions of the hidden layers are linear or nonlinear. In this study, AANN is used to capture the distribution of language-specific spectral and source feature vectors.

C. Identification using system features

The feature vectors given as input to train each language model are obtained by concatenating the feature vectors obtained from 6 speakers (3 male and 3 female). This is done so that the model captures the variability within the language such as, gender differences and content. Around 30 seconds of speech data from each speaker is used for training. The structure of the neural network used to capture the distribution of vocal tract system feature vectors in the feature space is *12L 38N 4N 38N 12L* where L refers to a linear unit, N to a nonlinear unit and the numbers represent the number of nodes in a layer. The 12-dimension system features (WLPCC) obtained for each speaker are given to the AANN in a randomized fashion. Each AANN is trained using 200 epochs. Two separate models for each language are obtained by training two AANN models using WLPCCs derived from clean and noisy speech. If a system is to perform accurately under noisy conditions, it needs to be trained on speech recorded under these conditions. Hence, noisy speech models are used. During testing, features derived from utterances of 20 secs duration, are used to compute confidence scores against models of each language and to identify the language of that utterance. The test speech utterances are different from those used for training.

D. Identification using source features

The structure of the AANN used is *20L 40N 10N 40N 20L*. The LP residual, that represents the excitation source features is down-sampled to 4 kHz sampling frequency to emphasize only those regions with high signal-to-noise ratio are used. Blocks of 20 samples (5 msec) of the normalized LP residual, with a shift of one sample, are applied in succession to train the model. A model is trained using 500 epochs. Two models for each language are obtained using the LP residual derived from clean and noisy speech.

During testing, the error between the output of the AANN and the input is used to compute the confidence score C_i of that frame which is inversely proportional to the error. The average confidence scores of all the N frames is obtained as $C = 1/N \sum_{i=1}^{N} C_i$, where N is the total number of frames/block. The average confidence scores of the test utterance against all the models are compared to evaluate the performance of the LID systems. The performance is evaluated in terms of percentage of the number of test utterances accepted out of the total test utterances used for this study.

4. Results and Discussion

The performances of the language identification systems based on the system and source features derived from the clean speech obtained from throat and close-speaking microphones are given in Tables 1 and 2 respectively. We see that the performance of the language identification system using the throat microphone speech is similar to that using the close-speaking microphone speech for both system as well as source features. As the scores obtained for both the system and source feature-based models form independent evidences, the two scores can be combined. The combination logic is a simple addition of the two scores to improve the performance of the LID system. The block diagram depicting the LID system based on combined logic is shown in Figure. 3.

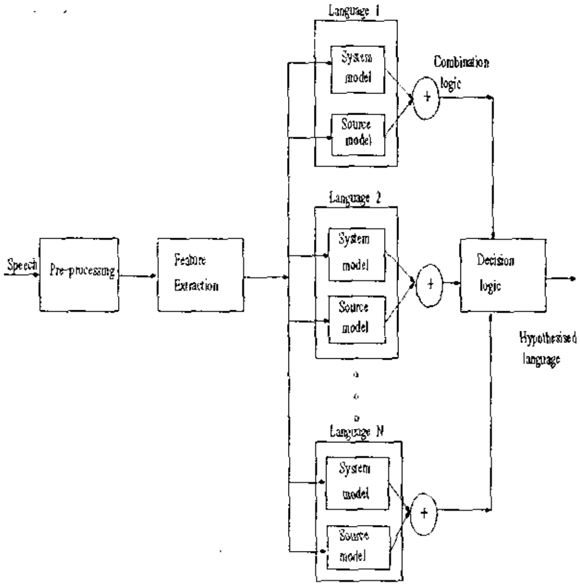

Fig. 3: Block diagram of the LID system using combined evidences.

Tables 3 and 4 show the performance of the language identification systems, where the language models trained using noisy speech are tested against noisy utterances. We observe that the performance of the throat microphone-based language identification systems is similar to that of clean speech based language identification systems (refer Table 1 and 3). But, the performance of the close-speaking microphone-based language identification systems degrades. This is due to the presence of significant noise levels in the speech.

