ICASSP (IEEE), pp. I 289-292, 1994

ANALYSIS OF PHONEME-BASED FEATURES FOR LANGUAGE IDENTIFICATION

Kay M. Berkling, Takayuki Arai, and 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

This paper presents an analysis of the phonemic language identification system introduced in [5], now extended to recognize German in addition to English and Japanese. In this system language identification is based on features derived from a superset of phonemes of all three languages. As we increase the number of languages, the need to reduce the feature space becomes apparent. Practical analysis of single-feature statistics in conjunction with linguistic knowledge leads to 90% reduction of the feature space with only a 5% loss in performance. Thus, the system discriminates between Japanese and English with 84.1% accuracy based on only 15 features compared to 84.6% based on the complete set of 318 phonemic features (or 83.6% using 333 broad-category features [4]). Results indicate that a language identification system may be designed based on linguistic knowledge and then implemented with a neural network of appropriate complexity.

1. INTRODUCTION

In [5] we introduced a language-identification system based on phoneme recognition. In that system we employed a separate phonemic front end for each language to be recognized, and computed features based on the outputs of all these front ends. With this approach, the number of inputs to the language classifier is proportional to the sum of the numbèrs of phonemes occurring in each of the languages to be recognized. Clearly, systems cannot scale well with such a large number of features. Since the phonemic level of transcription allows for a wide variety of allophonic variations it is not known *a priori* to which degree the allophonic realizations overlap across the languages. Rather than merging phonemes across languages we study ways of systematically reducing the number of feature inputs to the language classifier. This is accomplished by linking feature analysis to linguistic knowledge as outlined in section 2. Section 3 describes the automatic language identification (LID) system. Section 4 describes the approaches that were employed in order to reduce the feature space. Results obtained with these approaches are reported in section 5 followed by a discussion in section 6.

2. LINGUISTIC BACKGROUND

The superset of all phonemes spanning the three languages treated in this paper may be divided into the following two groups [2],

1. mono-phonemes: phonemes whose acoustic realizations in one language overlap little or not at all with those in another language (e.g. the English r vs. the German r).
2. poly-phonemes: phonemes whose acoustic realizations are similar enough across the languages to be equated (e.g. sh in English and German).

2.1. Mono-phonemes

Phonemes known to have different acoustic realizations in the languages to be identified are expected to contribute to the LID process. Below we list some of the mono-phonemes found in our set of three languages [1].

English vs. German German can be distinguished from English by the occurrence of language-specific phonemes, such as:

- The uvular fricative kx as in *Dach* is specific to German.
- The palatal fricative cx as in *ich* is specific to German.
- The interdental voiced dh and voiceless th fricatives are specific to English.
- The phoneme r in German has a complicated set of allophones which varies widely among speakers of different regional dialects. Whereas the (American) English r is an alveolar retroflex, frictionless continuant, the German r can be uvular, dental or alveolar, rolled, flapped, or fricated.
- Finally, unlike English, German has rounded high front vowels (*Guete*), mid front vowels (*Goethe*), and open central vowels (*Vaeter*).

Japanese vs. English There are several characteristics of Japanese that distinguish it from English at the phonemic level:

- The high back vowel u is unrounded.
- The Japanese r has allophones including an alveolar dental voiced flap and a lateral flap.
- The Japanese n is a voiced nasal with dental or velar allophones predictable by context. The former is close in production to the alveolar English n.

0-7803-1775-0/94 $3.00 © 1994 IEEE

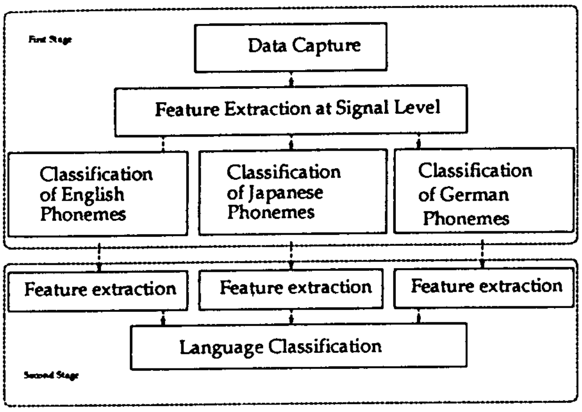
First Stage
Data Capture
Feature Extraction at Signal Level
Classification of English Phonemes
Classification of Japanese Phonemes
Classification of German Phonemes
Feature extraction
Feature extraction
Feature extraction
Language Classification
Second Stage

Figure 1. Modules of the Two-Stage LID System

Japanese vs. German Japanese differs from German in place of articulation of the n, the occurrence of palatal and uvular fricatives in the German speech as explained above, and the presence of rounded vowels in German.

