Language Identification: Insights from the Classification of Hand Annotated Phone Transcripts

Timothy Kempton, Roger K. Moore

Department of Computer Science, University of Sheffield, UK

t.kempton@dcs.shef.ac.uk, r.k.moore@dsc.shef.ac.uk

Abstract

Language Identification (LID) of speech can be split into two processes; phone recognition and language modelling. This two stage approach underlies some of the most successful LID systems. As phone recognizers become more accurate it is useful to simulate a very accurate phone recognizer to determine the effect on the overall LID accuracy. This can be done by using phone transcripts. In this paper LID is performed on phone transcripts from six different languages in the OGI multi-language telephone speech corpus. By simulating a phone recognizer that classifies phones into ten broad classes, a simple n-gram model gives low LID equal error rates (EER) of <1% on 30 seconds of test data. Language models based on these accurate phone transcripts can reveal insights into the phonology of different languages.

1. Introduction

Automatic Language Identification (LID) of speech has been a topic of research since the 1970s [1]. Although Language Identification can also refer to text classification, most research has been done on speech. In particular, the NIST Language Recognition Evaluation, currently held about once every two years, focuses on telephone speech. There has been a renewed interest in LID recently. At the NIST meeting in 2005 the number of participating sites had doubled [2] and this trend appeared to continue in 2007.

One type of system that has consistently performed well through all the evaluations is the phonotactic approach to LID [1, 4, 5]. This technique splits the problem into two stages; phone recognition, and then language modelling of the tokenized phones. This is known as PRLM (Phone Recognition followed by Language Modelling). Work by Matějka et al. [3, 4] demonstrates that improving the accuracy of the phone recognizers significantly increases the overall accuracy of the LID system. In the study reported in this paper a highly accurate phone recognizer is simulated using hand annotated phone transcripts (Figure 1). This isolates the language model and gives an upper bound to the performance expected on a complete LID system.

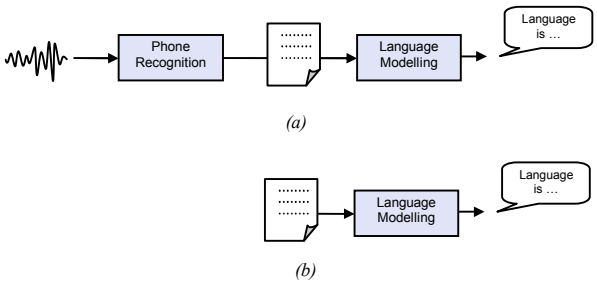

Figure 1: *a) PRLM, b) This study: A perfect phone recognizer is simulated with phone transcripts*

There have only been a few studies on the analysis of phone transcripts for LID. Interestingly, the most relevant paper on the subject was one of the earliest LID studies; House and Neuburg [6] hampered by the lack of accurate phone recognizers used phone transcripts instead. These transcripts were relatively short and many were derived from written text. Since then, a portion of the OGI Multi-language Telephone Speech Corpus [7] has been transcribed phonetically, giving a much richer source of transcripts. These transcripts have primarily been used to train phone recognizers. At OGI itself, there was interest in the statistics of the corpus and Muthusamy [8] looked at that distribution of broad phone classes – these were automatically segmented because the transcription effort was still taking place. As more transcripts became available, Berkling [9] was able to look at clustered phones and sequences of phones, picking out discriminative features for spoken LID between English and German. The corpus now contains over 600 verbatim phonetic transcripts.

The study reported in this paper investigated LID purely on the OGI phonetic transcripts themselves. This corpus was used because it remains one of the most detailed phonetic transcriptions of multilingual spontaneous speech available. Using OGI also allows a loose comparison with previous NIST LID evaluations because the OGI corpus was used in the early workshops. The OGI transcripts provide a valuable resource for getting closer to the true phonology of a language. Unlike language models based on inaccurate phone recognizers, language models based on accurate transcripts provide a reliable representation of the phone patterns. Inspecting these language models can potentially give new insights into the differences between languages. This paper is therefore intended as an update to the work of House and Neuburg [6]. The main differences are that more data is used and n-grams rather than HMMs are used for language modelling.

