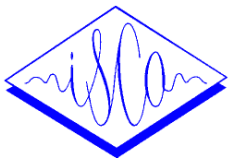

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

Language Identification of Six Languages Based on a Common Set of Broad Phonemes

Kay M. Berkling (berkling@cse.ogi.edu), Etienne Barnard (barnard@cse.ogi.edu)
Center for Spoken Language Understanding,
Oregon Graduate Institute of Science and Technology,
20000 N.W. Walker Road, P.O. Box 91000, Portland, OR 97291-1000, USA

Abstract

We describe a system designed to recognize the language of an utterance spoken by any native speaker over the telephone. Our previous work based on language-specific phonemes [5] is extended to include sequences of all lengths of language-independent speech units. These units are derived by clustering phonemes across all languages in the system (Hindi, Spanish, English, German, Japanese, and Mandarin). Our language-identification results based on broad-phoneme occurrence statistics indicate 90% accurate distinction between English and Japanese, which is comparable to results obtained when using language-specific phonemes. By relaxing the precision of language-dependent phonemes into language-independent broad phonemes we thus retain language discriminative power. The degree to which the precision can be relaxed while retaining sequences of broad phonemes that can discriminate between languages is an indication of the accuracy with which the phoneme segmenter and recognizer have to recognize the incoming speech.

1. Introduction

The work presented here is based on the hypothesis that incorporating linguistic knowledge is an essential part of a robust and extensible language identification system as envisioned in Fig. 1.

A system like this requires a maximal number of stages to be language independent. Rather than designing several language- dependent speech recognizers and running these in parallel, we seek to design a single recognition module which is language independent. We want to show here that this is feasible by recognizing the incoming speech in terms of language-independent speech units. The question therefor becomes one of finding the appropriate speech units.

Even though previous work indicates that best performance in language identification is achieved by using phonemes as the unit of speech, ([5],[3],[7]), there are a number of reasons why one may prefer to use a less detailed distinction between different speech untis. These reasons include the following:

- The proliferation of phonemes with an increasing number of languages, which both increases the computational cost of algorithms and decreases the classification accuracy of the phonetic recognizer.

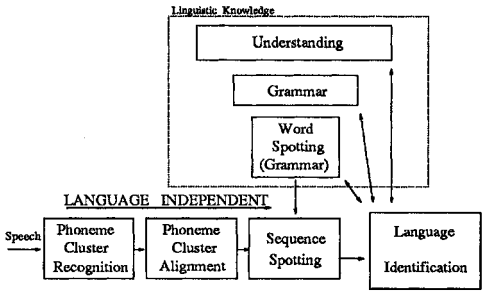

Figure 1. Modules of the LID System: The system consists of a phoneme recognizer, followed by an automatic alignment of the speech with the recognized phonemes. Finally features are derived based on the occurring sequences as well as mono-phonemes in the utterance.

- The absence of a common phoneme set to allow cross-lingual statistical analysis of linguistically meaningful phoneme sequences.
- The lack of discriminative training of phonemes across languages.
- The sometimes arbitrary choice of phoneme recognizers from a subset of the languages in the system.

As was shown in [1], most of the language dependent information is concentrated in the mono-phonemes (phonemes occurring in only one language) and not in the poly-phonemes (phonemes similar across languages). Hence, with regard to language identification, there exists a degree of redundancy in recognizing speech at the phoneme level. By clustering phonemes across all languages in the system and building a recognizer to classify these 'broad phonemes', we explore this redundancy, while solving some of the problems of more detailed phonetic approaches described above. As the number of clusters is decreased (i.e. as more phonemes are clustered together), higher accuracy of broad-phoneme recognition results. Therefor the goal is to have the largest possible clusters while retaining the ability to derive language-discriminating sequences in terms of these clusters. This serves to establish an upper bound on the required grain of the phoneme clusters, and thus the accuracy of the automatic segmenter.

Below we first show how we use clustering to derive our

language-independent set of broad phonemes. A "word-spotting" system based on such phonemes is then described, and it is shown that fairly broad categories maintain much useful information for language identification.

2. Broad Phonemes

2.1. Multi-Language Telephone Speech Corpus

The speech used here is taken from the six languages English, Japanese, German, Hindi, Mandarin, and Spanish in the OGI Multi-language Telephone Speech Corpus ([6]). A subset of the utterances which consist of a mix of unconstrained- and restricted-vocabulary speech, was phonemically labeled with 'Worldbet' as described in [4]. This set of labels is designed to capture phonetic similarities and differences not only within but across languages and has a mapping to the IPA and to the TIMIT symbol set[2].

