Language Identification Using Spectral Vector Distribution Across the Languages

T.Nagarajan and Hema A. Murthy
Dept. of Computer Science and Engineering,
Indian Institute of Technology, Madras,Chennai - 600 036, India.
Email: *hema@lantana.iitm.ernet.in*

Abstract

Automatic spoken language identification is the task of identifying the language from a short duration of a digitized speech signal. One of the important language identification cues is the differences in phoneme frequencies among different languages. Instead of explicitly using phonemic differences between languages, we develop a language identification system that only uses the differences in "usefulness of a language independent set of spectral vectors". Three different methods are developed to define the usefulness measure using (i) a common set of spectral vectors and (ii) pairwise common set of spectral vectors. The goal is to ultimately derive a language model from the given speech signal itself. The performance of each method is analyzed for different durations of speech signals and it is shown that even for 0.8sec of test speech utterances the performance is quite promising. If a separate set of spectral vectors for each pair of languages is used, the performance of the system is considerably improved.

1 Introduction

Language identification (LID) systems fall into two main categories based on the way with which languages are modeled : Explicit LID systems and Implicit LID systems. Systems that require segmented and labeled speech data with transcribed text can be considered as "Explicit LID systems". And, the systems that require only the digitized speech signal and the corresponding true identities of the languages being spoken are termed as "Implicit LID systems", in which language models are derived only from the speech signal.

There are variety of sources of information that the humans can use to distinguish one language from another. One of the important LID cues is the differences in phoneme frequencies among different languages [Muthusamy *et al.*, 1994]. But to incorporate this information into the LID system, there should be either language-dependent phone recognizers of all the languages under consideration or

at-least one language independent phone- recognizer as front end. Such phone recognizers, segmented and labeled speech data is required. Further in such systems, including a new language is also not a trivial task. In this paper, the focus is on the development of an Implicit LID system where an attempt is made to derive the language model from the digitized speech signal itself. [Foil, 1986] and [Goodman *et al.*, 1989] exploited the frequency of occurrence of voiced sounds using formant locations to represent sounds and using Vector Quantization (VQ) distortion measure as the basis for language decisions. [Zissman, 1993] used a simple GMM classifier where the decision is made by calculating log-likelihood that a language model produced the unknown speech utterance. [Sugiyama, 1991] and [Balleda, 2000] performed VQ classification on different feature sets where the decision is based on accumulated distortion. [Sugiyama, 1991] has further shown the possibility of classifying the languages based on their VQ histogram patterns. These methods do not directly use the variation in spectral vector distributions of different languages. [Tucker *et al.*, 1994] has efficiently used the variation in phoneme frequency distribution by using the log-likelihood ratio, but a phone recognizer is required as a front-end for language identification.

In our work, it is assumed that if the phoneme distributions are different among different languages, the distributions of the spectral vectors, derived from speech signal, should also be different. With this assumption, a simple LID system has been developed, based on vector quantization, for three Indian languages namely Tamil, Telugu and Hindi.

2 System Description

2.1 Speech Corpus

For both training and testing, Indian television news bulletins have been used [DD News, 2001]. The database consists of news bulletins in eight different languages. Presently, only the news bulletins of three languages namely Tamil, Telugu and Hindi are considered since the database of other languages are yet to be prepared. During training, for each language, 2 male and 2 female speakers data, each of 15mins duration are used. During testing, 16 news bulletins (both male and female speakers) are used for each language. The training and testing set speakers are different. Here, the total duration of the speech signal of each news bulletin is split into small segments of approximately 2.5s each resulting in total of 4500 test segments.

2.2 Common Codebook based methods

The system consists primarily of four steps: Feature extraction, K-means clustering, usefulness calculation and language identification. In the first step, from the speech files of all the languages under consideration, 13 dimensional Mel Frequency Cepstral Coefficients are extracted (excluding the zeroth coefficient since it contains only the overall energy level information) from all the frames, with frame size of 16ms and frame-shift of 8ms. K-means clustering algorithm is used to cluster all the feature vectors, producing "N" cluster centers, where N is the code book size. In this work, the number of cluster centers is chosen as 64, which is approximately the union of phonemes in all the languages under consideration. Using this language-independent code book, as a next step, speech files corresponding to each language, L_i, is coded separately and the probability of occurrence, $p(s_k/L_i)$ of each

code book index is found out. The histograms of the codebook indices for three languages namely, Tamil, Telugu and Hindi is shown in Fig.1.