2. Method

The OGI Multi-language Telephone Speech Corpus [10] includes phonetic transcriptions for six languages: English (208), German (101), Hindi (68), Japanese (64), Mandarin (70), and Spanish (108) with the number in parentheses giving the number of transcripts available. Each transcript is taken from the 'story' section of the recording - a spontaneous monologue just under a minute long. The last 20 transcripts of each language were used for testing and the rest were used for training. This number was chosen because the original OGI test set for audio files numbered 20 for each language.

Table 1. *The different phone sets*

'CV3'	'SO3'	'House5'	'Expanded10'
Vowel	Sonorant	Vowel	Open vowel
			Close vowel
Consonant		Sonorant Consonant	Approximant
			Nasal
	Obstruent	Fricative	Voiced Fricative
			Voiceless Fricative
		Plosive	Voiced Plosive
			Voiceless Plosive
			Closure
Silence	Silence	Silence	Silence

The phones are clustered into language-independent broad phone classes. Language-dependant phones were not used despite their success in current LID systems [3] because phone mappings from one language to another for the common language models are not well defined for transcript LID. With language-independent phones it is generally more realistic to simulate an accurate broad phone recognizer with a few classes than an accurate, more finely grained phone recognizer. Using a fewer number of broad classes also reduces the data sparsity problem of longer n-grams.

Four different sets of broad phone classes were investigated. These are shown in Table 1 and reflect different phonetic resolutions. CV3 refers to a consonant-vowel three-class set (with silence being the third class). Similarly, SO3 refers to a sonorant-obstruent three-class set. House5 refers to the same five-class set used by House and Neuburg [6] with Expanded10 adding slightly more detail. In this study diphthongs were interpreted as a sequence of two vowels. The table is loosely based on the sonority hierarchy with the most sonorant classes at the top. Decisions on the different sets and how to cluster the phones were based on phonetic studies [11, 12], transcription guidance [13, 14] and previous experiments by other researchers [6, 16].

The language model used here is the same one that underlies many current PRLM systems. It is a simple n-gram language model using Katz back-off with Good-Turing discounting. N-grams from unigrams to 5-grams were investigated.

A language model was trained for each of the six languages. When testing on a transcript each model produced a log-likelihood score. A final score for each language was then calculated from a ratio of the best and second best log-likelihood scores. A simple ratio was used at this stage to remain robust for the small amount of training data.

3. Results

3.1. Core results

Experiments were conducted to investigate the effect of varying two different factors; phonetic resolution, and the size of the n-gram. As in the NIST evaluations, performance was tested on different lengths of test data; 3 seconds, 10 seconds, and 30 seconds. These different time lengths could be specified because the OGI transcripts were time-aligned for each phone. The 3s utterances were contained within the 10s utterances which were in turn contained within the 30s utterances. This resulted in 20 test transcripts for each duration. Results are shown in Tables 2 and 3. Table 2 shows the results for the three-class phone sets and Table 3 shows the results for the more detailed sets. An EER (Equal Error Rate) is shown for each experiment.

Table 2. *EER percentages for the 3 class sets*

	CV3			SO3		
	3s	**10s**	**30s**	**3s**	**10s**	**30s**
Unigram	45	43	36	43	35	29
Bigram	36	26	14	33	25	17
Trigram	31	20	10	33	23	13
4-gram	28	19	8	33	23	13
5-gram	28	18	8	33	23	16

Table 3. *EER percentages for the 5 and 10 class sets*

	House5			Expanded10		
	3s	**10s**	**30s**	**3s**	**10s**	**30s**
Unigram	35	23	15	23	17	16
Bigram	24	10	3	8	1	0
Trigram	20	5	1	7	1	0
4-gram	18	4	2	7	0	0
5-gram	20	7	4	6	0	0