2.2. Hierarchical Clustering of Phonemes

There are 247 distinct phonemes for which more than 20 examples occur in the set of 115613 distinct phonemes occurring in our labeled training set. To obtain a feature vector for clustering these phonemes, each phoneme is represented with 56 PLP coefficients computed from within a 174 msec window, centered on the middle frame (as indicated by the hand labels). Features representing all occurrences are averaged resulting in a vector used as representative of this phoneme in the clustering algorithm.

Clustering is done across all phonemes based on minimum Euclidean distance. Merged phonemes are represented with their average at the next level. Phonemes are found to cluster into natural classes according to manner and place of articulation, and the presence or absence of voicing. An example of the resulting tree structure is shown in Fig. 2, within the broad category of fricatives. By looking at the clustering it can be seen that /V/ and /G/ cluster closely together. (/V/ is a voiced unaspirated bilabial fricative as in sa-b-io, /G/ a voiced velar fricative as in a-g-uda). However, both phonemes occur only in Spanish and by clustering them together we still retain a monophoneme. On the other hand /T/ the interdental voiceless fricative occurring only in Engish and Spanish is clustered with /f/ which occurs in all languages. Thus this monophoneme is lost. However, for telephone speech the distinction between the two phonemes is virtually impossible and therefor no real loss in incurred by clustering them into one broad-phoneme. The German /C/ and /K/ (palatal and uvular voiceless fricatives), are borderline cases. They are not clearly separated from the other phonemes. Yet they are not so closely associated with another phoneme that they cannot be maintained as monophoneme in the broad-phoneme set.

2.3. Pruning the Clustering Tree

The phoneme clusters, established by pruning the clustering tree, will be used as the broad phonemes in terms of which the incoming speech is finally represented. Thus branches of the tree are pruned at a level broad enough to hide detail less important to the task of language identification, yet fine enough to detect language-specific detail, for instance phonemes or clusters of phonemes occurring in only one language.

The following criteria are indicators of the importance of each phoneme for language identification.

- the frequency of occurrences of each phoneme
- the number of languages the phoneme occurs in
- the degree of confusion with other phonemes

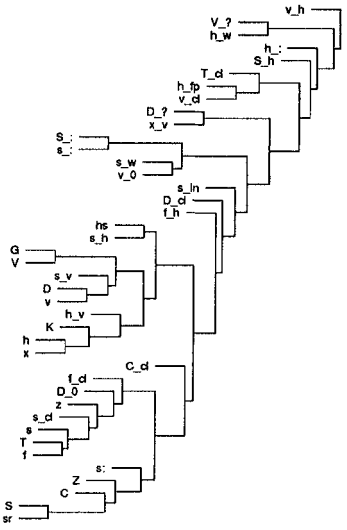

Figure 2. Phoneme Recognition: The superset of all phonemes from all N languages is clustered hierarchically. The level of phoneme clustering determines the granularity with which the utterances will be aligned. There exists a trade off between language specific detail and the accuracy of phoneme-cluster recognition which determines the cutoff line.

Taking into account these factors, we have manually created a manageable set of 84 language-independent broad phonemes which are valid across six languages while still retaining discriminative power.

2.4. Broad-phoneme Classification

A single three layer feedforward neural network was trained to discriminate between the 84 broad-phoneme classes on a frame-by-frame basis. The classifier performs with 32% frame accuracy, when evaluated on a test set of hand labeled speech from each language. Table 2.4 shows the number of broad phonemes within each of the categories of speech.

Table 2 details the Worldbet labels of the monophonemes which are retained in the 84 broad phonemes.

2.5. Language Identification from Phoneme Occurrence Statistics

As a preliminary investigation into the utility of these broad-phoneme classes, we performed the following experiment: Three simple features were extracted from each output of the broad-phoneme classifier, and used to classify the language of the underlying utterance. These features were:

- The average output activation **AVG** across the complete utterance,
- the maximum output activation **MAX**, and
- the variation in output activation **VARH**,

as explained in [5]. This results in a 3 x 84 element feature vector representing an utterance. Language classification is

CLASSES	NUMBER OF CLUSTERS
Vowels	17
Plosives	24
Frictives/ Affricates	23
Nasals/ Approx/ Flaps	12
Diphthong	7
non-speech	1
Total	84

Table 1. The number of broad phonemes within each of the categories of speech.

