Language Identification Using Minimum Linguistic Information

Diamantino Caseiro[1], Isabel Trancoso[2]

INESC/IST
INESC, Rua. Alves Redol, nº9,
1000 Lisbon, PORTUGAL
Phone: +351.1.3100268
Fax: +351.1.3145843
[1] *email: dcaseiro@speech.inesc.pt*
[2] *email: imt@inesc.pt*

Abstract: *Automatic spoken language identification is the problem of identifying the language being spoken from a sample of speech by an unknown speaker. Current language identification systems vary in their complexity. The systems that use higher level information have the best performance. Nevertheless, that information is hard to collect for each new language. In this work, we present a state of the art language identification system, which uses very little linguistic information, and so easily extendable to new languages. In fact, the presented system needs only one language specific phone recogniser (in our case the Portuguese one), and is trained with speech from each of the other languages.*
We studied the problem of language identification in the context of the European languages (including, for the first time, European Portuguese), which allowed us to study the effect of language proximity in Indo-European languages. The results reveal a significant impact on the identification of some languages. With the SpeechDat-M corpus, with 6 European languages (English, French, German, Italian, Portuguese and Spanish) our system achieved an identification rate of about 80% on 5-second utterances.

Keywords*: Automatic spoken language identification, Continuous speech recognition, Multilingual speech processing, Spoken language engineering.*

1. INTRODUCTION

When designing a language identification system we face the problem of scalability. To cover a significant amount of the thousands of languages commonly spoken, one is confronted with two problems: on one hand the problem of devising a system efficient enough to be able to identify the language in a short amount of time; on the other, the problem of collecting the information needed to train such a system. The collection of speech data is, in itself, a hard enough problem. Techniques that require hand labelling of the speech material and other linguistic data are very hard to extend beyond the most common languages.

In this paper, we present a system with state of the art performance that uses a minimum amount of linguistic information and requires only speech data to be extended to new languages. By contrast, the best systems reported in the literature make heavy use of linguistic data.

The best systems use multiple large vocabulary continuous speech recognisers [1][4]. These systems include a complete word recogniser for each language, and use word and sentence level language modelling. Due to the difficulty of adding a new language, those systems are generally limited to a very small set of languages. To build such a system, one requires a large amount of labelled speech to train phone recognisers, in particular if the system uses context dependent phone recognisers. In addition, large amounts of text are required to train language models of word-n-grams.

A particularly successful approach is parallel language dependent phone recognition followed by language modelling [8][7]. This type of approach exploits the phonotactic properties of the languages, and does not need to recognise words. The recognition and language modelling are done at the phone level. This approach is able to achieve identification rates in excess of 80%, using 10-second utterances in 6 languages [7]. The biggest drawback is the requirement of labelled speech for a large subset of the languages used. As it is based on multiple language-specific phone recognisers, it requires labelled speech to train those recognisers.

It is possible to obtain the same level of identification using only one set of language independent phone recognisers [3]. By performing multiple recognitions of the input utterance by the same models, and constraining each recognition by a different phone-bigram grammar (obtained from manually labelled transcriptions), Navrátil obtained multiple phone streams of the same utterance. Those streams were then fed to stream-specific language models. The likelihood of each language was determined by a weighted combination of the likelihoods of the languages in each stream. Nevertheless, this system still requires labelled speech in each language to model the language independent subword units. And requires textual data and pronunciation dictionaries to create the phone-bigram grammars.

We continue in Section 2 by describing the SPEECHDAT corpus, and the separation of training and test sets. In Section 3, we describe a baseline system, and, in Section 4, the proposed system. In Section 5, we show the results attained with the systems. Finally, concluding remarks and plans for future work are presented in Section 6.

2. TRAINING AND TEST CORPUS

The SPEECHDAT corpus [5] allowed us to investigate the problem of automatic identification of European languages, through the public telephone network. It includes utterances from about one thousand speakers from each of seven European countries. In English, French, German, Italian, Portuguese, Spanish, and Swiss French. In this work we used the first six languages.

We selected nine utterances from each speaker in each language from the set of phonetically rich sentences. The utterances in this set are read sentences with an average length of 5 seconds. They were randomly separated into training and evaluation sets. As illustrated in figure 1, we were forced to reject a significant amount of data to guarantee that there was no speaker or sentence overlap between the train and test set.

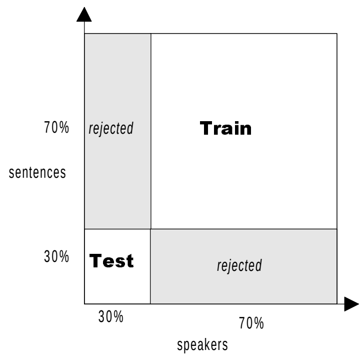

Figure 1. Illustration of rejected material.

