Automatic Language Identification Using Syllabic Spectral Features

Kung-Pu Li

ITT Aerospace Communication Div., San Diego, California 92131

ABSTRACT

Automatically identifying a language from just the acoustics is a challenging problem. Speaker differences are usually greater than language differences. This study has developed a text-independent system that is capable of performing both speaker and language identification. The system utilized different feature sets to observe changes in recognition performance to identify which set of features is suitable for language identification. Through these experimental results, the spectral features at the syllabic level have proven to be reliable for distinguishing languages. Performance on a five language database has exceeded 95% identification accuracy. Two other telephone-speech databases were also tested.

INTRODUCTION

A multilingual person has no problem identifying languages he understands. At the present, an automatic speech understanding system or a word spotting system even for a single language is still at the research stage. An automatic language identification without language understanding is a challenging problem. In the language identification domain, we must assume no test speaker's speech exists in the reference set. The comparison between a test sample and reference samples are always from unconstrainted utterances of two different speakers. Therefore, the differences between two utterances encompass text differences, speaker differences, environmental differences, and language differences. The main issue is how to extract language differences for the language identification. From past research, all speech recognition, speaker identification, and language identification operate within the same set of short term spectral representation. There is no certainty that speech recognition, speaker identification, and language identification can be treated as separate problems as in the past. This study intends to clarify some of these issues.

House and Neuberg [1] published the earliest language recognition research using Hidden Markov Models, HMM. Later, several studies [2,3], using either segmental information or time-frequency spectral patterns, demonstrated some encouraging results. Recently there were several explorations into using phonetic-acoustic level features [4,5] for language identification. The statistical approach of segmental information [6], and the use of an HMM or Gaussian mixture Markov model on speech parameters [7] have also being investigated. However, they assume that the model of a language structure can be trained as speaker independent.

The language identification system faces the following problems: (1) the segmentation process must be speaker and language independent, (2) length utterance must be suitable for statistical estimation, (3) a single statistical language model must cover the whole population of speakers, and (4) features must be effective for language recognition. When using spectral features, it is important to distinguish speaker and language differences to overcome some of those problems. In other words, a system should use the same underlying structure and the same database, from which we can test the system's capabilities of speaker discrimination and language recognition. We can assume that each speaker has individual speech and language characteristics. In the most extreme case, each speaker has his own individual language model. During the process of matching a target speaker, the measure represents the similarities of both acoustical and language structural characteristics. If two messages from two different speakers match well at the acoustic and segmental levels, the possibility of the two utterances being the same language is higher than that from two different languages. In theory, this process can reduce speaker differences. In addition, the same system can provide a comparison study on speaker *and* language recognition. When the feature set is changed to increase language differences, the speaker verification performance is less affected than the language identification. This phenomena can be used to determine appropriate feature sets for language identification.

APPROACH

The baseline system used in this study is based on a speaker identification with the nearest neighbor algorithm [8], which is a non-parametric approach using the averaged nearest neighbor distance to identify speakers. The basis of this measure and its interpretation in terms of classical pattern recognition is discussed in detail in the reference. The major advantage of the algorithm is that training samples from each speaker can be limited.

0-7803-1775-0/94 $3.00 © 1994 IEEE

All properly selected samples form their own local distribution without any statistical estimation. The message difference measure becomes a matching of distributions between a reference and an unknown. As described in the reference[8], each message difference contains *forward* and *backward* distances. These distances form a number of different measures. With a subsequence normalization of environments, including variations of length, contents, and speaker biases in those measures, the accuracy of matching a target speaker becomes excellent[8].

The variation of spectral differences between any pair of speakers is much larger than the language differences obtained from a single bilingual speaker [9]. Speaker and language differences are difficult to separate, but can be distinguished with the proper decomposition. Therefore it can be assumed that the two closest speakers have similar speaking characteristics and/or a similar language structure. Reducing the differences between two speakers to find the remaining language differences becomes an essential process in dealing with the speaker variation problem. For a speaker identification system, it can find matched utterances for the same speaker. Language identification can then be determined by finding the best matching speaker from a known language group.

