Phonotactic spoken language identification with limited training data

Marius Peche, Marelie Davel and Etienne Barnard

Human Language Technologies Research Group, Meraka Institute, Pretoria, South Africa

mpeche,mdavel,ebarnard@csir.co.za

Abstract

We investigate the addition of a new language, for which limited resources are available, to a phonotactic language identification system. Two classes of approaches are studied: in the first class, only existing phonetic recognizers are employed, whereas an additional phonetic recognizer in the new language is created for the second class. It is found that the number of acoustic recognizers employed plays a crucial role in determining the recognition accuracy for the new language. We study different approaches to incorporating a language for which audio-only data is available (no pronunciation dictionaries or transcriptions) and find that if more than about 2 000 training utterances are available, a bootstrapped acoustic model for the new language can improve accuracy substantially.

Index Terms: spoken language identification, generalization, resource scarce languages

1. Introduction

Spoken language identification (LID) has matured significantly in the past decade; on comparable data, current systems are an order of magnitude more accurate than state-of-the-art systems around 1996 [1, 2]. The best current systems combine acoustic [3] and phonotactic [2] information sources, and achieve equal error rates of less than 10% on the ten-second segments contained in the NIST 2003 benchmarks.

The component that contributes most to the accuracy of these systems employs phonotactic features in the form of bigram statistics. In the most successful configuration – known as "Parallel Phone Recognition" (PPR) [4] – phone recognizers in a number of languages are used to convert an input utterance into a set of phoneme strings (one for each recognizer). A classifier such as a naive Bayesian classifier or a support vector machine (SVM) is then used to determine the language spoken, based on the statistics of the bigrams that occur in these phoneme strings.

The goal of the current research is to investigate how a phonotactic language identification system can be applied to environments where limited training data is available. Such environments will be increasingly important as speech-processing systems are extended beyond the approximately twenty major languages that are currently the focus of most research efforts. For the majority of the approximately seven thousand living languages [5] limited resources (such as lexicons, pronunciation dictionaries and transcribed speech corpora) are available [6], and it is both practically and theoretically interesting to understand how speech-processing systems can be applied when limited resources are available.

In Section 2 we describe a number of methods that can be used to apply PPR language identification when limited resources are available, and highlight the issues that need to be understood in each of these approaches. Section 3 describes the experiments that we have undertaken to address these issues, and Section 4 contains the results of these experiments. Conclusions from the experiments and future work are discussed in Section 5.

2. Approaches

A basic PPR system, designed to recognize N languages, contains two types of models: (1) a set of M acoustic models that function as phone recognizers, and (2) a set of language models or classifiers that characterize the observed phonotactics of each language when recognized in terms of each of the phone recognizers.

The resource requirements of these two model classes are significantly different – whereas the language models or classifiers can be trained with untranscribed data from any language, the acoustic models typically require pronunciation dictionaries as well as transcribed speech. In addition, the amount of speech needed for the training of acoustic models is typically substantially larger, in order to obtain speaker-independent phone recognition of sufficient accuracy. Fortunately, there is no requirement that N and M be equal. That is, a language can be recognized even if no phone recognizer for that language is available [4], since the language models or classifiers can be constructed based on the phone strings produced by the existing acoustic models. The first approach to adding a new language that we investigate therefore uses no new acoustic model (in the language being added).

However, intuition as well as some experimental results [7] indicate that the addition of acoustic models relevant to the target language should improve the accuracy of the LID system. Although there is strong evidence that this improvement is positively correlated with the accuracy of the additional phone recognizer [8], it is conceivable that even a recognizer with low accuracy would be of some additional value, and initial results confirming this expectation have been reported [9]. To investigate this possibility in the current context, we have investigated two approaches to the construction of acoustic models when limited resources are available.

In the first approach, we assume that orthographic transcriptions of a limited number of utterances in the new target language are available. No other resources are assumed present. An orthography-based model, with one three-state HMM model per letter, is therefore constructed from these utterances, as in [10].

