Multi-lingual Speech Recognition

S. Harbeck and E. Nöth and H. Niemann

Universität Erlangen-Nürnberg,
Lehrstuhl für Mustererkennung (Informatik 5)
Martensstraße 3, D-91058 Erlangen, F.R. of Germany
E-mail: snharbec@informatik.uni-erlangen.de

Abstract. We present the speech recognition module for a multilingual information retrieval dialog system. The system can handle dialogs in any of a predefined set of languages. Recognition is done (theoretically) by running n monolingual recognizers in parallel and deciding for the language which has the best score. A method is presented where the n recognizers are combined into one recognizer. Only allowing transitions between the words from one language, each hypothesized word chain only contains words from one language. It is shown that the approach is as fast and accurate as the respective monolingual recognizers are, because the $n-1$ wrong languages fall out of the search beam at an early state in the recognition process. Results are presented for German/English and Slovenian/Slovak utterances.

1 Introduction

Many scenarios for human machine information retrieval dialog systems concern a travel domain [1, 3]. These scenarios are inherently multi-lingual because many users of such systems would be foreign tourists. Actually, automatic speech understanding (ASU) opens up new possibilities, especially for tourists who only speak their own "small" language like Czech or Finnish and who have consequently trouble getting any information in a foreign country. ASU over the telephone could provide a real improvement, available around the clock and everywhere.

A crucial problem in this scenario is the question, which language is spoken by the current user. In [4] we presented two approaches to language identification in the context of the multi-lingual and multi-functional speech understanding and dialog system SQEL (**S**poken **Q**ueries in **E**uropean **L**anguages). The system is being developed in the EC funded Copernicus project COP-1634. Partners are the Universities of Erlangen (Germany), Kosice (Slovak Republic), Ljubljana (Slovenia), and Pilsen (Czech Republic). The system is intended to handle questions about air flight (Slovenian system) and train connections (German, Slovak, and Czech system) in these four languages. Basis of the system is the EVAR system, the architecture of which is based on the German Sundial demonstrator (ESPRIT project P 2218) [2].

The two system architectures presented in [4] differ in the way the language identification is done:

- Explicit language identification with an additional language identification module before the speech recognition
- Implicit language identification by using n different language recognition systems in parallel and selecting the word chain with the least costs as the correct one

With the language identification module the spoken language can be identified and one of n monolingual recognizers can be started. One advantage of such a system is, that it can be used to reject languages where no speech recognizer is yet available. So, for example, the SQEL system can answer with a prerecorded utterance in Polish when Polish is detected and the system does not yet have a Polish speech understanding module. One disadvantage is the not error free decision of the commonly available language identification systems. Errors in the language identification module can not be recovered by the overall system. So, when an English speaker is identified as speaking German and the German word recognizer is used for automatic word recognition, this will result in a totally incorrect word chain and the failure of the whole dialog.

The second approach has the disadvantage of needing a lot of transcribed training material for all the possible languages. There is no solution to how to react to languages unknown to the system as described in the previous case. The idea is to let all available word recognizers run in parallel and after all word recognizers are finished to select the best one. This selection is not a trivial task because of different normalizations inside the word recognizers. In [6] different approaches using neural networks were presented but none of them gave satisfactory results. Another problem is the computational load, because only one recognizer is working on the correct language and the other $n - 1$ will do unnecessary work.

In this paper we concentrate on the second approach. We present a method for building a multi-lingual recognizer from n monolingual ones. It is shown that the computational load can be reduced to the order that one monolingual recognizer requires. The rest of the paper is organized as follows: In section 2 we will present our integrated approach for multi-lingual speech recognition systems. In section 3 we will present experiments on Slovak and Slovenian inside SQEL, and for German and English. In section 4 we will discuss the results and conclude with an outlook to future work.

2 Design of an integrated multi-lingual speech recognition system

We want to use all knowledge sources that are available as early as possible, i.e. apply n speech recognizers for the language identification process. To reduce the computational load mentioned above, we build a recognizer that contains all words from all languages in its dictionary. By using a stochastic bigram language model that only allows transitions between words within one language, each hypothesized word chain will only contain words of one language.