From the studies conducted, we can infer that the vocal tract system and excitation source features extracted from the throat microphone speech contains significant information to distinguish languages. The performance of the identification systems using throat and close-speaking microphone speech signals, when the signals are recorded under noise-free conditions are almost similar. However, under noisy conditions, the performance of the identification system using throat speech is similar to that under noise-free conditions, while that using close-speaking microphone speech degrades.

TABLE 1: PERFORMANCE (IN %) OF THE LANGUAGE IDENTIFICATION SYSTEMS BASED ON SOURCE AND SYSTEM FEATURES OBTAINED FROM SPEECH SIGNALS RECORDED USING A THROAT MICROPHONE IN A CLEAN ENVIRONMENT

Language	System features	Source features	Combined system
Hindi	90	69	92
Tamil	95	74	95
Telugu	92	70	96
Kannada	95	72.5	95

TABLE 2: PERFORMANCE (IN %) OF THE LANGUAGE IDENTIFICATION SYSTEMS BASED ON SOURCE AND SYSTEM FEATURES OBTAINED FROM SPEECH SIGNALS RECORDED USING A CLOSE SPEAKING MICROPHONE IN A CLEAN ENVIRONMENT

Language	System features	Source features	Combined system
Hindi	87	71	91.5
Tamil	90	67	93
Telugu	96	77.5	100
Kannada	100	72	100

TABLE 3: PERFORMANCE (IN %) OF THE LANGUAGE IDENTIFICATION SYSTEMS BASED ON SOURCE AND SYSTEM FEATURES OBTAINED FROM SPEECH SIGNALS RECORDED USING A THROAT MICROPHONE UNDER NOISY CONDITIONS

Language	System features	Source features	Combined system
Hindi	89	67.5	91
Tamil	90	76.5	94
Telugu	85	69	93.5
Kannada	92	75	92.5

TABLE 4: PERFORMANCE (IN %) OF THE LANGUAGE IDENTIFICATION SYSTEMS BASED ON SOURCE AND SYSTEM FEATURES OBTAINED FROM SPEECH SIGNALS RECORDED USING A CLOSE SPEAKING MICROPHONE UNDER NOISY CONDITIONS

Language	System features	Source features	Combined system
Hindi	57	56.5	57.5
Tamil	61	59	61
Telugu	45	52.5	51
Kannada	66.5	62	71.5

REFERENCES

[1] Y. K. Muthusamy, E. Bardnard, and R. A. Cole, "Reviewing automatic language identification," *IEEE Signal Processing Magazine*, vol. 11, pp. 33–41, Oct. 1994.

[2] M. A. Zissman, "Overview of current techniques for automatic language identification of speech," in *IEEE Intl Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 60–62, Dec. 1995.

[3] A. Shahina, B. Yegnanarayana, and M. R. Kesheorey, "Throat microphone signal for speaker recognition," in *Proc. Int. Conf. Spoken Language Processing*, (Jeju Island, Korea), Oct. 2004.

[4] T. Nagarajan, "Implicit systems for spoken language identification," PhD thesis, Dept. Comp. Sci. Engg, Indian Institute of Technology, Madras, Jan. 2004.

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

[6] S. Guruprasad, "Exploring features and scoring methods for speaker verification," MS thesis, Dept. Comp. Sci. Engg, Indian Institute of Technology, Madras, Jul. 2004.

[7] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice Hall, 1993.

[8] L. Mary, K. S. R. Murthy, S. R. M. Prasanna, and B. Yegnanarayana, "Features for speaker and language identification," in *Proc. ODYSSEY - The Speaker and Language Recognition Workshop*, (Toledo, Spain), June 2004.

[9] B. Yegnanarayana, *Artificial Neural Networks*. Connaught Circle, New Delhi: Prentice Hall India, 1999.

[10] S. Haykin, *Neural Networks: A Comprehensive Foundation*. NJ: Prentice-Hall International, 1999.

[11] B. Yegnanarayana and S. P. Kishore, "AANN: An alternative to GMM for pattern recognition," *Neural Networks*, vol. 15, pp. 459–469, Apr. 2002.