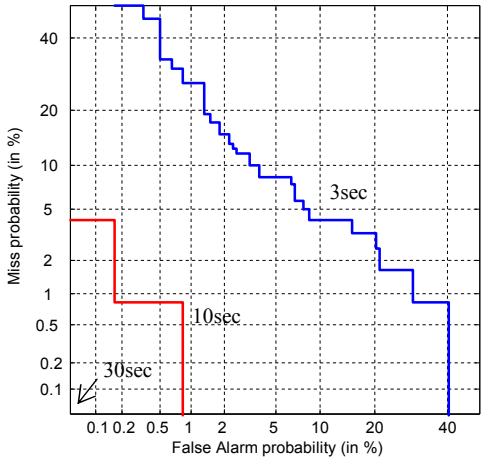

Figure 2: *DET plot for the Expanded-10 phone set using a trigram language model showing different lengths of test data*

There are some general trends for both factors which are not unexpected. As the phonetic resolution increases, LID accuracy also increases. As the n-gram size increases, LID accuracy generally increases up to trigrams but after that there is not much improvement. An example that works well is the ten-class set with a trigram model. A DET plot of this configuration is shown in Figure 2. There is no visible plot for 30s because there is 0% error on the test dataset of 120 test files. A 0% error should be interpreted as < 1% error because of the relatively small number of test files. The three timings and their error rates are helpful in giving a full picture of the accuracy.

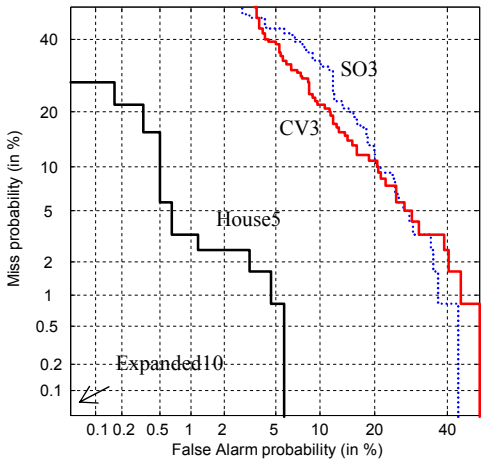

Figure 3: *DET plot showing different phonetic resolutions for the 30s bigram task*

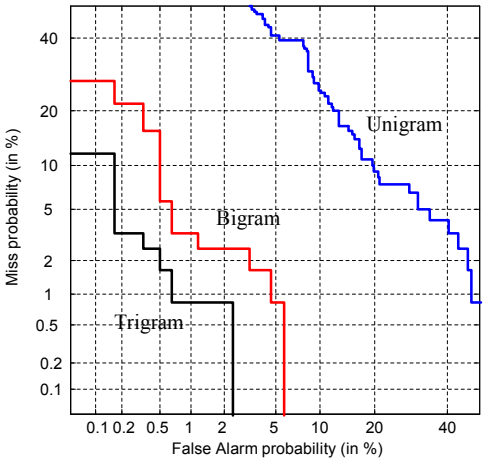

Figure 4: *DET plot of the House5 system showing unigrams, bigrams and trigrams on the 30s task*

Even with low phonetic resolution, the scores show a respectable accuracy. The consonant-vowel set is not as accurate as the sonorant-obstruent set for unigrams, but overtakes it as the n-gram increases in size. The five-class set is consistently better than the three-class sets. There is a similar improvement in moving to the ten-class set. There appears to be one anomaly in the unigram scores for 30s where the ten-class EER is higher than the five-class EER. This is due to an unusual kink in the DET curve at the EER point. Overall the ten-class set does perform better. This is confirmed by comparing the original six-way test errors (not shown on the table) which showed the ten-class set was indeed performing better for 30s overall. A comparison of the different phone resolutions are shown in Figure 3.