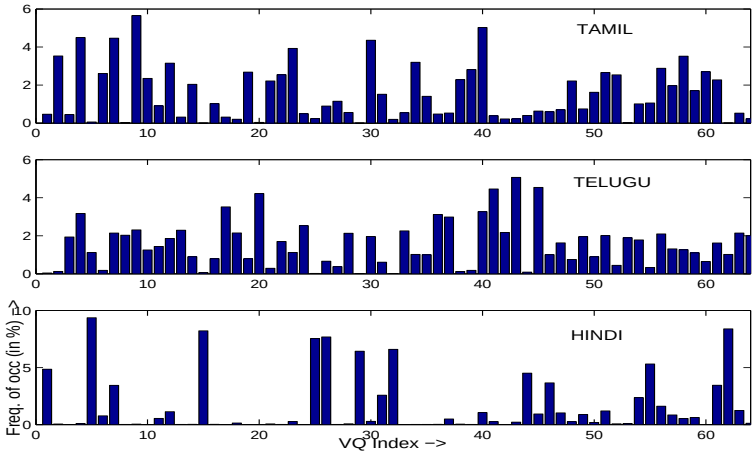

Figure 1: Freq. of occurrence (in %) of Tamil, Telugu and Hindi spectral vectors (Method I)

Using the probability of occurrence of each code book index in all the languages, the language identification is carried out in three different ways.

2.2.1 Method I

During testing, the test utterance is coded using the language-independent code book. This process creates a string of tokens, $s_1 s_2 s_3, ...s_p$, which we shall call the message 'S'. Now the task is to establish which of M languages generated that message, that is finding the language,L_i for which $p(L_i/S)$ is maximum.
From Bayes theorem,

$$p(L_i/S) = \frac{p(S/L_i).p(L_i)}{p(S)} \tag{1}$$

Since $p(S)$ is not dependent upon L_i, the problem is to find the language L_i for which $p(S/L_i).p(L_i)$ is maximum. We have assumed that $p(L_i)$ is the same for all the languages, even-though based on the geographical region there may be some variation. The problem now reduces to maximizing $p(S/L_i)$.

We have made another assumption that the tokens are generated by an independent random process, so that,

$$p(S/L_i) = p(s_1/L_i).p(s_2/L_i)...p(s_p/L_i) \tag{2}$$

$$log\ p(S/L_i) = \sum_{k=1}^{P} log\ p(s_k/L_i) \tag{3}$$

where,
- P is the number of observation symbols in the test utterance.
The decision is made on the following criterion,ie.

$$arg \max_{1 \leq i \leq M} log\ p(S/L_i) \tag{4}$$

where,
- M is the number of languages under consideration.

The overall language identification performance of this method for 2.5s test utterances is 90% and for 1.25s and 0.8s test utterances, the performance is 85.7% and 79.1% respectively as given in Table.1.

Table 1: Language identification performance (in %) for different duration of test utterances (Method I).

Language/ duration	performance for 2.5s	performance for 1.25s	performance for 0.8s
Tamil	96.2	85.7	73.8
Telugu	81.5	80.6	78.4
Hindi	92.5	91.0	85.2

2.2.2 Method II

In method I, during the testing stage, for each language, L_i, the $p(s_k/L_i)$ is directly used without considering the $p(s_k/L_j)$,i.e., the probability of occurrence of the index s_k in the language L_j. Suppose, if $p(s_k/L_i)$ is significantly larger than $p(s_k/L_j)$, then weightage can be given to that index s_k for the language L_i and vice versa. Based on this, the weightage or here, we call it as "Usefulness", of each of the code book indices for all the languages, are computed and used during testing.
For each language, L_i, the usefulness of the spectral vector, s_k, is computed in the following manner.

$$U(s_k, L_i) = \sum_{j=1 \& j \neq i}^{M} p(s_k/L_i)\ log\ \frac{p(s_k/L_i)}{p(s_k/L_j)} \tag{5}$$

$$= \sum_{j=1 \& j \neq i}^{M} U(s_k, L_i)|_{L_j} \tag{6}$$

for k = 1,2,...,N.

where,
- N is the CodeBook Size

- M is the number of languages under consideration.
- $U(s_k, L_i)$ is the Usefulness of the Spectral vector s_k for L_i
- L_i and L_j are the ith and jth languages
- $p(s_k/L_i)$ is the probability of the Spectral vector s_k belonging to L_i

