Eurospeech (ESCA), pp. 1307-1310, 1993

A COMPARISON OF APPROACHES TO AUTOMATIC LANGUAGE IDENTIFICATION USING TELEPHONE SPEECH

Yeshwant Muthusamy *Kay Berkling* *Takayuki Arai* *Ronald Cole* *Etienne Barnard*

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

A variety of approaches to language identification, based on (a) acoustic features, (b) broad-category segmentation, and (c) fine phonetic classification, are introduced. These approaches are evaluated in terms of their ability to distinguish between English and Japanese utterances spoken over a telephone channel. It is found that the best performance (86.3 % accurate classification of utterances with a mean length of 13.4 sec) is obtained when fine phonetic features are employed. In addition, the results show the importance of discriminatory training rather than likelihood estimation.

1. INTRODUCTION

As developments in telecommunications and long-distance travel cause national borders to become increasingly transparent, the ability to identify which language is being spoken is growing in importance. The utility of tasks such as directory assistance or automatic translation is, for instance, improved substantially by the availability of a means of identifying which language is being spoken. Given the scarceness of humans able to identify several languages, and the economic disadvantages of employing such people for this task, it is clear that automatic language identification is of much practical importance.

Automatic language identification is also very interesting for theoretical reasons. It resembles automatic speech recognition and automatic speaker identification in several ways, but also differs in important respects from both those tasks. It is, for example, neither necessary nor sufficient for the purposes of automatic language identification to recognize each of the phonemes occurring in an input speech signal.

Because automatic language identification is a relatively new pursuit, it is currently not clear which approaches to this task promise the best performance. As an initial attempt to understand what the possible approaches are, and what advantages and disadvantages can be expected from each, we present and compare three conceptually distinct techniques for performing this task. These techniques are based on (a) raw acoustic features, (b) broad phonetic categories, and (c) fine phonetic categories.

To simplify matters, we limit our task to the distinction between English and Japanese utterances. This will provide us with a basic comparison between the various approaches, which will be refined by the introduction of additional languages in future work. Also, our focus is on telephone speech, since this is likely to be the most important application of automatic language identification in the near future. The data employed are taken from the OGI-TS corpus [7], and are described in more detail in Section 3.

In section 2 we describe the three approaches to automatic language identification (LID) studied here. Section 3 describes the results obtained with these approaches, and a discussion is contained in Section 4.

2. THREE APPROACHES TO LANGUAGE IDENTIFICATION

This section describes (a) a one-stage approach based on raw acoustic features, (b) a two-stage approach – broad phonetic segmentation [5] is performed in a first stage before extracting features for language classification in a second stage –, and (c) another two-stage approach, in which the broad-category segmentation in (b) is replaced by a fine phonetic classifier.

All neural network classifiers used here are fully-connected, feed-forward networks trained using backpropagation with conjugate gradient optimization [1]. The number of hidden nodes was derived experimentally for each case.

In all cases the acoustic representation used is a seventh order Perceptual Linear Predictive (PLP) model [4], yielding 8 coefficients (including one for energy). This is computed on a 10 msec interval of speech, time shifted every 3 msec for the approaches based on acoustic features and broad-category segmentation, and every 6 msec for the fine-phonetic approach.

2.1. Identification using acoustic features

A baseline approach to using spectral features is to classify each frame as one of the two languages using a neural network, and accumulate the network output activations across all frames of the utterance for each language. The language with the maximum accumulated activation score is the winner.

The spectral features consisted of 56 averaged PLP coefficients – the abovementioned 8 coefficients averaged within each of 7 regions spanning a 171 ms window centered on the frame to be classified. The sampling intervals are shown in Figure 1. The objective was to provide substantial contextual information about the chosen frame to the network.

Frames were sampled at fixed intervals from each utterance. Several network configurations were experimented with. The best network performance on a frame-by-frame base was 59%, obtained from a network with 48 hidden nodes trained on data with a 24 ms sampling interval. The performance of this approach on whole utterances is listed in Section 3.

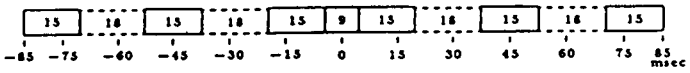

Figure 1. Spectral Experiment I: Sampling intervals for the PLP features. The solid boxes indicate the intervals over which PLP coefficients are averaged. Dashed boxes indicate intervals that are skipped.