2.2. Poly-Phonemes

A large set of phonemes, including stops, and fricatives: (p, t, k, b, d, g, f, sh, s, and h) appear to be poly-phonemes. These are also potentially useful for LID. We can, for instance, perform a statistical analysis on labeled data to detect phonemes which have a different frequency of occurrence between a pair of languages.

However, whereas mono-phonemes can theoretically identify a language after being detected once, features based on the occurrence statistics of poly-phonemes depend on the length of the utterance. Only over a longer utterance can a reliable statistic on frequency of occurrence for instance, be calculated. We therefore expect these phonemes to be of less importance in our current approach. On the other hand, poly-phonemes are extremely important in an approach such as word spotting, which we do not study here.

3. LANGUAGE IDENTIFICATION

We now describe the two-stage system depicted in Figure 1 for three languages. In the first stage, features are derived from an input speech signal in order to perform a frame-based phoneme classification. The features for LID are then derived from this classifier output and provided to a language classifier.

3.1. Phoneme Classification

For each sampled frame, 56 PLP coefficients from within a 174 msec window, centered on the frame to be classified, are computed and served as input to each of the phonemic classifiers.

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 and a German network are trained with respectively 25 and 42 output nodes representing each of the phoneme categories. The three phonemic classifiers perform with 48%, 46% and 39% accuracy, respectively, when evaluated on a test set of hand labeled speech from each language.

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 106 unigram features each are derived from the outputs of the three classifiers. For each of the 106 English, Japanese, and German phonemes three features are derived independently of each other, resulting in a 3 x 106 = 318 element feature vector representing an utterance. Language classification is performed by a network assigning an English, a Japanese, and a German language score to each incoming feature vector.

The three features extracted from each output of the first-stage classifiers are the average output activation **AVG** across the complete utterance, the maximum output activation **MAX**, and variation in output activation **VARH** as explained in [5].

Bigrams: 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 (derived from the labeled utterances) of 3 msec. and 300 msec., respectively, using a Viterbi search. That is, we use the frame-based output activations of the English phoneme recognizer to find the best scoring sequence of English phonemes to all utterances in either of the three languages. We now have a common set of phonemes based on which to derive features for language classification. The transition probabilities of 200 pairs (chosen for optimal performance [5]) were used as input to a neural-net classifier, trained to produce a language classification as output. Combining unigram and bigram features, the input feature vector to the language classifier now consists of the 318 unigram features concatenated with the 200 occurrence frequencies of the most common pairs.

4. FEATURE ANALYSIS

4.1. Statistical Analysis based on labeled files

The following statistical analysis of poly-phonemes was performed on labeled files in the OGI Multi-language Telephone Speech Corpus as described in section 5.1.1. In order to find those phonemes which differ in frequency of occurrence at the frame level, their average occurrence is calculated. Phoneme p occurs with frequency f_p^I in language I, (N^I denotes the total number of frames from language I and N_p^I denotes the number of frames from language I with label p):

$$f_p^I = \frac{N_p^I}{N^I} \qquad (1)$$

Table 1 lists significant phonemes p corresponding to ratios $\frac{f_p^{I1}}{f_p^{I2}}$ >= 2 as well as some other phonemes that are important for the later discussion.

The f_p^{Ii} indirectly capture both frequency and length of the poly-phonemes and their ratio identifies phonemes useful in discriminating between pairs of languages.

EN-GE	EN-JA	JA-GE
bx$^*_i$ (13) GE p (3) EN dx* (2) EN sh* (2) GE g* (2) GE k* (2) EN	f* (8) EN p* (7) EN sh*(2) JA k* (2) JA	f* (11) GE k* (4) JA p (2) GE b* (2) GE d (2) GE
f* (1.3) GE	t* (1.2) EN	h* (1.2) GE

Table 1. Phonemes that occur with significantly different frequencies in a language pair. The table reads as follows: Label (Ratio) Language with Larger Frequency of Occurrence. The * indicates a feature that was retained in the final LID system.