Because a minimum amount of speech is required to perform identification we removed the utterances with less than 2 seconds from the test set. We also removed the utterances with more than 10 seconds from the test set to bring the evaluation set within the upper limit of 10 seconds commonly used to evaluate language identification systems.

3. BASELINE SYSTEM

Our baseline system used the classic language dependent phone recognition followed by language modelling (PRLM) architecture [8]. In this architecture, the sequence of phones decoded by the recogniser is matched against a set of phone-bigram language models, one for each language. The output of each model is the likelihood of the sequence in its language

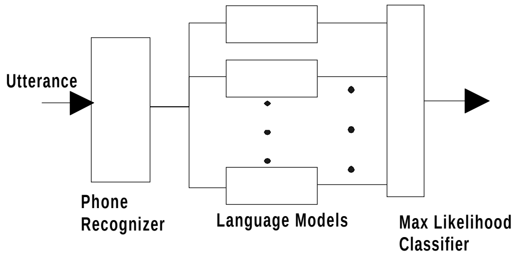

Figure 2. Baseline system architecture.

3.1. Parameter Extraction

From the train and test utterances we extracted vectors composed of 12 Mel-frequency cepstral coefficients, 12 delta-cepstral coefficients, energy and delta energy. In the cepstral analysis 10ms frames and 25ms Hamming windows were used. Mean cepstral removal was performed to reduce the channel effect.

3.2. Phone Recogniser

A continuous mixture HMM based phone recogniser decoded the incoming spoken utterance. This recogniser had male and female models of each of the 38 Portuguese phones and 2 non-speech units (silence and pause), totalling 78 different subword units. The models had the conventional three-state left-to-right architecture shown on figure 3. When tested with the Portuguese test set, the recogniser achieved 54.1% phone recognition correctness.

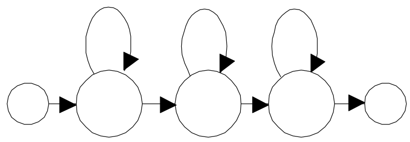

Figure 3. Three-state left-to-right HMM architecture.

3.3. Language Models

The decoded sequence was stripped of the sex information and was fed to a set of phone-bigram language models (see figure 2). The language models were based on interpolated phone-bigram probabilities. Let $A = \{a_1, a_2, \cdots, a_T\}$ be one sequence of phones. The likelihood for each language was computed as:

$$L(A \mid LM^l) = \frac{1}{T}\log(P(a_1 \mid LM^l)) + \sum_{t=2}^{T} \log\tilde{P}(a_t \mid a_{t-1}, LM^l)$$

where LM^l is the language model of language l and $\tilde{P}$ is the interpolated bigram model:

$$\tilde{P}(a_t \mid a_{t-1}) = \alpha P(a_t \mid a_{t-1}) + \beta P(a_t)$$

$$0 \le \alpha, \beta \le 1 \qquad \alpha + \beta = 1$$

α and β are empirical weights.

The identified language was selected using a maximum-likelihood classifier:

$$\hat{l} = \arg\max_{l} L(A \mid LM^{l})$$

3.4. Linguistic Information

The only linguistic information used was the one required to train the acoustic models: the orthographic transcription of the Portuguese utterances and a corresponding pronunciation dictionary.

4. PROPOSED SYSTEM

The best phonotactic language identification systems, like [7] and [3], combine in their architecture multiple simple modules similar to our baseline system. To achieve state of the art performance we decided to use an architecture similar to [3]. In this architecture the output of the phone recogniser consists of multiple phone sequences, each resulting of the Viterbi decoding constrained by a particular set of language-specific transition probabilities. To avoid Navrátil's requirement of original-label transcription, those probabilities were calculated by a bootstrapping process. First, the training data was decoded using a null-grammar constraining the utterances only to male or female phone models. The decoded phone sequences were then used to determine each language phone-bigram probabilities.

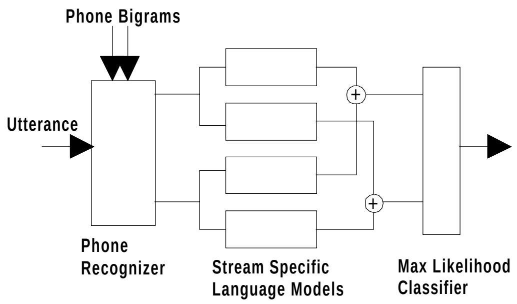

Figure 4. System architecture illustrated with two languages.

The language models used were similar to the ones used in the baseline system. But now, instead of n (number of languages) models, we have n^2. A set of n for each decoded sequence.

We determined each language score as the sum of its sequence specific likelihoods, and, like before, selected the identified language using a maximum-likelihood classifier.

5. RESULTS

The system was tested using a closed set of six European languages. Table 1 shows the identification rates of the baseline system. Portuguese is, naturally, the best identified language, because the acoustic models were trained only with Portuguese speech. French has the second best identification rate, because, the French portion of the corpus has a higher linguistic variety, and also because, English and German, and, Spanish and Italian, are very close languages.