Figure 1 illustrates a speaker identification system modified for the purpose of language identification. All

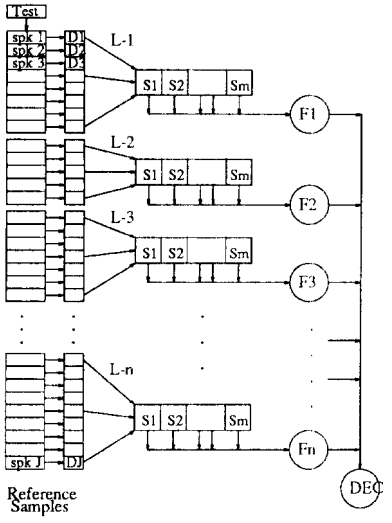

Figure 1. The basic language identification system.

of the speakers' models from the same language are simply grouped together. After an appropriate normalization of these message distances for comparable scores, each language group can then provide or choose a small set of best matched speakers to form the language score for an unknown utterance. Both the minimum and average scores in each language can be used to identify the language of the tested utterance. The *minimum* score is found when the system searches for the minimum message difference between the test sample and all reference speakers in each language group. The *average* score is the average of all message differences from all speakers in a language. Besides, an averages of N-best scores can also form many different scores with different N.

Several different sets of databases have been tested for this experimental study over the past several years. (A) IDA Five Language Database[a]: This digital data contains five languages (three European and two Asian languages), and was recorded from 82 male speakers with three different sessions for each speaker. (B) Three Asian Languages from the "Spoken Language Library" Database: This database contains 16 pairs of telephone conversations from 32 different male and female speakers, (C) Oregan Graduate Institute, OGI, Multi-language Telephone Speech Database: OGI [10] collected a set of multi-language telephone monologue speech, OGI_TS. The database for language identification experiments contains 10 languages and 900 different speakers from the recordings.

EXPERIMENTS

This experiment compares different input feature sets: i.e., 14 filter spectra plus relative amplitude value at the frame rate, and five-frame spectral samples at the syllabic rate. The spectral samples at the syllable rate are extracted by first automatically marking vocalic centers, then encoded by the change of spectra at the on-set of the syllabic nuclei with multiple spectral frames. The procedure, [11] which uses artificial neural networks to assist syllabic nuclei marking of the database, demonstrates a stable, language independent, and accurate process. The concept of this procedure is to devise a system that can be trained to extract some language structural segments without any high level language knowledge.

Speaker and Language Identifications

Both speaker identification and language identification experiments can be conducted with the IDA five language database. Speaker identification is determined by the best speaker (minimum) score among 82 speakers. The language identification results are simply determined by the best score among languages. The test samples for speaker identification include only the second and third sessions from each speaker. However, in language identification, the results are tested on three sessions for each speaker, while the reference uses only the first session and excluding the testing speaker. The

[a] The digitized data was provided by IDA, Princeton, NJ, and the original audio recordings are from Rome Lab, Rome NY.

results are summarized in Table. 1. The best results for language identification were achieved using syllabic features, while the best speaker identification used frame

	Lang. Id.		Spk. Id.	
	Frame Spect.	Syllabic Feature	Frame Spect.	Syllabic Feature
Olano Param.	73.6%	73.6%	77.4%	65.9%
Blind Deconv.	69.5%	85.8%	76.8%	73.8%
Z-norm.	69.9%	88.2%	78.7%	76.8%

Table 1. The comparison of speaker and language identification when the inputs are changed from spectrum in frame rate to syllabic spectral features.

rate spectrum. These comparisons clearly show significant performance improvement on language identification when the input features use the syllabic spectral features; but the recognition of speakers is not significantly affected.