As the size of n-grams increase, the accuracy also increases for all phonetic resolutions up to trigrams. A typical plot of the consistent improvement up to trigrams is shown in Figure 4 for the five-class set. After this point there is no improvement. For the consonant-vowel set and the ten-class set there is still some slight improvement up to 5-gram level.

3.2. Comparisons with other studies

Caution should be exercised when comparing these results with other studies because the test conditions vary slightly. However, the comparisons can still give a rough idea of how other LID systems compare with the upper bound performance of a broad phone recognizer.

The results in this study for LID on phone transcripts compare favourably with recent published results on spoken LID [4]. It is not possible to make an exact like-for-like comparison with the recent results because the evaluation data has not been transcribed manually. The only evaluation data to be transcribed phonetically is OGI, so comparisons with earlier evaluations are needed. The most recent evaluation on OGI data was the 1995 NIST evaluation on nine different languages [1]. The best system gave an approximate error of 23% on the 10 second nine-way test. The best bigram system in this paper gives an approximate error of 1% on a 10 second six-way test. In a pairwise comparison with English the best system showed a 4% error compared to <0.5% error in this study. Adding the three extra languages would increase the error rate slightly but extrapolating from the effect of gradually adding new languages indicates these error rates still compare favourably.

At the time of the 1993 NIST evaluation there were a few LID systems that were explicitly reported as using broad phone class recognition. However these struggled to get much lower than a 50% error rate on the 1993 10s ten-way task. Muthusamy's system [8] was one of these and he also reports a 33% error on a 10s four-way task. Both the ten-class and five-class systems investigated here score much better than this. Since the data was the same, this demonstrates it is possible to perform effective LID with the phone patterns alone using simple language models. It appears that the main limiting factor for these other LID systems is the accuracy of the phone recognition.

Comparisons with the House and Neuburg study are more difficult. They suffered from a lack of data and, for some of the tests, training data was also used for testing. The most suitable experiment for comparison is LID of five American Indian languages because training and test data were kept separate. If the experiment was similar to the others in the House and Neuburg study it is likely to be based on read speech. Test files were equivalent to about 30 seconds to 1 minute of speech; however, there was only one test file per language. Training was also based on an equally small amount of data. An HMM language model was used in their study. The result they give on a five-way test was 0% error on five test files. The closest comparison in this current study is a 30s six-way test where the five-class trigram model gives 3% error on 120 test files. These could be viewed as similar results. One additional way of comparing the two studies is to look at test-set perplexity scores. For the House and Neuburg HMM the average test perplexity for this five-symbol alphabet was 3.1. For the five-class trigram model the average test perplexity is slightly worse at 3.4. These results are not conclusive because different tests conditions could affect the figures, e.g. amount of training data, whether the speech is spontaneous or read, and the difficulty in comparing perplexities across different languages. However it is not easy to dismiss the HMM language model for this broad phone class problem. A further study using an HMM language model on the OGI transcripts would clarify the difference in performance.

