A PAIRWISE MULTIPLE CODEBOOK APPROACH TO IMPLICIT LANGUAGE IDENTIFICATION

T.Nagarajan and Hema A. Murthy

Dept. of Computer Science and Engineering
Indian Institute of Technology, Madras, Chennai. 600 036.
hema@lantana.iitm.ernet.in

ABSTRACT

Automatic spoken language identification is the task of identifying the language from a short duration of a speech signal. One of the important language identification cues is the differences in phoneme frequencies among different languages. Considering this, we develop a pairwise multiple codebook approach to language identification. This system is compared with the traditional single codebook per language system. Traditional VQ based language models are generally preferred since they do not explicitly require language models. The evaluation with Oregon Graduate Institute Multi-Language Telephone Speech Corpus shows that the multiple codebook system improves the performance by almost 7%.

1. INTRODUCTION

Automatic spoken language identification (LID) is the task of identifying the language from a short duration of the speech utterance. Humans are the best LID systems in the world today. Just by hearing one or two seconds of speech of a familiar language, they can easily identify the language. The sources of information used by humans to identify the language are several.

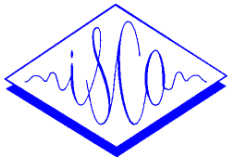
Any language is a sequence of phones/sound units and the differences between the languages can be at several levels. Hierarchically, we can say Phone level, Consonant-Vowel (CV unit) level, Syllable level, Word level and Sentence level. The possible differences among different languages at these levels are the inventory, the frequency of occurrence of different units in each set, the sequence of units(Phonotactics) and their frequencies of occurrence, the acoustic signature, the duration of a same sound unit in different languages and intonation patterns of units at each level. These are the whole possible sources of information for language identification.

The performance of any LID system depends on the amount of information extracted, the reliability of information extracted and how efficiently it is incorporated into the system.

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 are termed as **Explicit LID systems**. The systems that require only the speech signal and the corresponding true identities of the languages being spoken are termed as **Implicit LID systems**, in which language models are derived from the speech signal alone.

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 [1] and Martin [2] exploited the frequency of occurrence of voiced sounds using formant locations to represent sounds. A Vector Quantization (VQ) distortion measure was used as the basis for language decisions. Zissman [3] used a simple GMM classifier where the

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



Workshop on Spoken Language Processing
An ISCA-Supported Event
Mumbai, India
January 9-11, 2003

decision is made by calculating log-likelihood that a language model produced the unknown speech utterance. Sugiyama [4] and Jyotsna [5] performed VQ classification on different feature sets where the decision is based on accumulated distortion. Sugiyama [4] has further shown the possibility of classifying the languages based on their VQ histogram patterns.

The performance of the human listeners on language identification task has been analyzed by Muthusamy et al [6] and it has been observed that humans can identify the language of an utterance even if they do not have any knowledge about that language. This suggests that humans are able to extract and use some language specific patterns from the speech signal itself to recognize it. Using the same rationale, in this paper we describe two LID systems in which the language models are derived from the speech signal alone and no additional higher level knowledge of the language is supplied.

In this work, we have made two major assumptions that, if the phoneme frequencies are different among different languages, then the spectral vector distribution should also be different and if the acoustic signature of the same phoneme is different among languages, that should also manifest in a shorter segment of the speech signal.

2. SYSTEM DESCRIPTION

2.1. Speech Corpus

The Oregon Graduate Institute Multi-language Telephone Speech Corpus, which is designed specifically for LID research, is used for both training and testing. This corpus currently consists of spontaneous utterances in 11 languages: English (E), Farsi (Fa), French (Fr), German (G), Hindi (H), Japanese (J), Korean (K), Mandarin (M), Spanish (S), Tamil (T) and Vietnamese(V). The utterances were produced by ∼90 male and ∼40 female, in each language over real telephone lines. In our work, presently 10 languages are used (except Hindi, since the corpus does not contain sufficient data). To maintain the homogeneity in training and testing across languages, for each language 40 male speakers/15 female speakers are used for training and 15 male speakers/5 female speakers are used for testing. The rest of the data is used as development set. All the training and test set speakers are different.

2.2. Why vector quantization ?.

In this work, a new LID system has been developed, which captures the variation in the spectral vector distribution using a set of language-independent spectral vectors derived from all pairs of languages.

