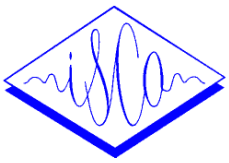

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

AUTOMATIC SEGMENTATION AND IDENTIFICATION OF TEN LANGUAGES USING TELEPHONE SPEECH

Yeshwant K. Muthusamy and Ronald A. Cole

Center for Spoken Language Understanding
Oregon Graduate Institute of Science and Technology
19600 NW Von Neumann Drive, Beaverton, OR 97006-1999

10.21437/ICSLP.1992-309

ABSTRACT

This paper extends our previous work on automatic language identification using 4 languages and high-quality speech, to automatic identification of 10 languages using telephone speech. The systems described here consist of two parts: (a) segmentation of telephone speech into seven broad phonetic categories and (b) classification of languages using feature measurements derived from the broad phonetic categories. Both the segmentation and classification stages use fully connected, feed-forward neural networks. When tested on new speakers from the 10 languages, the multi-language segmentation algorithm agrees with the hand-labels 79.8% of the time. Classifiers were trained to identify (i) all 10 languages, (ii) each language vs. all others, (iii) the pairs English-L, where L is one of the remaining 9 languages, and (iv) the triples English-L-*Other*, where *Other* consists of the remaining 8 languages. Performance varied from 47.7% for the single 10-language network to 88.6% for the English-Tamil network. Classification performance of human listeners on short excerpts of speech is also reported.

INTRODUCTION

A segment-based approach to automatic language identification assumes that each language in the world has a unique acoustic structure, and that this structure can be defined in terms of phonetic and prosodic features of speech. Phonetic, or segmental features, include the inventory of phonetic segments and their frequency of occurrence in speech. Prosodic information consists of the relative durations and amplitudes of sonorant (vowel-like) segments, their spacing in time, and patterns of pitch change within and across these segments.

To the extent that these assumptions are valid, languages can be identified automatically by segmenting speech into broad phonetic categories, computing features that capture the relevant phonetic and prosodic structure, and training a neural network classifier to associate the feature measurements with the spoken language.

We present preliminary results of a set of automatic language identification experiments in which features derived from broad phonetic categories are used to discriminate among utterances from speakers of the 10 languages. We also describe perceptual experiments that examined human classification performance on 1-, 2-, 4- and 6-second excerpts of speech from each of the 10 languages.

MULTI-LANGUAGE TELEPHONE SPEECH CORPUS

The segmentation and classification algorithms were developed and evaluated using the OGI Multi-language Telephone Speech Corpus, described in [5]. Human perceptual experiments were also performed with excerpts of speech from this corpus. The languages in the corpus are: English, Farsi (Persian), French, German, Korean, Japanese, Mandarin Chinese, Spanish, Tamil and Vietnamese. For these experiments, utterances from the first 70 valid calls in each language were used.

BROAD PHONETIC CATEGORY SEGMENTATION

Segmentation is performed by a fully-connected, feed-forward three-layer neural network that assigns 7 broad phonetic category scores to each 3 ms time frame, similar to the one described in [4]. The 7 broad phonetic categories are: vowel (VOC), fricative FRIC), stop (STOP), pre-vocalic sonorant (PRVS), inter-vocalic sonorant (INVS), post-vocalic sonorant (POVS), and silence or background noise (CLOS).

The input to the segmenter consists of 120 spectral features derived from a PLP analysis [3] of the waveform. The features were empirically derived to capture the contextual information in the vicinity of each frame [2].

Segmenter Training

The segmentation algorithm was trained and tested on utterances from the first 25 valid calls in each language. The training set consisted of 300 utterances; 2 per call for 15 calls from each language. The development test set consisted of 100 utterances; 2 per call from a different set of 5 calls in each language. The final test set also consisted of 100 utterances; 2 per call from yet another set of 5 calls in each language. The average duration of the utterances was 4.0 seconds. The training and test utterances were hand-labeled by experts with the seven broad phonetic categories.

Since it was not feasible to train the network on each time frame of each utterance, frames were chosen at random from the hand-labeled utterances in the training set. The network was trained using backpropagation with conjugate gradient optimization [1]. The frame-by-frame outputs of the segmenter were converted into a time-aligned sequence of the 7 broad phonetic category labels using a Viterbi search with duration and bigram constraints.