Our other approach does not assume the availability of orthographic transcriptions; it is bootstrapped from the acoustic models of an existing phoneme recognizer. That is, all of the training utterances are recognized with the existing recognizer (using Viterbi alignment); these recognition results are employed as "transcriptions" for training a new set of acoustic models, using embedded re-estimation, as in [9]. (In princi-

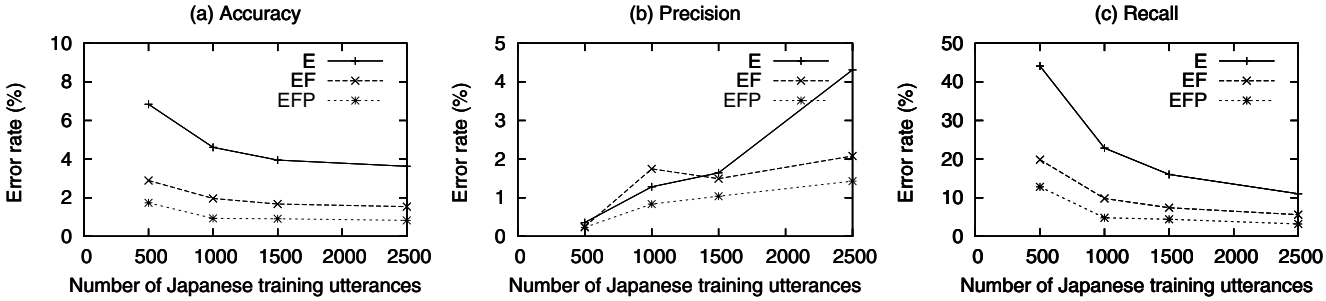

Figure 1: *Error rates when no Japanese acoustic models are constructed. An increasing amount of Japanese training data is used to train the language classifier of an English-only (E), an English-French (EF), and an English-French-Portuguese PPR system.*

ple, this process can be repeated with transcriptions repeatedly derived from the updated acoustic models, but we have only experimented with one cycle of re-estimation.)

The phonetic recognizer created in this fashion is used in the same way as the existing recognizers – that is, for each target language, bigram statistics produced by the recognizer are computed and used for language classification (in conjunction with the statistics produced by each of the "conventional" phone recognizers).

3. Experimental design

3.1. Corpora

Because of their role as world languages that are widely spoken in Africa, our initial LID system was designed to distinguish between English, French and Portuguese. We therefore trained phone recognizers and language classifiers for these languages, using the GlobalPhone corpus [11] for both training and test data. The amount of training data employed in each language is summarized in Table 1. To assess the addition of a new language, we chose to use a language that is linguistically dissimilar from these languages, but with data recorded under acoustically similar circumstances. We therefore selected Japanese, which is also contained in the GlobalPhone suite of corpora.

Table 1: *Amount of data used to train the different phone recognizers.*

Language	# utterances	# hours	# speakers
English	10,219	20.0	83
French	8,380	21.6	80
Portuguese	6,037	14.4	77

3.2. Language classifier

A number of approaches to language classification from phone strings have been proposed [4]. The most recent published results [2] as well as our own experiments indicate that a support vector machine (SVM) using bigram frequencies as input functions optimally in this regard. Thus, the features that are calculated for every utterance are the counts of each of the bigrams, normalized by the total number of bigrams recognized. These features are computed for all phonetic recognizers, and concatenated into a single vector; the total number of features is therefore

$$D = \sum_{i=1}^{M} P_i * (P_i - 1), \qquad (1)$$

where P_i is the number of phones (including the silence phone) in the recognizer for language i, and M is the number of phoneme recognizers (as above). Since P_i is around 45 (on average), there are approximately 2 000 features per recognizer – hence the importance of a classifier such as the SVM which generalizes well in high-dimensional feature spaces.

When using an SVM, the most important design choices are related to (a) the shape and width of the kernel function employed, and (b) the value of the margin-accuracy trade-off parameter. After some experimentation, we decided to employ a Gaussian kernel. Classification accuracy was found to be fairly insensitive to the width and trade-off parameters; reasonable values for these were thus chosen prior to formal experimentation, and these values were used throughout.

4. Results

4.1. Experiment 1: Using existing acoustic models

In the first experiment, only English, French and Portuguese acoustic models are used during language identification; no Japanese acoustic models are constructed. During training, Japanese training data is recognized by the other acoustic models, and the resulting phone strings are used to train the language classifier.