2.2. Identification using broad phonetic categories

2.2.1. Broad-Category Segmenter

Broad-category 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 of the utterance [6]. The broad phonetic categories are: vowel (VOC), fricative (FRIC), stop (STOP), pre-vocalic sonorant (PRVS), inter-vocalic sonorant (PRVS), post-vocalic sonorant (POVS), and closure or silence or background noise (CLOS).

The input to the segmenter consists of 120 spectral features derived from a PLP analysis [4] of the waveform. The features were empirically derived to capture the contextual information in the vicinity of each frame [2].

A Viterbi search, which incorporates duration and bigram probabilities from ten languages, uses these frame-based output activations to find the best scoring sequence of broad phonetic category labels spanning the utterance.

Several networks were trained and evaluated on the development set. The best network performance was 71.6% on the development set.

2.2.2. Language Classification

Two types of classifiers were trained based on the output of the broad-category segmenter – one used various measures based on bigram occurrences, and the other employed a window of outputs from the broad-category classifier as input.

Broad Category Bigrams There are 20 legal segment-pairs of the seven broad phonetic categories, VOC, FRIC, CLOS, STOP, PRVS, INVS and POVS. Four feature sets based on segment-pairs were examined:

- Segment-pair Frequency (SPF): number of occurrences of each segment-pair per second of speech, and
- Segment-pair Ratio (SPR): ratio of the number of occurrences of each segment-pair to the total number of segments in the utterance
- Segment-pair Median Duration (SPMD): median duration of each segment-pair in an utterance
- Segment-pair Duration Ratio (SPDR): ratio of the total duration of all occurrences of a segment-pair in an utterance to the total utterance duration

There was some separation in the distribution for SPF, SPR and SPDR for the following segment pairs: 1)POVS-FRIC, 2)CLOS-STOP 3)VOC-CLOS, and 4)VOC-POVS. Of these, VOC-POVS had the maximum difference in distributions. Interestingly, no such separation was evident for the SPMD feature set indicating that English and Japanese differed only in the number of VOC-POVS pairs rather in their median durations. This yields 12 features to be used for training and testing of the neural network classifier.

Window of Segments Features from a moving window of $N = 15$ segments were presented to the network at a time. For each segment in a window, the following feature measurements were made:

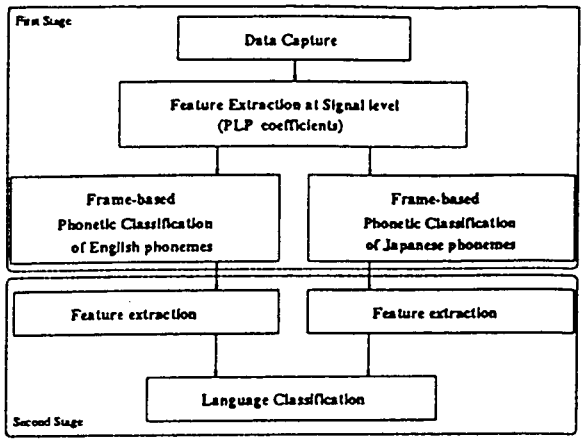

Figure 2. Modules of the Two-Stage Language Identification System

- The segment duration
- A vector of length seven, representing the broad phonetic label with the averaged frame-based scores from the segmenter network for each of the seven phonetic categories.

The network produced 2 language activation scores for each window of N segments, and the window was shifted frame-by-frame over the whole utterance. The 2 language scores were accumulated as the window progressed through the utterance. When the end of the utterance was reached, the language with the maximum accumulated activation score was taken as the system response.

The network had 120 ($= 15 \times 8$) input neurons, 32 hidden neurons, and 2 output neurons. The percentage of segment windows correctly classified was 66.8%.

2.3. Identification using fine phonetic categories

This section describes a two-stage system depicted in Figure 2. In the first stage, features are derived from the signal in order to perform a frame-based phoneme classification of the incoming speech signal. Features are then derived from this classifier output and used to perform language classification in the second stage.

2.3.1. Phoneme Classification

For each sampled frame, 56 ($= 8 \times 7$) PLP coefficients within a 174 msec window, centered on the frame to be classified, were computed and served as input to each of the phonetic classifiers [3].

The English classifier assigns 39 phonemic category scores to each 6 msec time frame. The 39 labels provide a quasi-phonemic level of description, in which most allophonic variations are ignored. Similarly a Japanese network is trained with 25 output nodes representing each of the phoneme categories. The English and Japanese phonetic classifiers perform with 48% and 46% accuracy, respectively, when evaluated on a test set of hand labeled speech from each language.