The usefulness of all the spectral vectors, calculated using eqn.6, is shown in Fig.2. If the frequency of occurrence of a spectral vector s_k in one language, say L_i is greater than other language, say L_j, then s_k gets a positive weight for L_i and negative weight for L_j. In the testing stage, the test utterance is coded using the language-independent codebook and the decision is made based on the following criterion, ie.,

$$arg \max_{1 \leq i \leq M} \left[\sum_{k=1}^{P} U(s_k, L_i) \right] \tag{7}$$

where,
- P is the number of observation symbols in the test utterance.

Since the log-likelihood ratio is used to weight each of the spectral vectors, there is a considerable improvement in the language identification performance as given in Table.2.

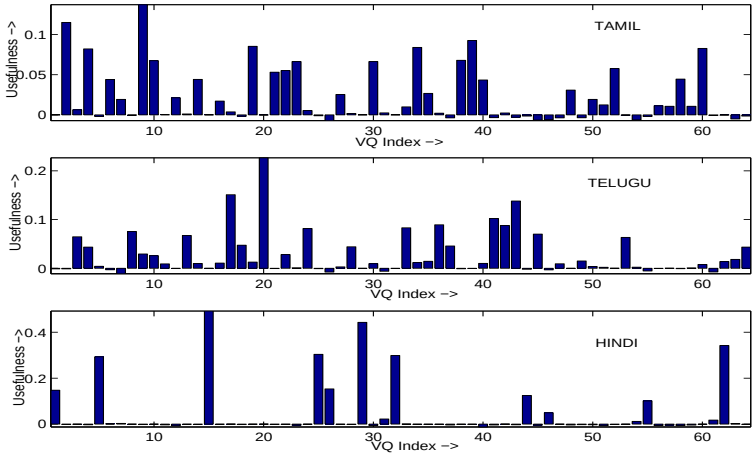

Figure 2: Usefulness of Tamil, Telugu and Hindi spectral vectors (Method II)

Table 2: Language identification performance (in %) for different duration of test utterances (Method II)

Language/ duration	performance for 2.5s	performance for 1.25s	performance for 0.8s
Tamil	98.0	93.4	89.0
Telugu	91.1	84.6	84.2
Hindi	98.6	98.0	97.5

2.2.3 Method III

In method II, for each of the spectral vectors, s_k, of each language, L_i, the usefulness is computed by taking sum of the log-likelihood ratios. Suppose, if $p(s_k/L_i)$ is greater than or less than $p(s_k/L_j)$ of all the other languages, then the usefulness of that particular spectral vector s_k is further strengthened, so that one can conclude whether s_k is useful or not useful to a particular language. On the other hand, if $p(s_k/L_i)$ is greater than $p(s_k/L_j)$ of some languages and less than $p(s_k/L_j)$ of other languages, then clearly the usefulness of s_k for LID can not be established. To address this issue, in Method III, for every pair of languages under consideration, say L_i and L_j, the usefulness of the spectral vector, s_k in the codebook, is computed by the following pair of equations.

$$U(s_k, L_i)|_{L_j} = p(s_k/L_i)\ log\ \frac{p(s_k/L_i)}{p(s_k/L_j)} \tag{8}$$

$$U(s_k, L_j)|_{L_i} = p(s_k/L_j)\ log\ \frac{p(s_k/L_j)}{p(s_k/L_i)} \tag{9}$$

for k = 1,2,...,N.

where,
- N is the Codebook Size
- $U(s_k, L_i)|_{L_j}$ is the Usefulness of the Spectral vector s_k for L_i when it is competing with L_j
- L_i and L_j are the ith and jth languages
- $p(s_k/L_i)$ is the probability of the Spectral vector s_k belonging to L_i

In this case, the number of usefulness functions is equal to $M(M-1)$, where M is the number of languages. The weights calculated by considering languages in pair is shown in Fig.3 and Fig.4.