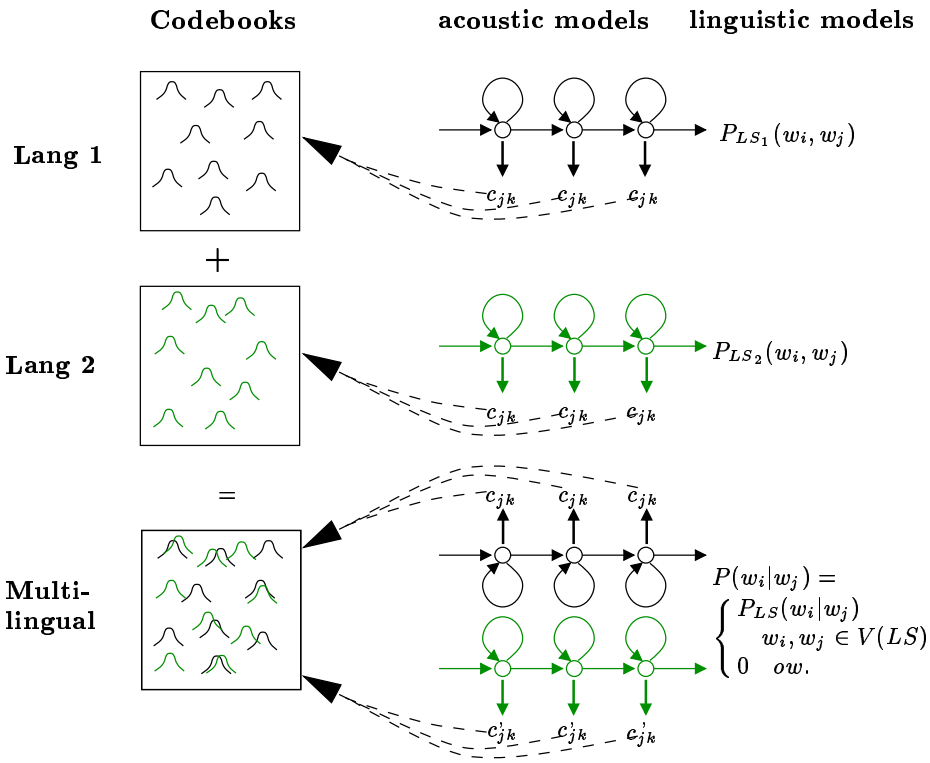

Fig. 1. Construction of the multilingual word recognition system using trained monolingual word recognizers for each language.

The basis for our multi-lingual speech recognition system are monolingual speech recognizers. We use semi-continuous hidden Markov models for acoustic and bigrams for linguistic modeling. The monolingual recognizers are trained in the ISADORA [5] environment which uses polyphones with maximum context as sub word units. The construction of the multi-lingual speech recognizer is as follows (see also figure 1):

1. Increase the number of codebook density functions to reflect the language dependent codebooks. For example when having two different languages with a codebook of 256 density functions per language, then the multi-lingual recognizer will have 512 density functions.
2. Add special weight coefficients to the HMM output density functions to reflect the increased number of available density functions. The new weight coefficients are set zero, so that every density function belonging to different languages has no influence on the output probability of the HMM.
3. Construct a special bigram model which consists of the monolingual bigrams and does not allow any transitions between the languages as shown in equa-

tion 1.

$$P(\text{word}_{language_i} | \text{word}_{language_j}) = 0 \quad \text{for } i \neq j. \tag{1}$$

One might argue that we simply increase the size of the word recognition vocabulary. But the combination of the multi-lingual bigram model and the beam search algorithm in the forward decoding decreases the computational effort. At the beginning of the recognition process every word of the multi-lingual vocabulary is possible, so that there are a lot of different search paths. After a few seconds the most probable paths will be in the correct language. The acoustic models of the other languages should result in paths with lower scores. The beam search algorithm is used to restrict the search space to paths through the word hypotheses graph which contain more reliable hypotheses. Experiments showed that this suboptimal search strategy has no bad effects on the word recognition rate. So using the beam search strategy in forward decoding only paths of the correct language should be expanded. After a few words it should be as fast as the monolingual speech recognition system. In figure 2 the number of states inside the beam for the multi-lingual and the monolingual speech recognizers are compared for one English sentence. At the beginning of the sentence all available languages are possible. Therefore, the number of states is significantly higher than in the monolingual case. After a short time (less than 2 seconds) all states of the wrong language are pruned and the number of states inside the beam is the same as the one for the monolingual recognizer.

