LANGUAGE IDENTIFICATION FROM SHORT SEGMENTS OF SPEECH

Jyotsana Balleda, Hema A Murthy and T.Nagarajan

Department of Computer Science and Engineering,
Indian Institute of Technology, Madras, Chennai - 600 036

e-mail : hema@lantana.iitm.ernet.in

ABSTRACT

Automatic language identification (LID) from the spoken speech utterance is a challenging problem. In this paper, we present an LID system that works for South Indian languages and Hindi. Each language is modeled using an approach based on Vector Quantisation [1]. The speech is segmented into different sounds (CVs) and the performance of the system on each of the segments is studied. Our studies indicate that the presence of some CVs is crucial for each language. We also find that for the same Consonant and Vowel (CV) combination, the quality of the sound is different in different languages. We show that once the speech signal is segmented into CVs, it is possible to perform LID on very short segments (100-150ms) of speech itself.

1. INTRODUCTION

Automatic Language Identification (LID) is the task of identifying the language being spoken from a short duration of speech by an unknown speaker. Just as in the case of most natural input systems, humans are the best LID systems. The purpose for the development of LID systems today are several. Particularly, in a country like India where a multitude of languages exist, once the language of a person is determined, better services can be provided. In the Indian context, except for the top 3-5% of the population, the rest can hardly speak English. Although Hindi is the national language, it is spoken by hardly 25% of the population. So most people who belong to the other 95% or 75% get marginalised in Indian society. The advent of the automated systems in most public places has further marginalised 95% of the population as they only cater to the English speaking elite. It is not always possible to have trained human operators who can assist in different Indian languages.

An automatic LID system can be extremely useful in a public place to provide assistance. Once the language of the speaker is identified future responses from the system can be made in the language of the speaker itself. Also LID systems can be used as a front-end in language translation systems.

The focus of the research in this paper deals with the development of a LID system to classify South Indian languages and Hindi. A *Vector Quantization* (VQ) based model is used to perform LID. In Section 2, we discuss the challenges in LID, and in Section 3, we discuss the database that was used in the study. Section 4 explains the model that was used in the development of the LID system. Section 5 evaluates the performance of our system for different languages. Section 5 also gives a linguistic analysis of the languages used in the study. Section 6 gives the short segmental analysis on the labelled data. Although the results that we have obtained (average of 84% for all the five Indian languages) are encouraging, we do feel that the performance can be further improved by using language specific information. In Section 7, we discuss the results and suggest directions in which the research must progress to improve the performance of the LID system.

2. THE LANGUAGE IDENTIFICATION PROBLEM

This paper presents an automatic LID system for Indian languages. Indian languages have a significant number of sounds that are common. Some languages have special sounds in addition, for example Tamil. Further, languages are subject to dialectical variations.

To cover a significant number of languages that are commonly spoken, one is confronted with two problems: the problem of devising a system efficient enough to identify the language, given a short duration of the speech signal and the problem of collecting the data needed to train and test such a system [2].

The best systems use multiple large vocabulary continuous speech recognizers [3]. These systems use word or sentence level modeling which restrict the ad-

dition of a new language. And such systems are further limited to a small vocabulary.

In this work, we present a system that uses a minimum amount of linguistic information and can be extended to all Indian languages. This paper uses VQ based modelling for each of the languages. During training separate codebooks were generated for each of the languages using a particular feature. During testing, the same features that were used in the training were extracted from the test utterance and compared with the codebooks obtained during training for each of the languages and a decision is made.

3. THE DATABASE USED IN THE STUDY

The system is developed for five Indian languages. The database consists of four South Indian languages belonging to the Dravidian group, namely, Kannada, Malayalam, Tamil, Telugu and one North Indian language belonging to the Aryan group, namely, Hindi.

The database consists of speech utterances from read text. For each language, data has been collected from five native male and five native female speakers. It was further ensured that the text read by different speakers were different. The text of the sentences were chosen randomly. No effort was made to choose a phonetically balanced set of sentences. The training utterances have an average length of ten seconds and the testing utterances have an average length of five seconds.

For each speaker, 60 seconds of data was collected. It was further ensured that the text sentences used in training and testing were different.

4. BASELINE SYSTEM

The language identification system used as a baseline for this paper consists of VQ based modeling for the languages for which the training data is available, along with the classifier that evaluates the minimum distortion of the unknown speech segments with respect to these models. The classifier hypothesises the identification of an unknown utterance by determining which model reduces the distortion for the test utterance. In both training and testing, the speech waveforms are digitized and preprocessed by a front-end interface that extracts a set of filter-bank based cepstral coefficients from each frame of data. The parameters that control the front-end processing (e.g. frequency range, filter shape, filter resolution) are set *a priori* and can be modified if desired. During the training phase, feature vectors extracted from the speech signal for a given language are clustered in the feature space using the k-means algorithm [1] [4]. The VQ codebook size is 64 and the length of the feature vector is 22. This process is repeated for all the languages.