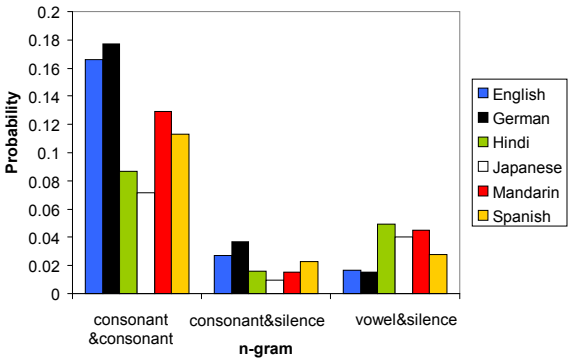

Figure 5: *CV3 discriminative n-grams*

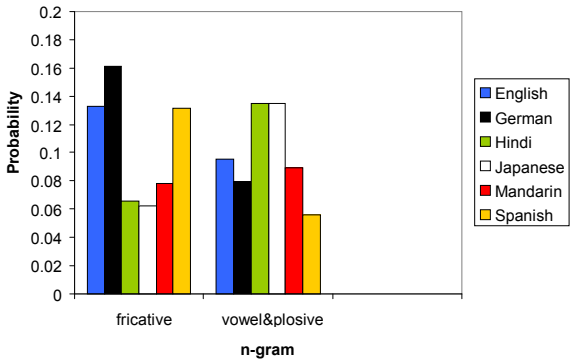

Figure 6: *House5 discriminative n-grams*

3.3. Inspecting the language models

In some previous experiments on broad phone classes, language models and corpus statistics have been inspected to seek insights on the differences between languages [6, 8, 9]. This study has the advantage of working from a large number of accurate transcripts so that the patterns found do give a picture of the genuine differences between languages. On inspecting the unigrams and bigrams in this experiment, a number of them where found to show particularly good discrimination. The most prominent of these are highlighted below.

Bigrams are only highlighted if they exhibit an effect independent of the unigram distributions. For example, although the distribution of consonants and vowels over the six languages are very similar for unigrams, there is much more variation among the bigrams. Some examples of the CV3 bigrams that appear to provide good discrimination between languages are shown in Figure 5. English and German have a high proportion of double consonants with Japanese having the least. This observation fits in well with the phonology of these languages; it is well known that English and German often have multiple consonant clusters whereas Japanese usually does not. It can also be seen that English and German are more likely to finish with a consonant before a silence (a pause or breath) whereas the other languages are more likely to finish with a vowel.

Sonorants and obstruents (not shown) have similar distributions at the unigram level. There are slight differences in the ratio of sonorants to obstruents with Mandarin having the most number of sonorants and German having the least. At the bigram level Mandarin is the only language that sticks out. The biggest difference is that Mandarin very rarely has two obstruents together when compared to the other languages.

The five-class phones show a large variation on the proportion of fricatives, with the European languages having about twice as many fricatives than the others in the group. This can be seen in Figure 6. It can also be seen that the vowel-plosive bigram is useful for discrimination.

Figure 7 shows the ten-class phones. It can be seen that there are some voiced fricatives in the European languages, some in Hindi but apparently none in Japanese or Mandarin. Nasals also show good discrimination. The bigram of double close-vowels reveals that these don't often occur in German and Hindi when compared to the other languages.

These observations are consistent with the OGI transcripts and the broad phone classes, but they do raise some important

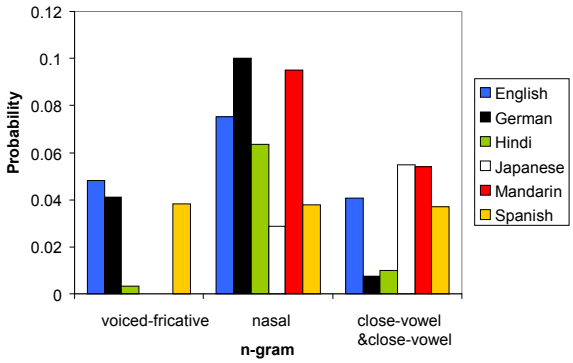

Figure 7: *Expanded10 discriminative n-grams*

questions about the phone classes. For example, although Japanese has an apparent voiced fricative phoneme /z/ all the realizations in the OGI corpus are affricates and are therefore grouped with the plosives (this is consistent with the groupings in a number of phonetic studies e.g. [11, 12]). Another issue is the voicing of plosives. Some realizations of the Mandarin phoneme /p/ can be acoustically very close to some realizations of the English phoneme /b/. It may be unrealistic to expect a phone recognizer to discriminate between the two. A few changes may be needed to the ten-class set if it is to realistically simulate an accurate phone recognizer. Even if an accurate phone recognizer can only produce up to five classes it has been shown that some interesting observations about the phonology of the language can be made.