3 Experiments

We tested our multi-lingual speech recognizer approach on two different databases. The first database was collected inside the SQEL project and contains read data from the languages Slovak and Slovenian, the second contains spontaneous corpora for German (EVAR) and English (ATIS) (see also tabular 1).

Language	Amount of training data	Number of speakers (calls)	Amount of test data	Number of speakers (calls)	Level of spontaneity
Slovak	4.5 h	30	40 min	4	read
Slovenian	4.5 h	42	40 min	6	read
German	7 h	804	1 h	234	spontaneous
English	7 h	46	2 h	30	spontaneous

Table 1. Description of training and test sets used for the multi-lingual speech recognizer.

In our first experiment we took the baseline system as described in section 2 and evaluated it on the SQEL database (see table2). The time to recognize the

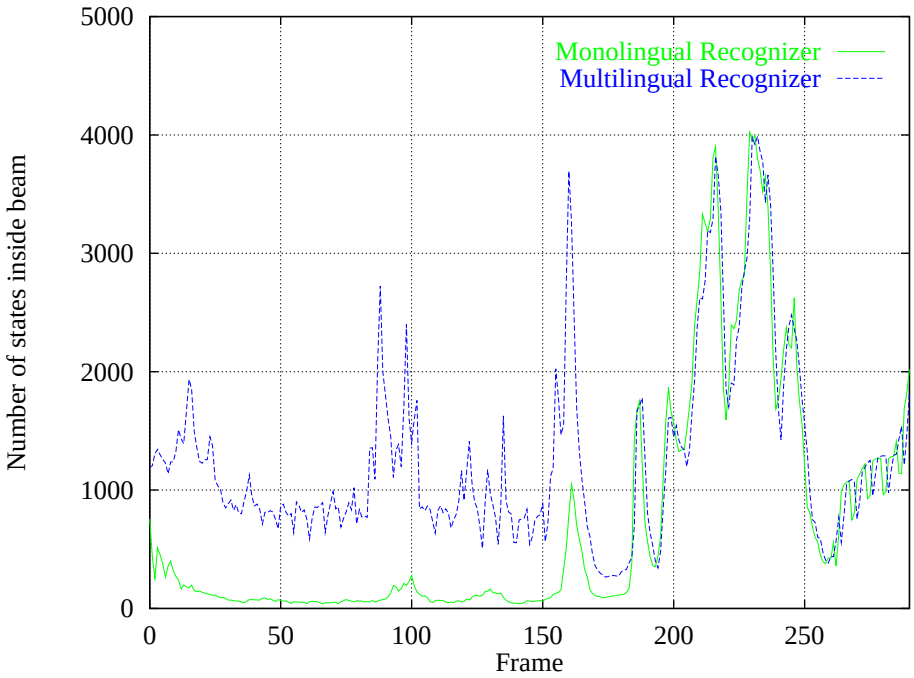

Fig. 2. The number of states inside the English monolingual and the German/English multilingual speech recognizer evaluated on an English sentence. At the beginning of the sentence the multilingual recognizer has much more states inside the beam because all languages are still possible. After 200 frames (2 seconds) all states of German models are pruned and the number of states of both recognizers are the same.

test sets were the same for the monolingual and the multi-lingual recognizers. Additionally the time of a recognizer evaluated on out-of-language is shown in table 2. For example the computation time of a Slovak recognizer on Slovenian utterances takes three times the computation time of a Slovenian recognizer. This is due to the fact that the beam width automatically adapts to the recognition.