In Fig. 1 we show the error rates that result from using different combinations of the English, French and Portuguese recognizers, when classifying a data set that contains these three languages as well as Japanese. All error rates are shown as a function of the number of Japanese training utterances used. (In each of the other three languages, the same training set of between 6,000 and 10,000 utterances per languages was used throughout.) Fig. 1 (a) shows the overall error for all four languages, whereas Figs. 1 (b) and (c), respectively, show the error rate in precision and recall for the added language (Japanese).

From both the accuracy and the recall measures, it can be seen that the addition of more acoustic models does indeed reduce the error rates of the classifier for all training-set sizes considered. The precision results are less informative in our experimental paradigm: because of the predominance of the training data in the three "well-resourced" languages, the error in precision is significantly lower than the recall error rate, and increases as additional Japanese training samples are added to the training set (since precision errors weigh more heavily then).

The confusion matrix in Table 2 shows the performance that can be achieved when 2 500 untranscribed utterances from a new language are added, using acoustic models from three languages. The resulting precision and recall are not much worse than those of the languages for which acoustic models are avail-

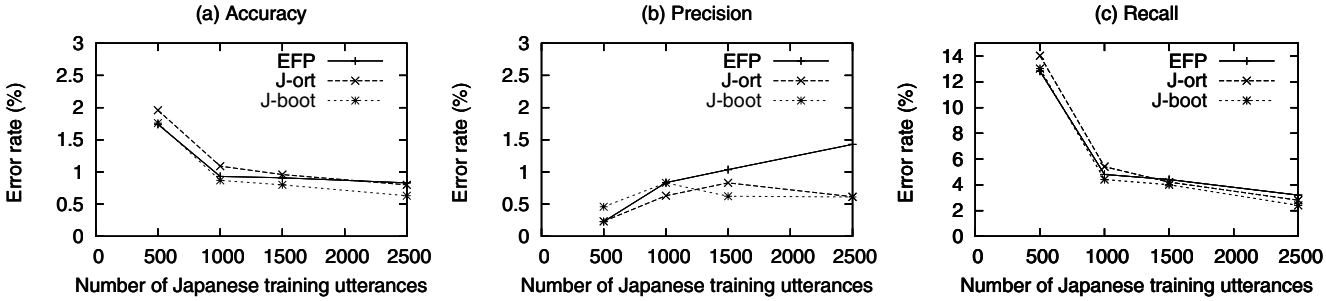

Figure 2: *Error rates when constructing a new Japanese acoustic model: either orthography-based (J-ort) or bootstrapped (J-boot) using an increasing amount of Japanese training data. Results are compared with the English-French-Portuguese PPR system (EFP) of Fig. 1.*

able: the recall accuracy for Japanese is about the same as that for Portuguese, and the error in precision is about 60% higher.

Table 2: *Confusion matrix when only English, French and Portuguese acoustic models are used. Columns correspond to the true class of an utterance, and rows represent the classification result*

	English	French	Portuguese	Japanese
English	**1701**	0	9	0
French	0	**1396**	3	4
Portuguese	2	1	**984**	12
Japanese	0	0	7	**483**

4.2. Experiment 2: Adding an acoustic model

In the second experiment, new acoustic models are added according to the two approaches described in Section 2, namely (1) using an orthography-based model and (2) bootstrapping transcriptions utilizing the English recognizer. Fig. 2 and Table 3 summarize the results obtained when comparing these two approaches with the best results obtained in Experiment 1 above. The addition of the orthography-based Japanese recognizer does not seem to be of much value, except for the case where 2 500 training utterances are used – there, the addition of the acoustic model reduces the error in recall from 3.2% to 2.8% and reduces the overall error rate from 0.83% to 0.80%.

The benefits of the bootstrapped acoustic model are more substantial for the amount of training data considered: even for 1 500 training utterances, the relative reduction of the error in recall and overall error rates are more than 10 %, and for 2 500 training utterances these relative reductions in error rate are around 20 %. Table 4 contains the resulting confusion matrix. Now, both the precision and the recall of the Japanese utterances are statistically indistinguishable from those of the Portuguese utterances. (Note that we have used the same utterances for estimating the acoustic and language models. With a small amount of training data this is the only available option, but it may be preferable to split the training data and use different sets for these two tasks when sufficiently many utterances are available.)