Average, Minimum, and Combined Scores:

We have found that the *minimum* score usually outperforms the *average* score for each language group. We also found that the combination of the minimum and average scores constantly provides the best results. The number of encoding frames for each syllabic nuclei also affects performance, as shown in Table 2. These results show that the optimal number of frames for each syllabic

Conditions	Performance		
IDA-5 Lang.	Average	Minimum	Combined
Four frames	84.1%	91.9%	94.3%
Five frames	84.1%	92.3%	94.7%
Six frames	85.8%	91.9%	95.1%
EV-36	86.2%	92.7%	94.7%
Opt-Wt	87.8%	93.5%	95.1%

Table 2. Experiments on different numbers of encoding and dimensional reduction of language identification on IDA five language database.

sample ranges from 5 to 6 frames. We also found that the results of a principle component analyses (EV-36) with and without optimal weighting (Opt-Wt) did not result in reduction in performance when the feature dimension was reduced from 75 to 36.

Encoding Syllabic Features

There were two kinds of encoding schemes also tested: *variable frame encoding* and *arc length encoding*. A summary of the best results on both IDA five and SLL three languages databases is shown in Table 3. The similarity of performance on both databases indicates that the change from high-quality reading speech to telephone conversational speech may not greatly influence or alter the performance. Note, the number of errors for each case is so small that any further development of the algorithm, and improvement of performance, requires either a new and/or a larger database.

DATABASE	Variable Frame Encoding		Arc-Length Encoding
	Trials	Percent	Percent
IDA Five Lang.			
5 Lang(82X3)	246	94.7%	93.9%
3 Lang(63X3)	189	99.5%	98.4%
2 Lang(48X3)	144	99.3%	99.3%
SLL 3 Lang(32)	32	93.8%	96.9%

Table 3. Present performance of language identification on different sampling techniques, and two different databases.

Experimental Results on OGI_TS Database

This same experimental procedure has been applied to the OGI_TS database. The training data are constrained with 25-50 sec./speaker from the designated training data, and 449 speakers from 10 languages are used for references. The testing data are from the development data in OGI_TS corpus. Two test sets are used (1) *Whole file:* where the duration of the test samples is between 25 and 50 sec., and (2) *10 sec. interval:* where the testing files are the hand marked segments of 10 seconds of speech from data in (1). The results with the OGI_TS database show a quite different trend. The minimum scores are very sensitive to the channel as well as the speaker variations. Performance was greatly degraded due to telephone channel and background noise variations. We have implemented several N-best averaged scores where N varies from 1/2 to 1/5 of the total number of speakers/language in the reference set. These N-best scores outperform either the average score or the minimum score.

Figure 2 shows the OGI_TS ten language closed set identification results, which are selected from the best

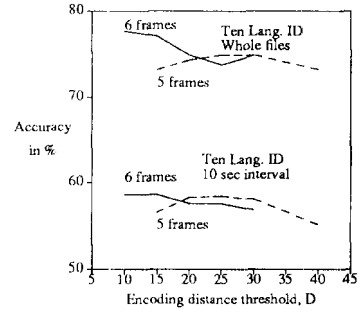

Figure 2. Ten language identification (closed set) results on OGI_TS database vs the encoding distance thresholds.

performance of different scoring techniques. The horizontal coordinate is the distance thresholds D for the variable frame encoding in syllabic feature extraction. Error analysis shows that the system relies on the number of each gender in the reference set within a language, and that errors are also affected by the closeness of the two

target languages. However, due to insufficient and unbalanced numbers of samples, speakers, and languages in the existing databases, we can only vaguely identify such factors. Additional data for female speech may verify this concern.

