SPOKEN LANGUAGE IDENTIFICATION USING THE SPEECHDAT CORPUS

Diamantino Caseiro, Isabel Trancoso

INESC/IST

INESC, Rua Alves Redol, nº9, 1000 Lisbon, Portugal.

E-mail: {dcaseiro, Isabel.Trancoso}@inesc.pt Phone: +351 1 3100268, Fax: +351 1 3145843

ABSTRACT

Current language identification systems vary significantly in their complexity. The systems that use higher level linguistic information have the best performance. Nevertheless, that information is hard to collect for each new language. The system presented in this paper is easily extendable to new languages because it uses very little linguistic information. 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.

With the SpeechDat-M corpus, with 6 European languages (English, French, German, Italian, Portuguese and Spanish) our system achieved an identification rate of 83.4% on 5-second utterances, this result shows an improvement of 5% over our previous version, mainly through the use of a neural network classifier. Both the baseline and the full system were implemented in realtime.

1. INTRODUCTION

When designing a language identification system, we face the problem of scalability. In order 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 [2][8]. 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. In order 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 [10][11]. 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 [10]. 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 [4]. 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. In addition, it 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 full system with multiple decoders and a neural net classifier. In Section 5, we present the realtime implementation of the systems. In Section 6, we show the results attained with both systems. Finally, concluding remarks and plans for future work are presented in Section 7.

2. TRAINING AND TEST CORPUS

We used the SPEECHDAT corpus [9] to investigate the problem of automatic identification of European languages. This corpus was collected through the public telephone network and includes utterances from about one thousand speakers from

each of seven European countries. The collected languages are English, French, German, Italian, Portuguese, Spanish, and Swiss French. In this work, we used only the first six languages.

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

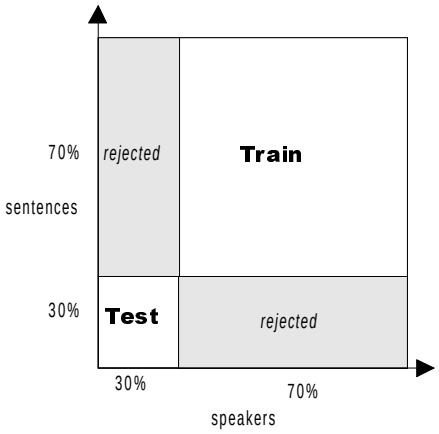

Figure 1: Illustration of rejected material.

3. BASELINE SYSTEM

Our baseline system used the classic language dependent phone recognition followed by language modelling (PRLM) architecture [11]. 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.

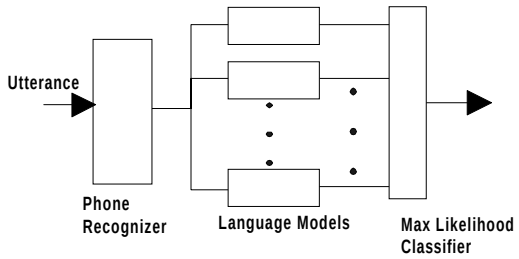

Figure 2: Baseline system architecture.

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.

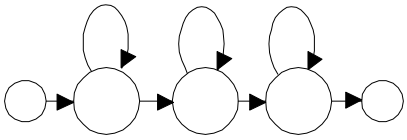

Figure 3: Three-state left-to-right HMM architecture.

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:

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

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

$$\tilde{P}(a_t \mid a_{t-1}) = \alpha P(a_t \mid a_{t-1}) + \beta P(a_t)$$
$$0 \le \alpha, \beta \le 1 \qquad \alpha + \beta = 1$$

where 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. FULL SYSTEM

The best phonotactic language identification systems reported in the literature, like [10] and [4], combine in their architecture multiple simple modules similar to our baseline system. In order to achieve state of the art performance, we decided to use an architecture similar to [4]. 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 transcriptions, 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.

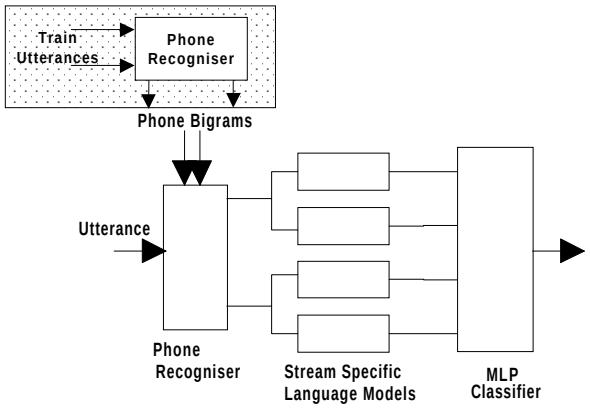

Figure 4: Full system architecture.

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

4.1. Neural Network Classifier

In [1] we determined each language score as the sum of its sequence specific likelihoods, and selected the identified language using a maximum-likelihood classifier.

There are several reports in the literature [10][5] which indicate that this selection technique is not the best approach and that a neural network classifier is able to achieve better results.

In order to improve the system, we used a multilayer perceptron (MLP) classifier which takes as input the output of the 36 language models, has 100 hidden units and 6 outputs. The output with the highest value identifies the language. This perceptron was trained using backpropagation and the method of adaptive step sizes [6][7].

The global identification results improved from 78.7% to 83.4%.