Segmenter Scoring

The segmenter was scored on the final test set using two different scoring procedures. In the first one, the labels output by the segmenter were compared frame-by-frame with the hand-labels, and the percentage of total frames in agreement was computed. In the second method, a string alignment and scoring program developed by NIST[1] was used, which measures the number of insertions, deletions and substitutions in the segmenter output with respect to the hand-labels. With the first method, the performance accuracy was 79.8%. This compares favorably with 85.1% for 4 languages using high-quality speech [4]. With the NIST algorithm, the performance accuracy was 72.2%, with 81.8% correct, 6.9% substitutions, 11.2% deletions, and 9.7% insertions.

LANGUAGE CLASSIFICATION

Data Sets

For the language classification experiments, only the spontaneous speech utterances from the first 70 valid calls in each language were used. The training set consisted of 2714 utterances (from 342 males and 158 females); 2-6 utterances per call for 50 calls in each language. The development test set consisted of 1120 utterances (from 151 males and 49 females); 2-6 utterances per call for 20 calls in each language. The average duration of the utterances was 13.4 seconds.

Feature Development

The language classifiers used features computed on the waveform itself and on the time-aligned sequence of broad phonetic labels. There are currently no spectral features based on PLP coefficients.

The design of the features was aided by statistical analyses of the broad phonetic sequences and phonological knowledge of the languages. For example, the presence of tonal languages like Mandarin Chinese and Vietnamese in the data set led us to design pitch features that captured the large variation in pitch within and across segments for utterances from these two languages. Similarly, the fact that syllables are about equally spaced in Japanese led us to include an "inter-segment duration difference" feature.

The current set of 194 features is described below. All the features were computed over the entire length of the utterance and were designed to yield the same number of values regardless of the duration of the utterance. The numbers in parentheses refer to the number of values generated.

- Intra-segment pitch variation: Average of the standard deviations of the pitch within all sonorant segments—VOC, PRVS, INVS and POVS (4 values)
- Inter-segment pitch variation: Standard deviation of the average pitch in all sonorant segments (4 values)
- Frequency of occurrence (number of occurrences per second of speech) and rate of occurrence (number of occurrences per segment of speech) of triples of segments. A total of 58 segment-triples were selected based on statistical analyses of the training data. (116 values)
- Frequency of occurrence of each of the seven broad phonetic labels (7 values)
- Frequency of occurrence of all segments (number of segments per second) (1 value)
- Frequency of occurrence of all sonorants (1 value)
- Frequency of occurrence of all consonants (STOPs and FRICs) (1 value)
- Frequency of occurrence of voiced consonants (1 value)
- Ratio of number of occurrences of each of the seven broad phonetic labels to the total number of segments (7 values)
- Ratio of number of sonorant segments to total number of segments (1 value)
- Ratio of number of consonant segments to total number of segments (1 value)
- Ratio of number of voiced consonants to total number of segments (1 value)
- Fraction of the total duration of the utterance devoted to each of the seven broad phonetic labels (7 values)
- Fraction of the total duration of the utterance devoted to all sonorants (1 value)
- Fraction of the total duration of the utterance devoted to all voiced consonants (1 value)
- Average duration of the seven broad phonetic labels (7 values)
- Standard deviation of the duration of the seven broad phonetic labels (7 values)
- Segment-pair ratios: conditional probability of occurrence of selected pairs of segments. The segment-pairs were selected based on histogram plots generated on the training set. Examples of selected pairs: POVS-FRIC, VOC-FRIC, INVS-VOC, etc. (22 values)
- Inter-segment duration difference: Average absolute difference in durations between successive segments (1 value)
- Standard deviation of the inter-segment duration differences (1 value)
- Average distance between the centers of successive vowels (1 value)
- Standard deviation of the distances between centers of successive vowels (1 value)

Classification Experiments

The following classification experiments were conducted, all using the same set of features:

- A single network to classify all 10 languages
- A single network to classify English, Japanese, Mandarin Chinese and Tamil (for the sake of comparison with the high-quality system described in [4]).
- Nine English-L' networks, where L' is one of the remaining 9 languages
- Ten L-*Other* networks, where L is one of the ten languages
- Nine English-L'-*Other* networks

[1] National Institute of Standards and Technology.

All the networks were trained with backpropagation with conjugate gradient optimization. Approximately equal numbers of utterances were chosen at random for the *Other* languages.