4.2. Analysis and Reduction of Features

Finding the best subset of the 318 variables is quite difficult. To search through all combinations of features would be prohibitive (the number of possible combinations to be searched for a 90% reduction in the number of features is equal to 10^{42}). Rather than using computationally expensive pruning techniques, feature selection is performed based on the following assumptions and consequently verified using linguistic knowledge.

1. The features are independent from each other. This is clearly not strictly true - average and maximal values for a particular phoneme will, for instance, be strongly correlated. This is, nonetheless, a useful first-order approximation.
2. The features are distributed normally. This is again not a good approximation for all features, but it gives a rough estimate that can be improved by visual inspection of the one-dimensional distributions.

With the aim of reducing the number of features to 5-10% of the original feature set, we arrive at several reductions in the feature space.

- **Set A:** Features with Bayes error less than 80%, calculated according to the assumptions above. Evident mistakes due to non-normal distributions were fixed by adding and deleting appropriate features.
- **Set B:** The subset of Set A, which agree with the linguistic predictions of mono-phonemes and polyphonemes [1] . (Table 2)
- **Set C:** When appropriate, the subset of Set B after performing an informal sensitivity analysis on the size of the weights used in the LID stage. (Table 2)

5. RESULTS

5.1. Training and test data

5.1.1. *Multi-Language Telephone Speech Corpus*

The classification algorithms were developed and evaluated using the OGI Multi-language Telephone Speech Corpus, described in [6].

	Feature	EN	JA	GE
EN vs. GE	**VARH**	er r w ow ay eh dx k th cl		kx m xy ö bx cl
	AVG	er ow eh h		f kx bx cl
	MAX	er r ay sh d cl		m g cl
EN vs. JA	**VARH**	*er *r aor *w ih *th dh *cl k	*e p *t *n *dx *cl	
	AVG	*er *r *aor *w *uw *f *cl	*e *t *cl	
	MAX	*er *r *aor *w *k *sh *cl	*e *t *cl	
JA vs. GE	**VARH**		*cl *e *n	f n m k xy oo *cl h
	AVG		*cl *e *n b *aa	h *n m ihw *xy *aa *cl
	MAX		cl n b	*h *n *m *xy *xw cl

Table 2. Features for SubSet B, and SubSet C (*) if applicable.

5.1.2. *Training and Test Sets: First stage*

Table 3 shows the number of frames in the training and test sets for the phonemic classifiers. The utterances refer to stories of up to 50 seconds of extemporaneous speech which have been hand labeled at the phonemic level.

	Train		Test	
	Utterances	Frames	Utterances	Frames
English	50	80502	20	7441
Japanese	35	20326	10	5138
German	45	71082	10	14567

Table 3. Division of Labeled Data into Training and Test set for the phonemic classifier

5.1.3. *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. 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.

5.2. Performance Analysis

5.2.1. *Language Identification with Complete Feature Set*

Table 4 displays system performance using unigram features separately and combined with bigram features. When applicable our approach is compared to that of Zissman [8] (who uses ergodic HMMs) and Muthusamy [4] (broad-category segmentation) For trilingual LID a different architecture combining three bilingual experts was also employed in order to let the architecture reflect the structure of the problem (Fig. 2). Performance improved slightly from 73.2% to 74.5% (unigrams only).

[1] The labels used here are similar to TIMIT labels. In addition *cl* stands for closure, *xy* and *xw* are dipthongs ending in a closed front vowel and closed back vowel, respectively.