4. Conclusions and further work

This paper has shown that high accuracy LID can be obtained on phonetic transcripts when using broad phone classes. Increasing the phonetic resolution of the broad classes increases the LID accuracy. Increasing the size of the n-grams up to tri-grams also leads to a greater LID accuracy. As expected these results compare favourably with published figures on spoken LID accuracy. The results can be viewed as defining an upper bound on the accuracy of a LID system that uses a broad class phone recognizer and an n-gram language model. The experiment here can be used as a baseline which other language models can be compared against.

Inspecting the language models shows each phone resolution exhibits n-grams that provide good discrimination

across languages. The n-grams often reflect known phonological rules, and have the potential to discover new rules or patterns. Some of the patterns uncovered are not immediately apparent in the finer phone classes, and are better represented in the broad phone classes e.g. the simple restriction on double obstruents in Mandarin is not obvious from the ten-class phone n-grams. This indicates that the levels of phone resolution may be complementary for LID.

Further work then, could include combining the different phone resolutions to see if they perform better together. Since some of the discriminative n-grams may be relatively easy to detect acoustically, such as estimating the proportion of voiced fricatives, these could also be tested on *spoken* LID systems. Combining these multiple streams of features is reminiscent of some previous work on LID with articulatory features [16], and it would be interesting to use articulatory features for transcript LID.

Further work should also include trying an HMM language model on this test because the comparison with the House and Neuburg study indicated that it may compete well on this broad phone class problem.

5. Acknowledgements

This work is funded by the UK Engineering and Physical Sciences Research Council (EPSRC).

6. References

[1] Zissman, M. and Berkling, K., "Automatic language identification' Speech Communication" 35(1), 2001, pp115-124

[2] Martin, A.F. and Le, A.N., "The Current State of Language Recognition: NIST 2005 Evaluation Results", Proc. IEEE Odyssey, 2006

[3] Matějka, P., Schwarz, P., Černocký, J. and Chytil, P., "Phonotactic Language Identification using High Quality Phoneme Recognition", Proc. Eurospeech, 2005, pp2237-2240

[4] Matějka, P., Schwarz, P., Burget, L. and Černocký, J., "Use of Anti-models to further improve state-of-the-art PRLM Language Recognition Systems", Proc. ICASSP, 1, 2006, pp197-200

[5] Martin, A. and Przybocki, M. "NIST 2003 language recognition evaluation", Proc. Eurospeech , 2003

[6] House, A. and Neuburg, E., "Toward automatic identification of the language of an utterance. I. Preliminary methodological considerations", J. Acoust. Soc. Amer., Vol. 62, 1977, p708.

[7] Muthusamy, Y.K., Cole, R. and Oshika, B., "The OGI multi-language telephone speech corpus", Proc. ICSLP, 1992, pp895-898

[8] Muthusamy, Y.K., "A Segmental Approach to Automatic Language Identification", Ph.D. Thesis, Oregon Graduate Institute of Science and Technology, 1993

[9] Berkling, K.M., "Automatic language identification with sequences of language independent phoneme clusters", Ph.D. Thesis, Oregon Graduate Institute of Science and Technology, 1996

[10] "OGI Multi-language Telephone Speech Corpus v1.2", http://www.cslu.ogi.edu/corpora/mlts/

[11] Ladefoged, P. and Maddieson, I., "The Sounds of the World's Languages", Blackwell Publishers, 1996

[12] Maddieson, I., "Patterns of Sound", Cambridge University Press, 1984

[13] Esling, J., "Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet", Cambridge University Press, 1999

[14] Hieronymus, J., "ASCII Phonetic Symbols for the World's Languages: Worldbet", Journal of the International Phonetic Association, 23, 1993

[15] Zissman, M. and Singer, E., "Automatic language identification of telephone speech messages using phoneme recognition and N-gram modelling", Proc. ICASSP, I, 1994, pp305-308

[16] Parandekar, S. and Kirchhoff, K., "Multi-stream language identification using data-driven dependency selection", Proc. ICASSP, 1, 2003, pp28-31