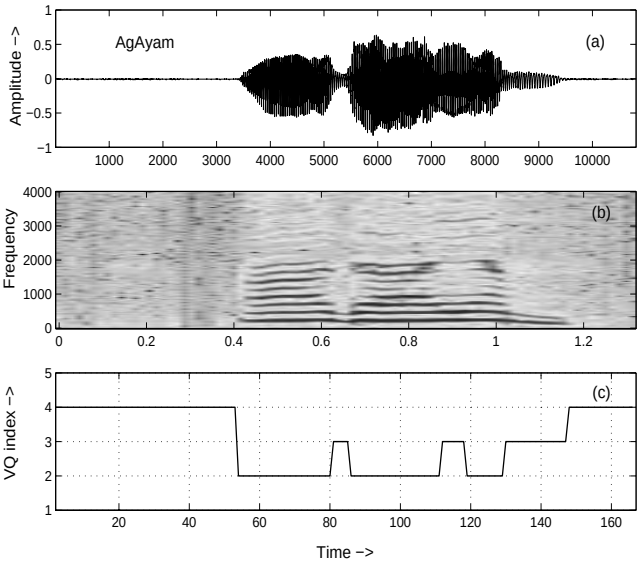


Figure 1: *Vector Quantization.*

The key advantage of the VQ representation is, discrete representation of speech sounds. By associating a phonetic label (or possibly a set of phonetic labels or a phonetic class) with each codebook vector, the process of choosing a best codebook vector to represent a given spectral vector becomes equivalent to assigning a phonetic label to each spectral frame of speech [7]. For example, let us consider a waveform for the sequence /SIL-A-g-A-y-a-m-SIL/, which is shown in Figure.(1a). The corresponding time-aligned spectrogram is given Figure.(1b). The sequences of indices after en-

coding, using a codebook with codebook size 4, is given in Figure.(1c). It clearly shows that at the transition of each phoneme a different index is assigned. But at the same time, since the size of the codebook is very small, for all the consonants present in the sequence, same index is assigned. From this, we can infer that using VQ with a small codebook size, it will be possible to classify the speech signal into broad phonetic classes. If the codebook size is too big then even for the same phoneme different indices may be assigned. From this we can conclude that if the size of the codebook is properly chosen, even without using phone-recognizers we will be able to derive information about the phoneme's acoustic signature and their frequency distributions.

2.3. Minimum distance classifier (MDC)

In our laboratory [8], a Spoken language identification system for Indian languages has been developed using vector quantization. That suggests that differences in languages exist at both acoustic and linguistic levels. In that work [8], it is observed that there are certain sounds that contribute to the performance of the language identification. It has also been established that the same sound has different acoustic signatures in different languages. As an initial study, the same system has been used but experiments have been run on OGI corpus. This is used as a baseline for comparison with the new system.

2.3.1. Baseline system

The language identification system used as a baseline for this work consists of VQ based modeling of the languages under consideration along with the classifier that evaluates the minimum distortion of the unknown speaker's speech utterance. The classifier hypothesizes the identification of the unknown utterance by determining which language model reduces the distortion for the test utterance.

During training, 'd' dimensional Mel Frequency Cepstral Coefficients (MFCC) are extracted from each speakers data of each language L_i. Cepstral mean subtraction(CMS) is employed to compensate for channel effects. The front-end parameters like, frame size, frame shift, number of channels, frequency range and the number of cepstral coefficients are tuned for the development set. The extracted feature vectors are clustered in the feature space using K-means algorithm, to get 'N' representative spectral vectors. The same process is repeated for all the languages and separate codebooks (or the language models) are derived.

During the testing phase, features (MFCC with CMS) are extracted from the test speech signal. Euclidean distance measure is used to find out accumulated distortion for each language model and the language model which gets the minimum accumulated distortion is declared as the language of the test utterance.

In the language identification task also, it is well established [3] that gender-dependent models always outperform the gender-independent models. We were motivated to use gender-dependent front-ends and back-ends for two reasons.

- Since the number of female speakers in each language is very small when compared to male data, the gender-independent models may be biased toward male data.
- The models can always be trained better if we do not mix up male and female data together.

During testing, the average pitch frequency of the test speaker is found using the pitch detection algorithm of Edinburgh Speech Tools [9]. If the pitch frequency is below 175Hz, the test signal is passed to the male models or else it is given to female models for language identification. The procedure is not to decide the test speaker's gender.

In this method, our intention is to find out the actual distance between the feature vectors extracted from the test utterance and the nearest entry in the codebook or the reference vectors. In other-words, the interest was on how close the test feature vector is from the reference vectors. Since the variations in the frequency of occurrence

of spectral vectors across languages is also one of the language identification cues, another system has been developed which is explained in the next subsection.