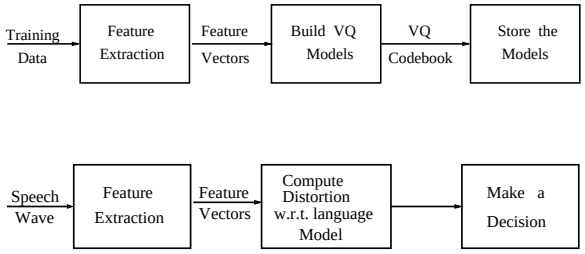


Figure 1: Baseline language identification system

During the testing phase, features extracted from the test utterance are clustered using the VQ codebook which were obtained during the training for each of the language models. The classifier minimises the distortion of the test utterance and the language is identified (Figure 1). The classifier minimises the measure using the Euclidean distortion measure. Data of three minutes duration is used for training and data of different durations is used for testing the performance of the system.

5. PERFORMANCE EVALUATION

The performance of the language identification system for five seconds test utterances is given in Table 1. The rows refer to the utterance of the language being tested and the columns refer to the language that is identified. All figures in the Table correspond to percentages. A total of 42 utterances of each five seconds duration, 240 sentences of two seconds duration and average of 230 sentences of one second duration were taken for each language for evaluating the system. Table 1 shows the performance of the system for five seconds test utterances. From Table 1 it is clear that there are significant differences in the acoustic signatures of the various languages. This is the reason why a VQ based model performs quite well. Also since the test utterance is very long (five seconds), it is possible that the acoustic signatures of a given language is completely available in the test utterance. Five seconds is normally very long for any identification system. We therefore reduced the length of the test utterances to two seconds and one second to evaluate the performance of the system. The utterances were chosen arbitrarily where the utterance does not correspond to a complete sentence. The performance of the system for different durations of the test utterances was studied [5] and it was found that the degradation was graceful.

It is well known that the linguistic content of different languages are also different. We therefore per-

Table 1: Performance evaluation of the language identification for five seconds of the test utterances. The performance is given as percentage of total number of test utterances in each of the languages.

Language Given	Language Identified (%)				
	Telugu	Tamil	Malayalam	Kannada	Hindi
Telugu	83.34	11.90	0.00	0.00	4.76
Tamil	2.38	88.09	0.00	0.00	9.53
Malayalam	11.90	7.14	76.20	0.00	4.76
Kannada	11.90	2.38	0.00	76.20	9.52
Hindi	0.00	2.38	2.38	0.00	95.24

formed a study based on linguistic criteria. The text of the sentences used in our study were first transcribed into a sequence of sounds. The basic unit CV has proved promising for continuous speech recognition for Indian languages [6], we therefore used this unit as the unit for the transcription. Using the basic C and V level transcription, the distribution of Vowels and consonants and distribution of consonants based on the place of articulation and manner of articulation were analysed for the languages considered for this study. From the analyses done in [5,ch.5], it was observed that eventhough there is a only a small variation in the distribution of vowels and consonants, there is considerable difference in the frequency distribution of consonants based on the place of articulation and the manner of articulation [5,ch.5].

6. ACOUSTIC ANALYSIS

To corroborate the evidence shown by the linguistic analysis, as a next step, the acoustic analysis was performed on the data.

To perform short segmental analysis, the speech data is manually segmented and labelled into the basic units. The labelled data is used and language identification is done on this small segment of the labelled data. The performance of the system is tested on a five second utterance of hindi by combining the evidences at the segmental level. Each labelled segment was tested against the trained models. The results are indicated in the (Figure 2), where 'Te' is for Telugu, 'Ta is for Tamil, 'M' is for Malayalam, 'K' is for kannada and 'H' is for Hindi as languages identified. Clearly, the overall evidence indicates that the language is Hindi. Table 2 gives the list of sounds that have contributed to the performance of the system. Although the sounds are common across different languages, they are acoustically different. Table 3 shows the confusion matrix of sound /da/. The rows refer to the sound of the language being tested and the columns refer to the language that is identified based on the sound. The diagonal elements represent the sounds being correctly identified. The sounds are clustered around already trained statistical models using three minutes of data. The system is tested with 100-150ms of data and decision is made. We can clearly see the /da/ sound contributing more than 50% to Telugu, Tamil, Kannada and Malayalam languages identification. Similarly, there are sounds that are unique to specific language as listed in Table 2.

Table 2: Sounds that contributed to the performance of our system.