Language	Result
EN	72,3%
ES	68,9%
DE	67,3%
PT	88,9%
IT	62,7%
FR	75,1%
Total	72,7%

Table 1. Baseline system results.

The results attained by the proposed system are shown on table 2.

	EN	ES	DE	PT	IT	FR
EN	**81.4%**	0.4%	*11.1%*	1.7%	2.7%	2.8%
ES	1.9%	**70.6%**	2.3%	3.7%	*15.9%*	5.6%
DE	*8.6%*	1.5%	**82.4%**	1.2%	2.0%	4.2%
PT	2.5%	3.1%	1.9%	**87.8%**	1.7%	3.1%
IT	4.9%	*14.4%*	4.1%	1.1%	**70.0%**	5.6%
FR	2.1%	2.9%	4.3%	1.2%	4.0%	**85.5%**
Total	79.6%					

Table 2. Proposed system identification confusion table.

The global results cannot be directly compared with published results using other databases because of the language choices and utterance lengths. Nevertheless, our results are close to those reported for 6-language tasks, with 10s utterances, using the NIST test set. This system shows a significant improvement over the baseline system. The error rate was reduced to less than 75% of the baseline system error.

The proximity between some languages can be clearly seen: Germanic languages were most often confused with each other (11.1% of the English utterances were mistaken as German, and 8.6% of German as English). As far as romance languages are concerned, we see that Spanish and Italian were also easily mistaken with each other (15.9% and 14.4%). The Portuguese and French results are harder to interpret. The confusion rates between these languages and the others were not significant enough for us to claim any particular proximity. In the Portuguese case, this might be due to the fact that the acoustic models were Portuguese and so we were able to identify it better. The French language is one of the romance languages with more Germanic influence [6], which might explain why it does not show any significant confusion trend.

Clearly, the duration of the utterances has a major impact on system performance. With the purpose of investigating this effect, we ranked all the utterances by duration and performed separate tests with the utterances in each 1-second interval.

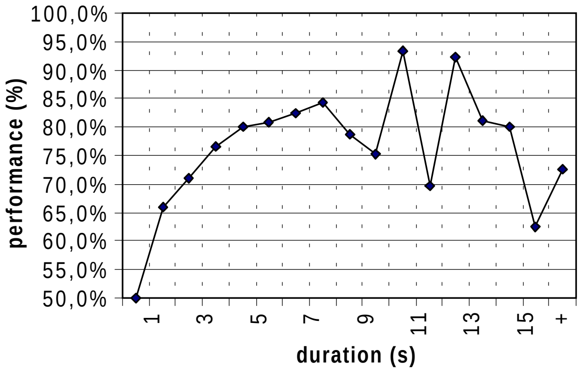

Figure 4. Evolution of the identification rate with utterance duration.

As expected, the identification rate increases with the duration of the utterances. From 8 seconds onwards the results were erratic due to the low number of utterances in each interval. The best significative result was 84.3% with utterances of 7 to 8 seconds duration. This result leads us to believe that the system may attain an identification rate in excess of 85% if tested with 10-second utterances.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we described a method for automatic language identification that uses a minimal amount of linguistic information, in order to extent this system to new languages, only speech data of these languages is required. Our results showed that linguistic proximity between languages can degrade the performance of automatic language identification systems. That proximity indicates that when speech data from more languages becomes available, hierarchical systems, which identify first the group of languages and them the language within the group, may be feasible.

The identification results can be improved by tuning the combination of the different language models: instead of simply adding each model contribution, we can weight each model with empirical weights, trained from a validation set. We plan to improve the system by using pseudo-words [2] to decide between the two best languages.

7. REFERENCES

[1] J. Hieronymous, S. Kadambe, “Spoken language identification using large vocabulary speech recognition.” Proc. ICSLP-96, Philadelphia, USA, 1996.

[2] M. Lund, H Gish, “Two novel language model estimation techniques for statistical language identification,” Proc. EUROSPEECH’95, Madrid, Spain, 1995.

[3] J. Navrátil, W. Zühlke, “Double-bigram decoding in phonotactic language identification,” Proc. ICASSP-97, Munich, Germany, 1997.

[4] T. Schultz, I. Rogina, A. Waibel, “LVCSR-based language identification.” Proc. ICASSP-96, Atlanta, USA, 1996.

[5] SPEECHDAT (M): EU-project LRE-63314.

[6] H. Walter, “L’Aventure des Langues en Occident”, Éditions Robert Laffont, 1994

[7] Y. Yan, E. Barnard, “An approach to automatic language identification based on language-dependent phone recognition,” Proc. ICASSP-95, Detroit, USA, 1995.

[8] M. Zissman, E. Singer “Automatic Language Identification of Telephone Speech Messages using Phoneme Recognition and N-Gram Modelling” Proc. ICASSP-94, Minneapolis, USA, 1994.