5. ONLINE SYSTEM

The online language identification system uses a distributed architecture, it consists of one client module, and one or more language identification servers. The client controls a voice acquisition board and runs on a Windows 95 PC The servers decode the acquired speech into phone sequences, and perform the language identification task. Language identification using only one phone recogniser (our baseline system) runs at approximately real time (a 16s utterance takes 15s to be decoded) on a Linux PC with a Pentium II processor at 300Mhz. The full system is implemented with one server serving as a controller which interfaces with the client application and the other servers performing the multiple phone recognitions, one server for each language.

6. RESULTS

The system was tested using a closed set of six European languages. Table 1 shows the identification rates of the baseline system.

Global	DE	EN	ES	FR	IT	PT
71.4%	70.3%	77.4%	66.3%	71.0%	65.2%	88.3%

Table 1: Baseline system identification results.

Portuguese is, naturally, the best identified language, because the acoustic models were trained only with Portuguese speech.

The results achieved with the full system are shown on table 2.

	DE	EN	ES	FR	IT	PT
DE	**83.9%**	*9.1%*	1.3%	3.9%	1.0%	0.9%
EN	*12.0%*	**81.8%**	1.6%	2.0%	1.7%	0.9%
ES	1.1%	1.0%	**84.5%**	2.0%	*6.6%*	4.7%
FR	3.8%	2.0%	3.8%	**86.6%**	3.4%	0.5%
IT	4.8%	2.9%	*14.0%*	6.4%	**70.4%**	1.5%
PT	1.1%	0.4%	5.2%	0.8%	0.0%	**92.5%**
Total	83.4%					

Table 2: Final 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 proximity between some languages can be clearly seen: Germanic languages were most often confused with each other (12% of the English utterances were mistaken as German, and 9.1% of German as English). As far as romance languages are concerned, we see that Spanish and Italian were also easily mistaken with each other (6.6% and 14%). The Portuguese and French results are harder to interpret. The confusion rates between these languages and the others were not significant and stable enough for us to claim any particular proximity. In this table, Portuguese shows some proximity with Spanish, yet this proximity was seldom revealed during development. The

French language is one of the romance languages with more Germanic influence, which might explain why it does not show any significant confusion trend.

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

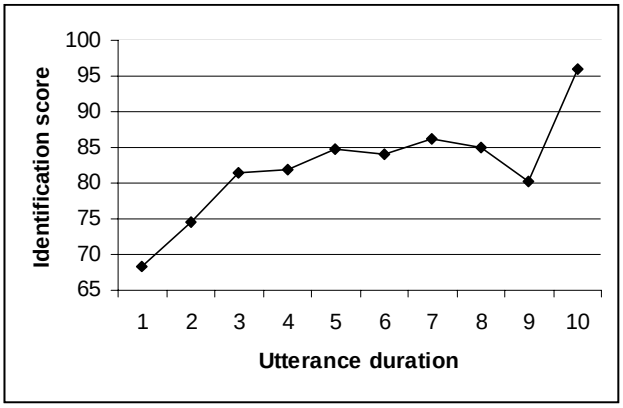

Figure 5: Evolution of the identification rate with utterance duration.

The results are shown on figure 5. 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 86.1% with utterances of 7 to 8 seconds duration.

7. CONCLUSIONS AND FUTURE WORK

In this paper, we described a method for automatic language identification that uses little linguistic information. In order to extend this system to new languages, only speech data in these languages is required. This method was implemented in an interactive realtime language identification system.

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 then the language within the group, may be feasible.

Acknowledgements

The authors would like to thank Portugal Telecom for their support of this work in the scope of the Speechdat project.

The present work was part of Diamantino Caseiro Ms thesis "Identificação automática da língua em fala contínua", and was sponsored by FCT scholarships (PRAXIS XXI/BM/6863/95, and PRAXIS XXI/BD/15836/98).

8. REFERENCES

1. D. Caseiro, I. Trancoso, "Identification of Spoken European Languages", Proc. EUSIPCO-98, Rhodes, Greece, 1998.

2. J. Hieronymous, S. Kadambe, "Spoken language identification using large vocabulary speech recognition." Proc. ICSLP-96, Philadelphia, USA, 1996.

3. M. Lund, H Gish, "Two novel language model estimation techniques for statistical language identification," Proc. EUROSPEECH'95, Madrid, Spain, 1995.

4. J. Navrátil, W. Zühlke, "Double-bigram decoding in phonotactic language identification," Proc. ICASSP-97, Munich, Germany, 1997.

5. J. Navrátil, W. Zühlke, "An Efficient Phonetactic-Acoustic System for Language Identification," Proc. ICASSP-98, Seattle, USA, 1997.

6. F. Silva, L. Almeida, "Acceleration Techniques for the Backpropagation Algorithm", Neural Networks, ED. L. B. Almeida and C. J. Wellekens, Springer, 1990.

7. F. Silva, L. Almeida, "Speeding Up Backpropagation", Advanced Neural Computers, ED. R. Echmiller, Elsevier, 1990.

8. T. Schultz, I. Rogina, A. Waibel, "LVCSR-based language identification." Proc. ICASSP-96, Atlanta, USA, 1996.

9. SPEECHDAT, "European Speech Databases for Telephone Applications", Proc. ICASSP-97, Munich, Germany, 1997. EU-project LRE-63314.

10. Y. Yan, E. Barnard, "An approach to automatic language identification based on language-dependent phone recognition," Proc. ICASSP-95, Detroit, USA, 1995.

11. M. Zissman, E. Singer "Automatic Language Identification of Telephone Speech Messages using Phoneme Recognition and N-Gram Modelling" Proc. ICASSP-94, Minneapolis, USA, 1994.