2.4. Frame indexing followed by language modeling (FILM)

By considering the differences in the frequency of occurrence of a set language-independent spectral vectors among languages as a language identification cue, we have already developed a system for Indian languages [10]. In that work, we have analyzed a series of approaches to handle the variations in the relative frequency distributions of spectral vectors and it was 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 used, the performance of the LID system can be considerably improved. From that observation, we have decided to derive codebooks for all pairs of languages in the OGI corpus also.

The system consists primarily of four steps : feature extraction, clustering, calculation of weights for cluster centers and language identification. In the first step, from the speech files of each pair of languages, 13 dimensional mel frequency cepstral coefficients are extracted (excluding zeroth coefficient since it contains only the energy level information) from all the frames with frame size of 20ms and frame-shift of 10ms. Cepstral mean subtraction is employed to compensate for channel effects. In the second step, K-means algorithm is used to cluster all the feature vectors, producing N cluster centers, where N is the codebook size. Considering the number of phonemes in each language, the codebook size is selected as 64. Since, feature vectors are extracted from each pair of languages, the resultant codebook contains language-independent spectral vectors. Using this set of spectral vectors in the third step, the speech files corresponding to each language L_i are coded separately and for each of the reference spectral vector, s_k, the probability of occurrence, $P(s_k/L_i)$ is computed. This probability values can be used in several ways to weight each spectral vector s_k of a language L_i. But here, the intention of deriving the codebook for each pair of languages is to find out the discrimination power of each spectral vector for language identification.

Among all the spectral vectors only a small set will contribute to the identification of a language. The spectral vectors, whose probability of occurrence in one language is very different from the probability of occurrence of the same spectral vector in the other language, do have greater discriminatory power. That is, if

$$P(s_k/L_i) >> P(s_k/L_j) \quad (1)$$

$$P(s_k/L_i) << P(s_k/L_j) \quad (2)$$

then that particular spectral vector, s_k contribute significantly to the identification of a language.

Based on this, for every pair of languages under consideration, say L_i and L_j, the weights or the usefulness of the spectral vectors in the codebook is computed by the following pair of equations.

$$U(s_k, L_i)|_{L_j} = p(s_k/L_i)\ log\ \frac{p(s_k/L_i)}{p(s_k/L_j)} \quad (3)$$

$$U(s_k, L_j)|_{L_i} = p(s_k/L_j)\ log\ \frac{p(s_k/L_j)}{p(s_k/L_i)} \quad (4)$$

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

The usefulness of the spectral vectors belonging to English when it it competing with French

and vice-versa are given in Figure.2. From the Figure, we see that, for example, the spectral vectors corresponding to the VQ indices 6,7,8 are in favour of English and 9,10,11 are in favour of French. We can further notice that some of the spectral vectors, for example from 50 to 64, are neither in favour of English nor in favor of French. This implies that the probability of occurrence of these spectral vectors in both the languages are negligible and in turn implies that number of phonemes in these languages is less than 64. But in some of the languages the number of phonemes may be slightly more than that of English or French. To maintain the uniformity, the size of the codebooks of all the pairs of languages is fixed to 64.

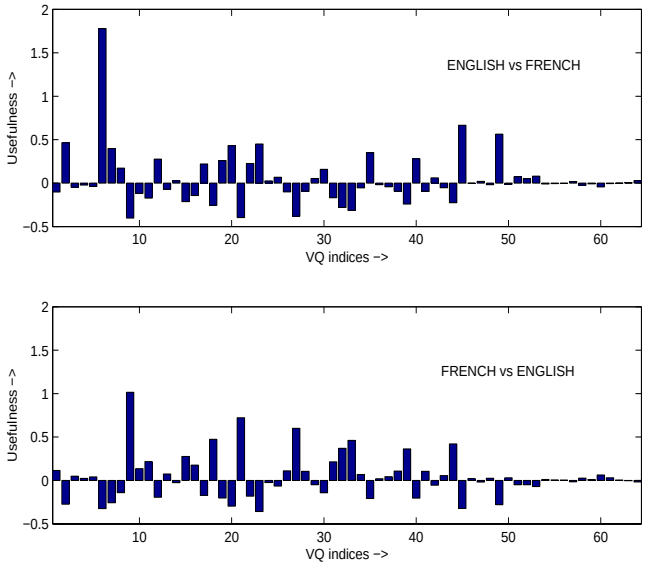


Figure 2: *Usefulness of spectral vectors for English and French.*

During testing, the tournament is between a pair of languages. The testing methodology adopted here is similar to the construction of a winner tree (see Fig.3). In Fig.3, each non-leaf node in the tree represents the winner of a tournament, and the root represents the overall winner, here the language of the test utterance. In this way, even-though the number of pairs of languages or the number of codebooks is equal to $\frac{M(M-1)}{2}$, where M is the number of languages considered, the number of tournaments (or comparisons) is $M-1$ only. The language which wins in the $M-1$ th tournament is declared as the language of the test utterance.