Language	MONO-Phonemes	% correct (avg/max)
English	r h_v d(>_r	51/59
Japanese	dz t	34/34
Spanish Hindi	nj/L hs V s_v x -	38/60
Mandarin	r cCc/tsc/tsrc tsR cCh ai_2/3/4 A_2/3/4 y_2 E_2/3/4	44/55
German	ö K C b_(yax > Y/> i/oax y: e:	37/71

Table 2. Monophonemes retained in the 84 broad phonemes. (/ indicates members of one broad-phoneme class)

performed by a network assigning an English, and a Japanese language score to each incoming feature vector. Training and testing of the language classifier is done on all free speech utterances from the six languages in the database. We get 90% recognition on English vs. Japanese. This compares to 85% recognition when using two language specific phoneme recognizers in parallel. For all six languages the performance is 53% on the test set.

We have created a set of broad phonemes which span all six languages and have the ability to discriminate between the languages in the system at the phoneme level. In the next section we explore the ability of these speech units to denote language-discriminative sequences.

3. Word Spotting

3.1. Occurrence Statistics of Sequences

Our initial studies have focused on the ability to distinguish between the different languages based on the sequences of broad-phoneme labels that occur in the *hand-labeled* speech files.

After mapping the labeled files into broad phonemes, frequencies of all sequences of arbitrary length occurring in a training set of utterances are recorded in a hash table. For each language the 10 most frequently occurring sequences for each length are extracted and recorded. We now have six lists of 'words', one list for each language. Taking each list at a time, we calculate how many times on average the words in this list occur in each of the files of each of the languages. If on the average, an English 'word' occurs more frequently in, say, Japanese than in English, it is discarded. In this manner, sequences are retained if they occur on the average at least once in each file and if they are more frequent in the language for which they are used as discriminators. It is found that sequences of length longer than six do not occur more than once on average per file for this set of broad phonemes. Table 3.1 shows a list of some of the words that were retained for each of the six languages.

For each of the files in the training set the count of all occurrences of sequences in their own language are calculated and averaged. This number represents a normalizing factor to be used in the test set. In order to score the labeled files in the test set, the normalized count of occurrences of 'words' is returned for each of the language- specific lists. The utterance is classified as the language which corresponds to the highest scoring list.

Language	'words','sub-words'
English	in,so,if,to,am can,and
German	ich,am,und, dann,bin
Hindi	hai,mai,ke, ana,ek,par
Japanese	des-, -merika -kon-, ne
Mandarin	wa-, -shjen , -jou, wan tjan,-qua,yen-,-tchou
Spanish	es,en,con,por

Table 3. List of most frequent "words" (i.e. words and subwords corresponding to common sequences in hand labels)

Having done word spotting with these broad phonemes, we are now able to reevaluate the significance of each of the phonemes. The number of mono-phonemes occurring in the most frequent words is very limited. It is therefore of interest to see how different levels of clustering will perform on the language identification of the labeled files.

3.2. Language Identification

Using the algorithm outlined in the previous section, we classify all utterances from the test set. Table 3.2 shows the results. Comparative results are shown for broader phoneme clusters obtained by truncating our cluster tree at higher levels as explained above.

# BP	EN	GE	HI	JA	MA	SP
84	100%	100%	100%	94%	100%	100%
60	88%	95%	89%	94%	94%	94%
18	72%	89%	100%	100%	94%	94%

Table 4. Language recognition based on sequences in labeled files. BC = broad-phonemes used.

It can be seen that the difference in performance is minimal, except for English. The small change in performance may be because only five of the monophonemes (cCc,E,A,ai, for Mandarin and C for German) occur in the most frequent word list.

4. Towards Language Identification on Unlabeled Speech

4.1. Automatic Segmentation

Based on the previous findings we have adopted the strategy of segmenting incoming speech files at an even broader level, distinguishing between only the listed 18 classes.

- non-speech
- Vowels: A, E, I, O, U, reduced vowel
- Diphthongs: xI, xU, xa
- Unvoiced fricatives
- Voiced fricatives
- Fricated r
- Affricates
- Closures
- Plosives
- Oral Resonant Consonants: r/l
- Nasal Resonant Consonants: n/m

A neural network was trained and classifies the categories in the test set with 47% accuracy at the frame level. A separate Viterbi search (incorporating duration and bigram probabilities from all languages in the system) is performed on the outputs of the broad classifier. That is, we use these frame-based output activations to find the best scoring sequence of phoneme labels spanning the utterance.