Besides the ten-language closed-set identification, several other language identification experiments were also tested. The best performance sorted from different scoring methods can be listed as follows: (1) English against one of nine languages and others, which is a three-class (*EN-L*-Others) separation with a fixed scoring, shows an average of 72.0% and 83.2% for 10 sec. intervals and whole files respectively, (2) All 45 language pairs separation (L_i–L_j) with a fixed scoring technique has an average of 81.9% and 92.2% for 10 sec. intervals and whole files respectively, (3) English vs one of nine languages as two-class (*EN-L*) separation with fixed scoring reaches 82.3% and 90.5% for 10 sec. intervals and whole files respectively, (4) English vs one of nine languages two-class (*EN-L*) separation with the language pair dependent scoring shows an average of 86.9% and 94.1% for 10 sec. intervals and whole files respectively, and (5) The best average English verification (EN-Others) ROC has shown an equal error rate of 78% and 90% for 10 sec. interval and whole file respectively. It is interesting to note that the performance differences between 10 sec. intervals and whole files are within the range of 7.2% to 12.0%, and error ratios varied from 1.4 to 2.3.

EXPERIMENTS ON TWO OTHER MODELS

There were two additional attempts to use different models or presentations to classify spoken languages. Both were speaker independent: (1) using an HMM to replace the syllabic encoding, and (2) using two stage back-error-propagation of a multiple-layer-perceptron neural network, BEP-MLP, to find the optimal discrimination of sequence syllabic features among languages.

From the IDA five language database, for HMM experiments, two languages were selected and divided into two separate parts: training and testing. The performance of three random divisions of data are consistently worse than the nearest neighbor results on the same data set. The error was about one and one half times the errors obtained from the nearest neighbor technique.

For BEP-MLP experiment, the training involved all speakers in the database. Although the recognition results are slightly better than those of the nearest neighbor technique without the tested speaker being in the reference set, the error rate of this result is about twice that of the nearest neighbor system which includes the tested speaker in the reference set.

DISCUSSIONS

The study has used a modified nearest neighbor speaker identification system for either speaker or language identification. The parallel experimental results show quite different amounts of language features at the syllabic level; however, speaker differences are still very strong. We find that the nearest-neighbor technique can overcome some of the problems when the two sets of features (speakers and languages) are overlapped. It has greatly reduced the effect due to large speaker variations. In conclusion, the syllabic spectral feature is suitable for use in language identification.

The average scoring helped improve performance when combined with the minimum score. This indicates that some common language characteristics exist in the feature space that are different from the differences between two closely-matched speakers in different languages.

Two other trained speaker independent language models (HMM and two stage BEP-MLP) show inferior results compared to the nearest neighbor technique. These results are evidence of the weakness in using a single language model to cover multiple speakers. Both techniques need much larger training sets to confirm their performance.

The best language identification system must be able to reduce speaker differences to show language distinction. We need a new set of databases that contain a sufficient number of speakers, including both male and females, and spontaneous conversational speech. The next step is to understand how many speakers are required in a reference set for the system to provide optimal performance.

The experiment shows the importance of minimum scoring compared with average scoring. The optimal result came from the combined score. Results also show that there are differences to discriminate different languages. The minimum score is the best if there are only speaker variations. The average score can be the best if there are no diverse groups of speakers and channel variations. In between these extremes, the averaging of n-best scores can handle certain outliers, channel changes, gender differences, and/or dialect differences. This may strongly suggest that the decision process for language identification must be able to reduce speaker, channel, and background noise variations. This study also demonstrated that speech, speaker, and language recognition technology is merged together as a single problem.

REFERENCES

[1]. House et al, JASA 62-3, pp. 708-13, 1977.
[2]. Li et al, ICASSP-80, pp. 884-7, 1980.
[3]. Leonard et al, TI Final Report, 1978.
[4]. Goodman et al, ICASSP-89, pp. 528-31, 1989.
[5]. Sugiyama et al, ICASSP-91, pp. 813-6, 1991.
[6]. Muthusamy et al, ICSLP-92, pp.895-8, 1992
[7]. Zissman, ICASSP-93, pp. II-399-402, 1993.
[8]. Higgins et al, ICASSP-93, pp. II-375-8, 1993.
[9]. Abe et al, JASA 90-1, pp. 76-82, 1991.
[10]. Muthusamy et al, ICSLP-92, pp. 1007-10, 1992.
[11]. Li, JASA 92-4(2) pp. A-2477, 1992.