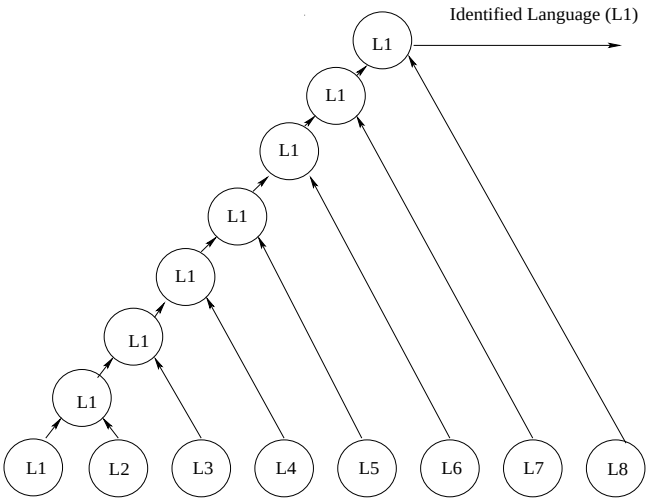


Figure 3: *Testing methodology.*

During each pairwise testing, say between the languages L_i and L_j, the test utterance is coded using the codebook $C_{L_iL_j}$. This process creates a string of tokens, $s_1...s_p$, which we shall call the message 'S'. Now the task is to establish which of the two languages generated that message, using the weights derived from the probability of occurrence of the spectral vectors, here the token s_k.

The decision for each pair of languages, will be based on the following criteria :

$$arg \max_{L_i,L_j} \left[\sum_{k=1}^{P} U(s_k, L_i)|_{L_j}, \sum_{k=1}^{P} U(s_k, L_j)|_{L_i} \right] \quad (5)$$

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

3. PERFORMANCE ANALYSIS

There are many ways to measure the performance of an LID system, including its accuracy and computational complexity. Computational complexity is often difficult to compare, since it depends on the specific implementation and the computing platform. Therefore, we will focus only on our

system's language identification accuracy, as measured by its 1-best accuracy and n-best accuracies.

3.1. Performance of baseline system

3.1.1. Optimization of parameters

The parameters of the system like, the frame-size, frame-shift, number of channels, frequency-range, feature vector dimension, number of training set speakers and the codebook size are tuned. For some of the parameters, since there is no theoretical justification, a line search is performed to determine the best parameter set. The list of experiments are tabulated in Table.1.

Experiments :
1. No. of training set speakers : 10,20,30,40
2. Codebook size : 8,16,32,64,128
3. Feature dimension : 12 to 20 (in steps of 1)

Table 1: List of experiments.

Based on the experiments and analyzes, the best parameter set is derived and listed in Table.2 and n-best performance for the baseline system is given in Table.3.

parameter	Value
Frame-size	20ms
Frame-shift	10ms
No. of channels	36
Frequency-range	300-3400
No. of training set speakers	40
Codebook size	128
Feature dimension	13

Table 2: Parameters after tuning.

n	1	2	3
Accuracy	53	66	80

Table 3: n-best language identification accuracy(in %) of MDC

3.1.2. Proximity analysis

We have made another analysis on how different the languages are. In each testing, the top-5 languages are considered to be acoustically similar to the test language and the bottom-5 languages are considered to be acoustically dissimilar. From all the results of each language testing (20 test speakers), based on the number of times each language appears in top-5 places, the languages are sorted and tabulated (see Table.4.).

Language	Similar	Dissimilar
E	E, G, Fa, S, J	Fr, M, V, K, T
Fa	Fa, G, K, S, M	J, Fr, V, T, E
Fr	Fr, K, G, M, S	J, Fa, E, V, T
G	G, Fr, K, S, M	J, Fa, E, V, T
J	J, K, S, M, V	Fr, Fa, G, T, E
K	K, Fr, J, S, M	V, G, Fa, E, T
M	M, V, J, K, S	G, Fa, Fr, T, E
S	S, E, Fr, Fa, J	G, M, V, T, K
T	T, S, V, K, M	J, Fa, E, G, Fr
V	V, J, M, T, S	K, E, F, G, Fr

Table 4: Proximity analysis.