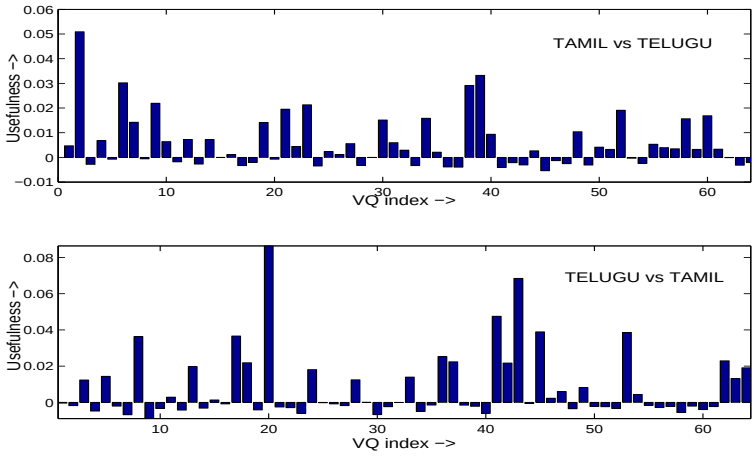

Figure 3: Usefulness of Tamil vs Telugu spectral vectors (Method III)

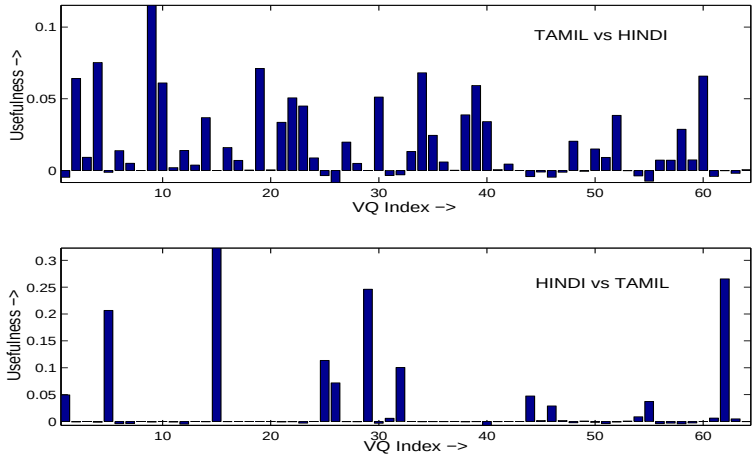

Figure 4: Usefulness of Tamil vs Hindi spectral vectors (Method III)

Table 3: Language identification performance (in %) for different duration of test utterances (Method III)

Language/ duration	performance for 2.5s	performance for 1.25s	performance for 0.8s
Tamil	99.3	97.2	95.4
Telugu	97.1	95.9	90.0
Hindi	99.4	98.2	96.1

The decision will be based on the following criteria :

$$arg \max_{L_i, L_j} \left[\sum_{k=1}^{P} U(s_k, L_i)|_{L_j}, \sum_{k=1}^{P} U(s_k, L_j)|_{L_i} \right] \tag{10}$$

where,
- P is the number of the observation symbols in the test utterance. Here, at one instant, only two languages are allowed to compete with each other. The winning language alone is allowed to proceed further to compete with the other language. Therefore, in this case also, the number of competitions/comparisons is equal to the total number of languages under consideration. In this method, since the usefulness of each of the spectral vectors is computed for pairs of languages the confusion between languages is greatly reduced and it is seen that for 2.5s test utterances, the overall performance has raised to 98.6% (Table.3).

2.3 Multiple codebook based method

There are two basic problems with common codebook based techniques. Firstly, if the number of languages are more, for the same codebook size, the statistics derived may not be optimum. This

problem may be overcome by increasing the codebook size. But that will increase the computation time during testing, which may make the system unsuitable for real-time applications. Secondly, among all the languages under consideration, if more than half of the number of languages belong to one familial group, the codebook itself will be biased. To avoid such problems, we have decided to go for multiple number of codebooks for language identification.

Unlike the common codebook based methods, here for each pair of languages, a separate codebook is generated. If "M" is the number of languages, then $\frac{M(M-1)}{2}$ code books are generated. Using each codebook, the usefulness of the all the spectral vectors of the corresponding two languages are computed and tested as explained in Method III.

Table 4: Language identification performance (in %) for different duration of test utterances (Multiple codebook case)

Language/ duration	performance for 2.5s	performance for 1.25s	performance for 0.8s
Tamil	99.36	97.0	96.6
Telugu	98.2	97.5	95.0
Hindi	99.38	99.0	97.1

The language identification performance (in %) of this method, for different durations of test utterance is given in Table.4 and the pair-wise language identification performance is given in Table.5. It is observed that the performance of this method is not much affected by the number of languages and even for 0.8s test utterances the overall performance is 96.2%.