Two Viterbi searches, each incorporating duration and bigram probabilities from the corresponding language, use these frame-based output activations to find the best scoring sequence of phoneme labels spanning the utterance.

2.3.2. Language Classification

Language classification is performed based on (a) unigram features and (b) bigram features.

Unigrams In the second stage three groups of 64 unigram features each are derived from the outputs of the two classifiers. For each of the 64 English and Japanese phonemes three features are derived independently of each other, resulting in a 3 x 64 = 192 element feature vector representing an utterance. Language classification is performed by a network assigning both an English and a Japanese language score to each incoming feature vector.

The three features extracted from each output of the first-stage classifier are described below.

Average Output Activation. (Average H)

$$\text{Average H}_i = \frac{1}{N}\sum_{t=1}^{N} d_t(p_i), \qquad (1)$$

where $d_t(p_i)$ denotes the activation of the i^{th} phoneme at time t, and N the number of frames over which the feature was extracted.

Maximum Output Activation. (Maximum)

$$\text{Maximum}_i = \max_{1 \le t \le N} d_t(p_i), \qquad \forall 1 \le i \le I \qquad (2)$$

where I is the number of phonemes. Rare occurrences of language-specific phonemes may be rendered invisible in the averaging process. This feature is designed to overcome such limitations.

Variation in Output Activation. (Variation H)

$$\text{Variation H}_i = \frac{1}{N}\sum_{t=1}^{N}[d_t(p_i) - \text{Average H}_i]^2 \qquad (3)$$

For completeness, the language classifier was trained on the 64-dimensional vectors resulting from each of the features in isolation as well as on the 192-dimensional vector resulting from combining all three features.

Bigrams To extend this approach, we also studied features derived from the transition probabilities between pairs of phonemes – i.e., using bigram probabilities.

The frame-by-frame outputs of the English phoneme classifier described above were converted into a time-aligned sequence of the phoneme labels by applying minimum and maximum duration constraints of 3 msec. and 300 msec., respectively. In principle, we would like to use the transition frequencies for all consecutive pairs of phonemes as features to the classifier, but the number of possible pairs is 39 x 38, which is too large given the amount of training data available. We therefore limited our features to N transition probabilities, namely those probabilities whose average values over the Japanese and English training sets differed the most. N was varied for optimal performance (see Section 3).

Thus, the transition probabilities of these N pairs were used as input to a neural-net classifier, which was trained to produce a Japanese-English distinction as output.

Combining unigram and bigram features Finally, both unigram and bigram features are combined. The input feature vector to the language classifier now consists of the 3 unigram features derived for each of the 64 phonemic outputs of the first-stage classifiers resulting in a 192 (64x3) element vector, concatenated with the N occurrence frequencies of the most common pairs.

Label	*Training Set*		*Development Set*	
	Tokens	*Frames*	*Tokens*	*Frames*
VOC	3274	6251	1297	5578
FRIC	1314	6336	473	5353
CLOS	2413	6530	787	5284
STOP	1150	6523	442	5795
PRVS	383	6568	124	4991
INVS	887	6424	390	6037
POVS	648	6462	265	5854
TOTAL	10069	45094	3778	38892

Table 1. Distribution of Tokens and Frames in the Training and Development Sets for the broad category segmenter

	Train		Test	
	Utterances	Frames	Utterances	Frames
English Phonetic Classifier	50	80502	20	7441
Japanese Phonetic Classifier	35	20326	10	5138

Table 2. Division of Labeled Data into Training and Test set for the phonetic classifier

An alternative way to combine unigram and bigram probabilities was also investigated. Language likelihoods were estimated with a Viterbi search using unigram and bigram probabilities. The Viterbi search maximizes

$$L_t = \max\left[L_{t-1} + \log\{P(p_j|p_i)\cdot o(j,t)\}\right], \qquad (4)$$

where L_t is logarithmic likelihood at frame t, $P(p_j|p_i)$ is the transition probability from phoneme p_i to p_j, and $o(j,t)$ is the activation of phoneme p_j from the phoneme classifier at frame t. For each language, $P(p_j|p_i)$ for all i and j were estimated from the training sets, and the classifier assigned an incoming utterance to the language with the largest likelihood according to this expression. (In practice, the likelihood for Japanese was scaled by a factor slightly larger than 1 to account for the fact that certain English phonemes do not occur in Japanese. This scaling factor was chosen to give optimal training-set performance.)