Language	Sounds that contributed for performance
Telugu	ko va ja sa da ga ra A ka la na si ya a ni du n ta
Tamil	O rai wwu yai zi mu wwa pa o ttu vu li du ke tti u m tta di ru ri ya pe ko pa xu la e i da ni xi xa
Malayalam	lla me LLA a de la li ya Ra La kku n na mu ma
Kannada	lle lli mu tra vi ki du ma e va ni ke lla re tu nnu pa li gi la ge ri ya ga ba A da ra ta ja ru sa
Hindi	au A ni di ca ke ma ki da ko ga hai hi u a ka ri se me la n pa

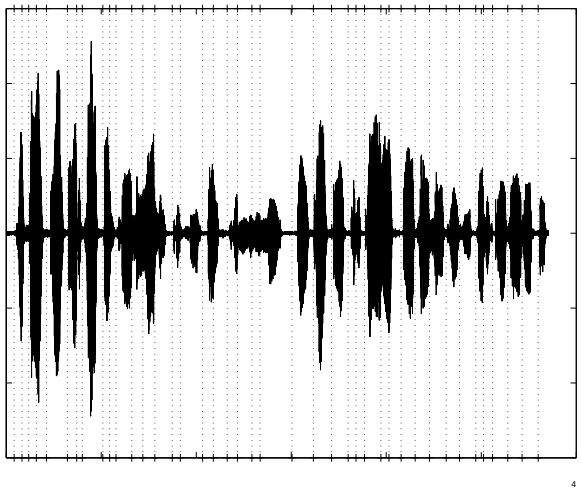
Table 3: Confusion matrix of /da/ sound

Language Given	Language Identified (%)				
	Telugu	Tamil	Malayalam	Kannada	Hindi
Telugu	58.54	14.63	4.88	12.76	9.76
Tamil	18.00	52.00	8.00	8.00	14.00
Malayalam	21.43	21.43	35.71	0.00	21.43
Kannada	3.64	9.09	12.73	56.36	18.18
Hindi	6.67	26.67	13.34	0.00	53.39

Table 4 gives the individual language performance of a system taking decision at the segmental level. The performance is still graceful.

7. CONCLUSIONS

This paper presents a language identification system for Indian languages. Five languages were chosen for



The sounds are hru da ya ko u r ja po n ca ne va lo ki sI e ka Da m nI me e kA e ka ru kA va ta A jA ne se hi di l ka do ra va

The recognition of the sounds are H Ta H H M M H H H H M M H H H Ta H M H H H H H H Ta Ta H H Te H Te M H H H Te H H H M

Figure 2: Evidences from labelled data of complete utterrance

this system. The results presented in this paper are quite promising. The performance of the system for Hindi is much better than that of the other languages. This is because amongst the group of languages chosen, Hindi is the only language that belongs to the Aryan group. The results reported in Section 6 do indicate that the acoustic manifestation of same linguistic unit is different in different languages. languages exist at both acoustic and linguistic levels. Further in Section 6, we observe that there are certain sounds that contribute to the performance of LID system. If automatic labelling and segmentation is done, key sound spotting can be done and language identification can be done with shorter utterances. We observe that even if the sounds are common across the languages, the quality of the sound is different (Table 3). The model used in this paper is borrowed from speech recognition [4], and we need to explore new models to capture the language specific features [7].

Table 4:
Performance of the system based on the segmental decision

Language	Performance (%)
Telugu	70.84
Tamil	83.34
Malayalam	61.90
Kannada	88.10
Hindi	79.17

8. REFERENCES

[1] R. M. Gray, "Vector Quantization," *IEEE ASSP Magazine*, pp. 4–29, April 1984.

[2] R. A. Cole, Y. K. Muthusamy, and B. T. Oshika, "The OGI multi-language telephone speech corpus," *In Proc. International Conference on Spoken Language Processing*, pp. 895–898, October 1992.

[3] T. Schultz, I. Rogine, and A. Weibel, "LVCSR - based language identification," *Proc. ICASSP Atlanta USA*, 1996.

[4] X. D. Huang, Y. Ariki, M. A. Jack , *Hidden Markov Models for Speech Recogntion.* Edinburg University press.

[5] Jyotsana Balleda, *A Spoken Language Identification System for Indian Languages.* MS dissertation, Indian Institute of Technology, Madras, 2000.

[6] C. Chandra Sekhar, *Neural Network Models for Recognition of Stop Consonant-Vowel (SCV) Segments in Continuous Speech.* PhD thesis, Indian Institute of Technology, Madras, 1996.

[7] Jyotsana Balleda, S.Vela and Hema A Murthy, "A new model for speaker identification using affine transformations," *Proc.SPCOM - 99, IISc, Bangalore,India*, pp. 121–124, July 1999.