Results

The networks were evaluated on the development test set. The 10-language network performed at an accuracy of 47.7%. The 4-language network performed at an accuracy of 70.6%. In comparison, the corresponding 4-language classifier trained on high-quality speech performed at an accuracy of 89.5% on test utterances that were 17.1 seconds long on the average. The results of the remaining three experiments are shown in Tables 1, 2, and 3:

- It can be seen that the English-L' classification (Table 1) is the least difficult, with peformances ranging from 70.0% (English-French) to 88.6% (English-Tamil), with a median accuracy of 78.0% (English-Japanese).
- Classification of individual languages against all others (*L-Other*) produces about the same level of performance (Table 2), from 69.5% (English-Other) to 86.0% (Tamil-Other), with a median accuracy of 78.4% (Vietnamese-Other, Spanish-Other).
- English-*L'-Other* classification is more difficult (Table 3), with performances ranging from 56.1% (English-French-Other) to 65.9% (English-Tamil-Other) with a median accuracy of 62.5% (English-Mandarin-Other).

Table 1: Results of the English-L' Experiment

Network	*Accuracy (%)*
English-Tamil	88.6
English-Vietnamese	80.2
English-Spanish	78.8
English-Mandarin	78.6
English-Japanese	78.0
English-German	77.7
English-Farsi	77.0
English-Korean	73.2
English-French	70.0

Table 2: Results of the *L-Other* Experiment

Network	*Accuracy (%)*
Tamil-Other	86.0
Mandarin-Other	83.9
German-Other	82.3
Farsi-Other	79.5
Vietnamese-Other	78.6
Spanish-Other	78.1
Japanese-Other	74.6
Korean-Other	70.1
French-Other	69.5
English-Other	69.5

Table 3: Results of the English-*L'-Other* Experiment

Network	*Accuracy (%)*
English-Tamil-Other	65.9
English-German-Other	64.4
English-Farsi-Other	63.9
English-Spanish-Other	62.7
English-Mandarin-Other	62.5
English-Japanese-Other	62.0
English-Vietnamese-Other	61.7
English-Korean-Other	60.1
English-French-Other	56.1

HUMAN LISTENING EXPERIMENTS

To determine human listening performance on excerpts of speech from the 10 languages, 7 male and 4 female monolingual native English speakers were presented with 1-, 2-, 4- and 6-second excerpts of spontaneous speech excised from the 10 languages.

Experimental Procedure

The experiment was conducted using an interactive graphics program that played excerpts of speech chosen at random from each of the 10 languages, and maintained a log of subject responses. Following a brief training session, subjects were presented with 760 different excerpts, 19 at each duration from each language. The subjects could listen to each excerpt as many times as they desired. After responding, they were given feedback on every trial. The subjects could also listen to an excerpt *after* making the choice—a feature that was included to aid in the learning process. Each block of 100 trials was considered a session, and the program automatically quit after every 100 trials, to ensure that the subjects did not get fatigued.

Results

The average listener performance for each language is shown for the four durations in Figure 1. The average performance over all languages rose from 37.0% to 43.0% to 51.2% to 54.6% as duration increased from 1 to 2 to 4 to 6 seconds respectively. As expected, the listeners recognized English with high accuracy. Also, note the relatively high performance on French, German and Spanish—languages that the listeners were most often exposed to, either through courses or by contact with foreign friends. Performance on Farsi, Korean, Tamil and Vietnamese—languages that the listeners had never been exposed to, was very poor.

Analysis of performance by each block of 190 trials revealed little evidence of learning during the experiment. For example, for 6-second excerpts, the average performance on Korean for the first and last 190 trials was 13.5% and 16.7% respectively.