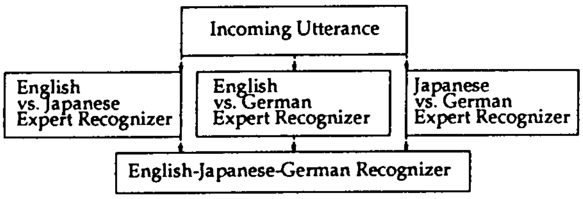
Incoming Utterance
English vs. Japanese Expert Recognizer
English vs. German Expert Recognizer
Japanese vs. German Expert Recognizer
English-Japanese-German Recognizer

Figure 2. Modules of the Two-Stage LID System

Classifier	Unigrams	Uni+Big	Muth.	Ziss.
EN-JA	84.6%	85.0%	83.6%	79%
EN-GE	85.5%	83.1%	73.6%	67%
JA-GE	78.6%	79.0%	-	-
EN-JA-GE	73.2%	74.2%	-	-

Table 4. Results

5.2.2. *Reducing the Features*

Table 5 shows the performance of the LID Systems as a function of the number of phonemes. It can be noted that the performance is highest when all features are used. In all cases however we have achieved a reduction of 90-95% with a performance loss no greater than 5%.

EN-JA		EN-GE		JA-GE	
Num	Uni	Num	Uni	Num	Uni
318	84.6%	318	85.5%	318	78.6%
56	83.3%	77	75.1%	70	72.4%
33	82.4%	32	82.6%	33	73.8%
27	82.4%	-	-	17	74.3%
m(15)	84.1%	m(18)	77.5%	m(12)	74.8%
p(12)	65.2%	p(14)	53.9%	-	-

Table 5. Reducing Features. Uni = unigrams, Num = Number of Features used, p = poly-phonemes, m = mono-phonemes

5.2.3. *Finding Major Features*

As is evident from Table 2, the number of mono-phonemes in the final feature set outweighs the number of poly-phonemes. In order to determine their contribution we retrained exclusively on the mono-phonemes from the smallest feature set. Clearly, the bulk of the information is contained in the mono-phonemes, but poly-phonemes contribute to an overall good performance as indicated by the results in Table 5.

6. CONCLUSION

In order to develop a system to distinguish among a larger number of languages, it may be useful to reduce the number of features despite a small loss in performance. As shown in this study, it is possible to limit the number of features by using known pair-wise contrastive mono-phonemes. An architecture of appropriate complexity can therefore be handcrafted and the system implemented in the described manner with neural networks. Additional features derived from poly-phonemes contribute to the discrimination process but do not contain the bulk of the information. Because these systems rely heavily on mono-phonemes, the number of features will increase steadily with the number of languages added. We therefore perceive the system described here as a building block in the hierarchical decision structure of a larger modular LID system.

6.1. Future Work

Future work will incorporate these building blocks and expand the feature set to include sequential information, such as prosody, keyword-spotting, syllable-spotting, and detection of phoneme sequences. One might also be interested in trying to capture the differences present in the signal within the poly-phoneme set to avoid this explosion of the feature space as languages are added.

7. ACKNOWLEDGEMENTS

We thank Beatrice T. Oshika for valuable conversations about the linguistic aspects of this paper, and her help with the transcription conventions. Thanks also go to the transcribers of the utterances. We also thank Ron Cole, who contributed several valuable ideas to this research. 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] Bernard Comrie. *The World's Major Languages.* Oxford University Press, 1 edition, 1990.

[2] P. Dalsgaard and O. Andersen. Identification of mono- and poly-phonemes using acoustic-phonetic features derived by a self-organizing neural network. In *International Conference on Spoken Language Processing*, pages 547–550, Banff, oct 1992.

[3] Peter Ladefoged. The revised international phonetic alphabet. *Journal of the International Phonetic Association*, 19(2):67–80, 1990.

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

[5] Y. K. Muthusamy, K. Berkling, T. Arai, R. Cole, and E. Barnard. A comparison of approaches to automatic language identification using telephone speech. In *Eurospeech*, volume 2, pages 1307–1310, sep 1993.

[6] Y. K. Muthusamy, R. A. Cole, and B. T. Oshika. The OGI multi-language telephone speech corpus. In *International Conference on Spoken Language Processing*, pages 895–898, oct 1992.

[7] Y. K. Muthusamy, Neena Jain, and Ronald A. Cole. Perceptual benchmarks for automatic language identification. In *International Conference on Speech and Signal Processing*, Alberta, Australia, apr 1994.

[8] Marc A. Zissman. Automatic language identification using gaussian mixtures and hidden markov models. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 399–402, apr 1993.