3. RESULTS

3.1. Training and test data

3.1.1. Multi-Language Telephone Speech Corpus

The segmentation and classification algorithms were developed and evaluated using the OGI Multi-language Telephone Speech Corpus, described in [7]. Both the training and test utterances were hand-labeled with the seven broad phonetic category or phoneme labels as necessary.

3.1.2. Training and Test Sets: First stage

To limit computational requirements, the fine phonetic and broad phonetic classifiers were trained from a randomly chosen subset of the frames from the training utterances.

Table 1 displays the distribution of tokens and frames of each broad phonetic category used to train the network for broad-category segmentation. The numbers in the *Frames* column for both the training and development sets include 3000 edge-sampled frames, the balance being made of randomly sampled frames. These frames were sampled from hand labeled data in ten languages.

Approach	Features	Performance
Acoustic features	PLP	70.0%
Broad Category features	Pairs Window	75.8% 83.2%
Fine Phonetic Unigram features	Average H Maximum Variation H Combined	76.2% 81.5% 80.2% 82.3%
Fine Phonetic Bigram features	20 pairs 100 pairs 200 pairs	74.0% 77.5% 79.3%
Bigram + Unigram	Viterbi likelihood 100 pairs + unigram 200 pairs + unigram	81.1% 85.5% 86.3%

Table 3. Summary of Performance for all Approaches

Table 2 shows the number of frames in the training and test sets for the phonetic classifiers. The utterances in the Table refer to stories of up to 50 seconds of extemporaneous speech which have been handlabled at the phonemic level.

3.1.3. Training and Test Sets: Second stage

The language classifiers were trained and evaluated on only the spontaneous speech utterances from the first 70 valid calls in each language (a subset of the calls in the training set was used to train the broad category and fine phonetic classifiers). The development test set consisted of 2-6 utterances per call for 20 calls in each language. The utterances ranged in duration from 1 second to 49 seconds with an average of 13.4 seconds.

3.2. Comparative results

Results obtained with the best classifier of each type described above are listed in Table 3. It can be seen that the best classifier using fine phonetic distinctions slightly outperforms the best classifier based on broad-category segmentation (86.3 % vs. 83.2 %), and that the classifier using only acoustic features is substantially inferior to these two methods.

Also, the combination of unigram and bigram features is substantially better than either feature type individually. Of the unigram features, the "maximum-output" feature was most useful.

4. DISCUSSION

In going from acoustic features to broad-category features to fine phonetic features, the inputs to our classifier become increasingly imprecise (since the fine phonetic classifiers are relatively unreliable, and the broad-category segmenter is also not perfect). However, our results indicate that these imperfections are more than compensated for by the more detailed information that the classifiers of lesser accuracy provide.

It is interesting to note that the combination of fine-phonetic unigram and bigram features using a neural-network classifier is significantly superior to the approach based on likelihood estimation by way of a Viterbi search. This result emphasizes the desirability of discriminatory training; this is hardly surprising, since it is hard to imagine a theoretically accurate way of computing the likelihood of utterances given a language identity.

Several extensions of this work are now being investigated. As was mentioned above, it is necessary to study the generality of our results with respect to other languages; we are therefore also comparing various approaches on the full ten-language set in the OGI Corpus [7]. In addition, the successes of the fine-phonetic approach have prompted us to design methods which perform even more detailed classifications in the first stage. These methods are now being evaluated.

5. ACKNOWLEDGEMENTS

We thank Beatrice T. Oshika for helping with the transcription conventions and Takayuki Arai for transcribing the Japanese utterances. Research was supported in part by the Department of Defense and grants from the Office of Naval Research and the National Science Foundation.

REFERENCES

[1] E. Barnard and R. Cole. A neural-net training program based on conjugate-gradient optimization. Technical Report CSE 89-014, Oregon Graduate Institution, 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*, pages 199–206, San Mateo, CA, 1991. Morgan Kaufmann Publishers.

[3] Mark Fanty, Philipp Schmid, and Ronald Cole. City name recognition over the telephone. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 549–552, April 1993.

[4] H. Hermansky. Perceptual Linear Predictive (PLP) analysis of speech. *J. Acoust. Soc. Am.*, 87(4):1738–1752, 1990.

[5] Y. K. Muthusamy. *A Segmental Approach to Automatic Language Identification*. PhD thesis, Oregon Graduate Institute, July 1993.

[6] Y. K. Muthusamy and R. A. Cole. Automatic segmentation and identification of ten languages using telephone speech. In *Proceedings International Conference on Spoken Language Processing 92*, Banff, Alberta, Canada, October 1992.

[7] 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*, pages 895–898, Banff, Alberta, Canada, October 1992.