5. Discussion

We have shown that languages with limited available resources can be added to a phonotactic language-identification system containing more resource-demanding components with almost

Table 3: *Detail of error rates (in %) when constructing a new Japanese acoustic model: either orthography-based (J-ort) or bootstrapped (J-boot) using 2,500 Japanese training utterances. Results are compared with the English-French-Portuguese PPR system (EFP) of Fig. 1*

System	Accuracy	Precision	Recall
EFP	0.83	1.43	3.21
J-ort	0.80	0.62	2.81
J-boot	0.63	0.61	2.41

Table 4: *Confusion matrix when a Japanese acoustic model is bootstrapped using the English recognizer.*

	English	French	Portuguese	Japanese
English	**1702**	0	9	0
French	0	**1396**	3	3
Portuguese	1	1	**988**	9
Japanese	0	0	3	**487**

no loss in accuracy to the latter languages, and good – though not quite equal – accuracy for the former. In achieving this performance, the number and diversity of available phonetic recognizers is an important requirement. Through use of a bootstrapping procedure, acoustic models can be created for a language for which neither transcriptions nor pronunciation dictionaries are available; such acoustic models have a significantly beneficial effect on overall accuracy. This result confirms the findings in [9], provides some additional insight on the relationship between the amount of training data and the benefits of various algorithmic components, and suggests that blind re-estimation is preferable to the use of orthography-based unit recognizers. In light of the improvements obtained, it is worthwhile to investigate alternatives and extensions to the methods employed - for example, MAP adaptation of acoustic models rather than embedded re-estimation and multiple iterations of estimating the transcriptions for untranscribed languages.

It would also be interesting to extend this work to systems that include an acoustic scoring component. In light of the superior performance achievable with phonotactic models [2], we expect that the improvements reported here will remain the most important contribution to overall accuracy. It will also be worthwhile to repeat these comparisons on other corpora, such as those used for the NIST evaluations. Our long-term aim is to apply these methods to resource-scarce languages, including the majority of languages spoken in Africa.

6. References

[1] Alvin F. Martin and Mark A. Przybocki, "NIST 2003 language recognition evaluation," in *Eurospeech*, 2003, pp. 1341–1344.

[2] Haizhou Li, Bin Ma, and Chin-Hui Lee, "A vector space modeling approach to spoken language identification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, pp. 271–284, January 2007.

[3] William M. Campbell, Joseph P. Campbell, Douglas A. Reynolds, E. Singer, and P. A. Torres-Carrasquillo, "Support vector machines for speaker and language recognition.," *Computer Speech and Language*, vol. 20, no. 2-3, pp. 210–229, 2006.

[4] M. A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Transactions on Audio, Speech and Language Processing SAP-4(1)*, pp. 31–44, January 1996.

[5] SIL International, "Ethnologue, languages of the world," http://www.ethnologue.org/.

[6] Ksenia Shalonova and Roger Tucker, "Issues in porting TTS to minority languages," in *LREC, SALTMIL workshop on Minority Languages*, Lisbon, 2004.

[7] E. Barnard and Y. Yan, "Toward new language adaptation for language identification," *Speech Communication*, vol. 21, pp. 245–254, 1997.

[8] Pavel Matejka, Petr Schwarz, Jan Cernock, and Pavel Chytil, "Phonotactic language identification using high quality phoneme recognition," in *Interspeech*, 2005, pp. 2237–2240.

[9] A. Montero-Asenjo, D.T. Toledano, J. Gonzalez-Dominguez, J. Gonzalez-Rodriguez, and J. Ortega-Garcia, "Exploring PPRLM performance for NIST 2005 language recognition evaluation," in *IEEE Odyssey 2006: The Speaker and Language Recognition Workshop*, June 2006, pp. 1–6.

[10] C. Schillo, G.A. Fink, and F. Kummert, "Grapheme based recognition for large vocabularies," in *ICSLP*, Beijing, China, October 2000, pp. 129–132.

[11] Tanja Schultz and Alex Waibel, "Fast bootstrapping of LVCSR systems with multilingual phoneme sets," in *Eurospeech*, Rhodes, Greece, September 1997, vol. 1, pp. 371–373.