Table 5: Language identification performance (in %) for pairs of languages (Multiple codebook based method) for 2.5s test utterances.

Ta . Te	Ta . Hi	Te . Ta	Te . Hi	Hi . Ta	Hi . Te
99.45	99.5	98.8	99.0	99.8	99.6

3 Discussion

From the performance of the common codebook based methods, described in Section 2.2, it is observed that, instead of using relative frequency distribution of spectral vectors of each language separately, if the usefulness, i.e., the log- likelihood ratio is used, there is a considerable improvement in the performance as we can see from Table.1, Table.2 and Table.3. It is further observed that, even if the duration of test utterance is reduced, the degradation in the performance is graceful. In Method I, described in Section. 2.2.1, as we reduce the duration of the test speech signal from 2.5s to 0.8s, the reduction in performance is 11 % (approx). But in Method II and III, it is just 4.5 % (approx.). Interestingly, in Method III, even for 0.8s of test speech signal, the overall performance is 93.8 %.

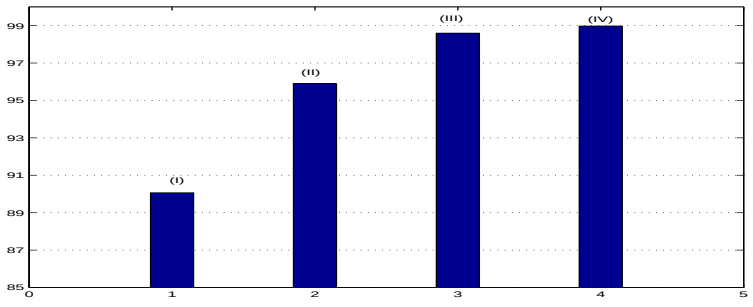

Figure 5: Performance (in %) of different methods for 2.5s test utterances. (I) - Method I. (II) - Method II, (III) - Method III. (IV) - Multiple codebook based method.

In the case of multiple codebook based method,described in Section.2.3, eventhough the improvement in performance is small compared to Method III (see Figure.5), if we increase the number of languages, it is expected that there may not be a significant degradation in performance since the problem is always treated as a two language problem.

From the performance of all the above explained methods, there is a general observation that the performance of the language Hindi, almost in all cases, is the best. The reason for this behaviour is, among the three languages considered for the study, the language Hindi belong to the Aryan group. But the other two languages, i.e., Tamil and Telugu, belong to the Dravidian group.

4 Conclusion

In this paper, we have analyzed a series of approaches to handle the variations in the relative frequency distributions of spectral vectors of different languages and it is shown that if the weight or the usefulness of each of the spectral vectors is derived for pairs of languages, there is a considerable improvement in the language identification performance. It has further been shown that instead of using a common set of language-independent spectral vectors for deriving the relative frequency distributions of them, if separate codebooks or set of spectral vectors, for each pair of languages are used, the performance of the LID system can be considerably improved.

References

[Muthusamy *et al.*, 1994] Yeshwant K.Muthusamy, Etienne barnard and Ronald A.Cole. Reviewing automatic language identification. In *IEEE Signal Processing Magazine*, pages 33–41, October, 1994.

[Foil, 1986] J.T.Foil. Language Identification using noisy speech. In *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, pages 861–864, April, 1986.

[Goodman *et al.*, 1989] F.J.Goodman, A.F.Martin and R.E.Wohlford. Improved automatic language identification in noisy speech. In *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, pages 528–531, May, 1989.

[Zissman, 1993] M.A.Zissman. Automatic language identification using Gaussian mixture and hidden Markov models. In *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, pages 399–402, April, 1993.

[Sugiyama, 1991] M.Sugiyama. Automatic language recognition using acoustic features. In *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, pages 813–816, May, 1991.

[Balleda, 2000] Jyotsna Balleda. A spoken language identification system for Indian languages. *M.S Dissertation, Department of Computer Science and Engg., Indian Institute of Technology, Madras* , June, 2000.

[Tucker *et al.*, 1994] R.C.F.Tucker, M.J.Carey and E.S.Paris. Automatic language identification using sub-words models. In *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, pages 301–304, April, 1994.

[DD News, 2001] Database for Indian Languages. *Speech and Vision Laboratory, IIT Madras*, 2001.