The word accuracy of the Slovenian sentences is the same as in the monolingual recognizer, but the accuracy for Slovak sentences decreased by 60 percent. The problem is that the beam search canceled all states of the correct language after a short time in almost all Slovak sentences. Once there are no Slovak states inside the beam, the recognizer cannot return to the Slovak language. When using two different beams inside the forward decoding, one very big beam for the first part of an utterance and one normal one for the rest of the utterance, we increased the recognition rate of Slovak with the side effect of higher computation time.

Tables 3 and 4 show the effect of using a multi-lingual silence category. Instead of using different silence models for each language all silence models of each language are in one common category. This method allows transitions between

Recognition rates (word accuracy)		
Monolingual Slovenian	91 %	
Monolingual Slovak		86 %
Multi-lingual	91 %	29 %

Computation time		
Recognizer	Slovenian	Slovak
Monolingual Slovenian	20 min	1 h
Monolingual Slovak	1 h	20 min
Multi-lingual	20 min	20 min

Table 2. Recognition rates for the multi-lingual word recognizer and the monolingual recognizer in the languages Slovak and Slovenian.

the languages by using a silence model during decoding. The word accuracy using the multi-lingual word recognizer and the computation time was almost as good as using the correct monolingual recognizers in both databases.

WA on Slovenian	**WA on Slovak**	**Computation Time**
90 %	87 %	40 min

Table 3. Using a multi-lingual silence category on the SQEL database.

Recognizer	**WA on English**	**WA on German**
Monolingual	63 %	71 %
Multi-lingual	64 %	71 %

Table 4. Using a multi-lingual silence category on the ATIS/EVAR task.

4 Conclusion

We presented an approach for multi-lingual speech recognition using one multi-lingual speech recognizer for different languages at the same time. This is done using a multi-lingual language model which does not allow transitions between languages. Since the beam search eliminates partial hypotheses with bad scores,

the size of the search space approaches that of the monolingual recognizers. The delay caused by increased vocabulary size should be small. In experiments on the SQEL database with the languages Slovenian and Slovak and on ATIS/EVAR database for English and German it is shown that we get the same performance with the same computational load as using monolingual speech recognizers.

We plan to use this approach for a multi domain application and want to extend the number of different languages inside the multi-lingual recognizer. Because with increasing number of languages the number of output probabilities is growing we want to examine an approach of sharing the same codebook or the same acoustic models for sub word modeling.

5 Acknowledgment

This work was partly funded by the European Community in the framework of the SQEL–Project (Spoken Queries in European Languages), Copernicus Project No. 1634. The responsibility for the contents lies with the authors.

References

1. F. Charpentier, G. Micca, E.G. Schukat-Talamazzini, and T. Thomas. The recognition component of the sundial project. In A. Rubio-Ayuso, editor, *New Advances and Trends in Speech Recognition and Coding*, NATO ASI Series F. Springer Verlag, Berlin, 1993.
2. W. Eckert, T. Kuhn, H. Niemann, S. Rieck, A. Scheuer, and E. G. Schukat-Talamazzini. A Spoken Dialogue System for German Intercity Train Timetable Inquiries. In *Proc. European Conf. on Speech Communication and Technology*, pages 1871–1874, Berlin, September 1993.
3. L.F. Lamel. Report on Speech Corpora Development in the U.S. *ESCA Newsletter*, 8:7–10, 1992.
4. E Nöth, S. Harbeck, H Niemann, V Warnke, and I. Ipšić. Language identification in the context of automatic speech understanding. In N. Pavesic and H. Niemann, editors, *3rd Slovenian–German and 2nd SDRV Workshop*. Faculty of Electrical and Computer Engineering, University of Ljubljana, Ljubljana, April 1996.
5. E.G. Schukat-Talamazzini, T. Kuhn, and H. Niemann. Speech Recognition for Spoken Dialog Systems. In H. Niemann, R. De Mori, and G. Hanrieder, editors, *Progress and Prospects of Speech Research and Technology*, number 1 in Proceedings in Artificial Intelligence, pages 110–120. Infix, 1994.
6. T. Schultz, I. Rogina, and A. Waibel. Lvcsr-based language identification. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, pages 781–784, Atlanta, 1996.