4.2. Language Identification

As explained in Section 3.1 we now derive sequences based on which the incoming utterances are classified. It was found that it is best to use labled data for obtaining these rather than aligned data. Since the aligned files are less accurate we compensate for that in part by extracting the top 30 occurring sequences for each length as opposed to the top 10 used before.

Tables 4.2 and 4.2 show the confusion matrix obtained using the top 30 occurring sequences in the labeled files to derive the sequence lists. It can be seen that the higher accuracy of the recognizers with a smaller number of classes outweighs the performance decrease due to the loss of mono-phonemes.

Language	EN	JA	MA	SP	HI	GE
ENGLISH	9	3	1	4	1	0
JAPANESE	0	9	4	1	2	2
MANDARIN	1	0	12	3	0	1
SPANISH	1	0	3	9	2	2
HINDI	1	0	0	6	11	0
GERMAN	0	1	3	1	1	14

Table 5. Confusion matrix for identification of files aligned with 18 broad phonemes (59% correct).

Language	EN	JA	MA	SP	HI	GE
ENGLISH	8	1	2	1	2	4
JAPANESE	1	7	4	0	5	1
MANDARIN	1	0	11	1	2	2
SPANISH	2	3	4	4	2	2
HINDI	2	0	4	1	10	2
GERMAN	0	1	5	4	1	7

Table 6. Confusion matrix for identification of files aligned with 84 broad phonemes (35% correct).

By comparing the list of sequences useful to the aligned files with the list used for the labled files, we can find which of the sequences the alignment was not able to identify reliably even though they are highly useful in the recognition process. This allows us to evaluate the deficiencies of the automatic alignment. For example the alignment has trouble with plosives follwed by r/l which are very common in English. In general it can be seen that the shortcomings of the automatic alignment lie with the oral and nasal resonant consonants, (n, m, r, and l), especially in certain context as was essential here. The second finding was that it is important for our purpose to recognize the correct broad vowel category between consonants. This includes an emphasis on reduced vowels as well. Hindi, Japanese, and Spanish for example share a lot of the same short consonant-vowel-consonant sequences in the top 30 occurrences, distinguished only by the vowel.

5. Conclusion

We have shown that good language identification is possible with broad labels, as long as accurate recognition is achieved. The degree to which we can decrease the granularity of the broad phonemes without losing language discriminative power is an indication of how accurate our broad-phoneme recognizer has to be. While retreating to broad-phonemes, we recognize the importance of monophonemes which are lost in the process. A final solution might entail a combined system of sequence spotters in which broad-phonemes are used and phoneme-spotters in which emphasis is placed on recognizing language-dependent phonemes. Future work will include the improvement of the automatic alignment by addressing the problems discussed in the last section. The sequence spotter will then be evaluated on automatically labeled data. Finally our current system requires the sequences of broad phonemes we use as 'target words' to be matched perfectly during recognition. By allowing partial matches, we hope to compensate for the imperfections of our phonemic recognition.

REFERENCES

[1] K. Berkling, T. Arai, E. Barnard, and R.A.Cole. Analysis of phoneme-based features for language identification. In *Proceedings 1994 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages I–289 – I–292, Adelaide, Australia, April 1994.

[2] James L. Hieronymous. Ascii phonetic symbols for the world's languages: Worldbet. *Journal of the International Phonetic Association*, 1993.

[3] L. F. Lamel and J-L. S. Gauvain. Language identification using phone-based acoustic likelihoods. In *Proceedings 1994 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages I–293 – I–296, Adelaide, Australia, April 1994.

[4] T. Metzler and T. Lander. The cslu labeling guide. Technical Report CSLU 003, Oregon Graduate Institute, 1994.

[5] Y. K. Muthusamy, K. M. Berkling, T. Arai, R. A. Cole, and E. Barnard. A comparison of approaches to automatic language identification. In *Proceedings Eurospeech 93*, pages 1307–1310, Berlin, Germany, September 1993.

[6] Y. K. Muthusamy and R. A. Cole. Automatic segmentation and identification of ten languages using telephone speech. In *Proceedings 1992 International Conference on Spoken Language Processing*, pages 1007–1010, 1992.

[7] M. A. Zissman and E. Singer. Automatic language identification of telephone speech messages using phoneme recognition and n-gram modelling. In *Proceedings 1994 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages I–305 – I–308, Adelaide, Australia, April 1994.