Table.4. shows that there is some acoustic similarity and dissimilarity visible within a group of languages. From the third row in the Table.4, we can conclude that German is similar to French and Tamil is very dissimilar to French. From the fourth row too, we conclude the same. That is, the languages which are dissimilar to French are dissimilar to German also, since German is similar to French. But in some cases no such relationship exists. For example, in the first row which is for English, Farsi is in the similar list, but in Farsi, English is in dissimilar list. But the overall picture shows that the languages can be grouped. This analysis suggests that a modular approach can be used as front-end and that languages can be grouped into two or three classes.

3.2. Performance of the FILM method

Based on the analysis done on amount of the training data required for MDC, here the number of speakers used during training for male and female cases are fixed to 40 and 15 respectively. Here also, the number of languages considered is 10. Since the codebooks are generated for all the pairs of languages, the total number of codebooks is equal to $\frac{M(M-1)}{2}$, where M is the number of languages. The n-best performance for this method is given in Table.5 and it is compared with the MDC method also. From the Table.5., we can observe that for FILM method the overall performance for 1-best case is significantly better. But the computational complexity for the FILM method is $\sim$5 times greater, since the number codebooks required for this is 45 for 10 languages.

Classifier	n-best performance in %		
	1	2	3
MDC	53	66	80
FILM	59.5	72	79

Table 5: n-best performance comparison.

The language-wise performance of this method is compared with the MDC method and is tabulated in Table.6. For all the languages except English and German, the performance of the FILM is significantly better than that of the MDC.

As we have done for MDC, for this method also, the proximity analysis is done. (see Table.7.) Interestingly, for most of the languages, e.g., French and German, the similar and dissimilar languages are exactly the same as MDC case except a slight differences in their places. This analysis further strengthens the observation which we have derived from MDC's proximity analysis.

4. CONCLUSIONS

In this paper, we have defined a new approach to LID using a VQ based approach. Here, the language models are derived from speech signal alone and no additional higher level linguistic information is supplied. It is observed that when we use the variations in the frequency of occurrence of different spectral vectors as the LID cue, the performance is better. Even though the performance of our system for 10 languages is just 53% for MDC and 59.5% for FILM, it is noticed that if we consider the first 3-best languages, the performance is 80% for MDC. Further it has been concluded from the proximity analyses that a modular approach can be used as a front-end for language identification. As mentioned earlier, the information derived from one source will not be sufficient for a reliable language identification task. Information from other sources like variations in the intonation patterns and speech rate should also be extracted and a method of combining the ev-

Language	1-best performance in %	
	MDC	FILM
En	65	40
Fa	40	60
Fr	80	80
Ge	35	30
Ja	60	60
Ko	40	50
Ma	55	55
Sp	40	65
Ta	50	75
Vi	65	80
Average	53	59.5

Table 6: Language-wise performance comparison.

Language	Similar	Dissimilar
E	E, J, Fa, G, S	V, Fr, M, T, K
Fa	Fa, S, T, G, M	K, E, Fr, V, J
Fr	Fr, G, K, S, M	J, T, V, Fa, E
G	G, K, S, M, Fr	J, Fa, V, T, E
J	K, J, M, V, S	Fa, T, G, Fr, E
K	K, M, S, G, J	V, Fr, Fa, T, E
M	M, K, V, J, Fa	S, T, G, Fr, E
S	S, G, T, J, Fa	E, V, M, Fr, K
T	T, Fa, V, S, K	M, J, E, G, Fr
V	V, M, J, S, Fa	K, T, G, E, Fr

Table 7: Proximity analysis.

idence from these sources should be derived and used to improve the performance.

REFERENCES

[1] J.T.Foil, "Language identification using noisy speech," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, Apr. 1986, pp. 861–864.

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

[3] M.A.Zissman, "Automatic langauge identification using gaussian mixture and hidden markov models," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, Apr. 1993, pp. 399–402.

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

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

[6] Yeshwant K.Muthusamy, Etienne barnard and Ronald A.Cole, "Reviewing automatic language identification," *IEEE Signal Processing Magazine*, pp. 33–41, Oct. 1994.

[7] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, New Jersy, 1993.

[8] Jyotsna Balleda, Hema A. Murthy and T.Nagarajan, "Language identification from short segments of speech," in *Proceedings of Int. Conf. Spoken Language Processing*, Beijing, China, October 2000, pp. 1033 – 1036.

[9] "Cstr web page, University of Edinburgh," *http://www.cstr.ed.ac.uk*.

[10] T.Nagarajan and Hema A.Murthy, "Language identification using spectral vector distribution across the languages," in *Proceedings of Int. Conf. Natural Language Processing*, Dec. 2002, p. (accepted for presentation).