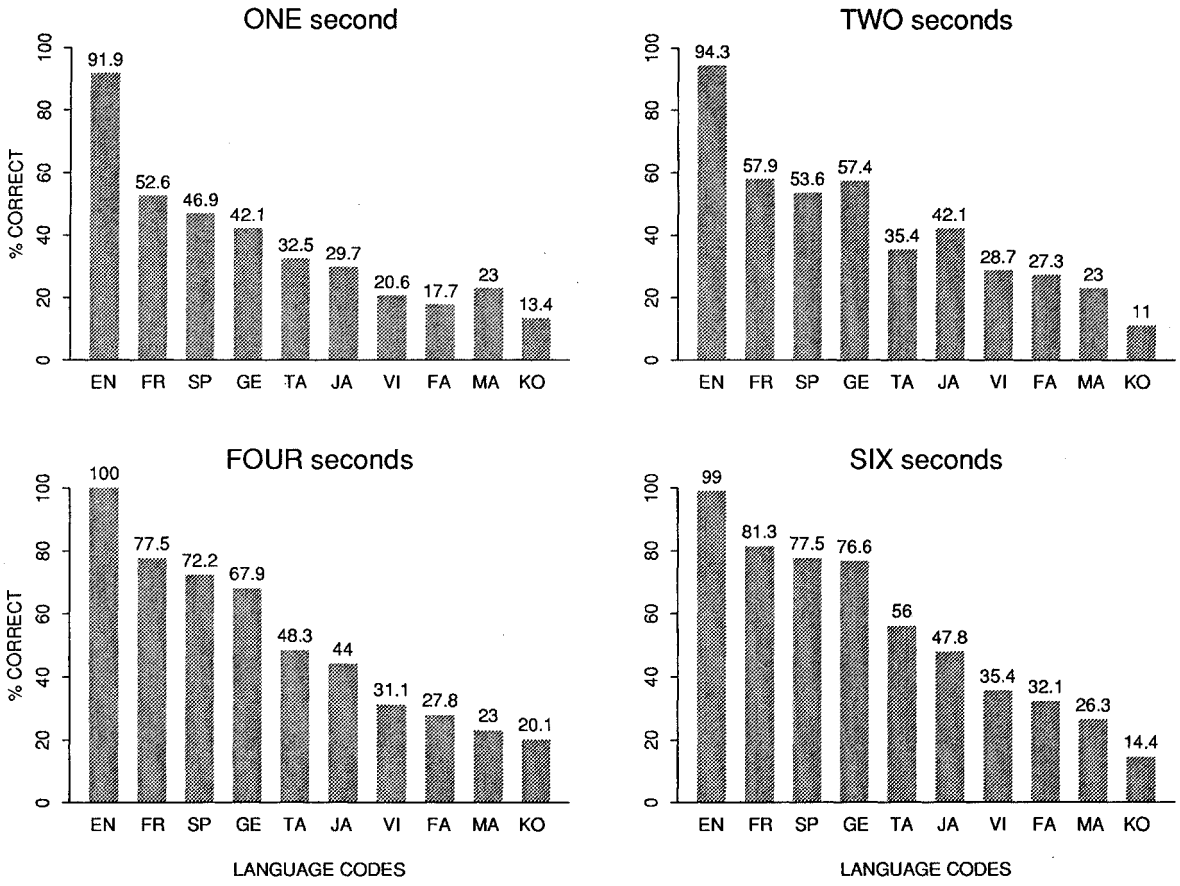

Figure 1: Average Listener Performance for the Four Durations

SUMMARY AND FUTURE WORK

The results of the classification experiments indicate that performance on constrained tasks (fewer languages) is reasonably high, and drops steeply as more languages are added. The feature analysis and development process is by no means complete. The current feature set does not include any spectral information. The addition of PLP-based features and pitch contours (that capture the intonation contour across the utterance) is likely to improve the classification performance of all networks. This research is now underway.

ACKNOWLEDGEMENTS

The authors thank Mark Fanty and Todd Leen for their useful comments and discussions, and Terri Durham for running the perceptual experiments.

REFERENCES

[1] E. Barnard and R. A. Cole. A neural-net training program based on conjugate-gradient optimization. Technical Report CSE 89-014, Department of Computer Science, Oregon Graduate Institute of Science and Technology, 1989.

[2] Mark Fanty, Ronald A. Cole, and Krist Roginski. English alphabet recognition with telephone speech. In J. E. Moody, S. J. Hanson, and R. P. Lippmann, editors, *Advances in Neural Information Processing Systems 4*, San Mateo, CA, 1992. Morgan Kaufmann Publishers.

[3] H. Hermansky. Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America*, 87:1738–1752, April 1990.

[4] Y. K. Muthusamy and R. A. Cole. A segment-based automatic language identification system. In J. E. Moody, S. J. Hanson, and R. P. Lippmann, editors, *Advances in Neural Information Processing Systems 4*, San Mateo, CA, 1992. Morgan Kaufmann Publishers.

[5] Y. K. Muthusamy, R. A. Cole, and B. T. Oshika. The OGI multi-language telephone speech corpus. In *Proceedings International Conference on Spoken Language Processing 92*, Banff, Alberta, Canada, October 1992.