Multilingual Speech Recognition in the Context of Multilingual Information Retrieval Dialogues

Stefan Harbeck, Elmar Nöth and Heinrich Niemann

University of Erlangen-Nürnberg
Chair for Pattern Recognition – Martensstraße 3 – 91058 Erlangen – Germany
(snharbec,noeth,niemann)@informatik.uni-erlangen.de

Abstract. The multilingual speech recognizer implemented inside the SQEL Copernicus project is based on a combination of several monolingual recognizer within one recognizer using a special bigram grammar. Only allowing transitions between the words from one language, each hypothesized word chain contains words from just one language and language identification is an implicit by-product of the speech recognizer. Using this concept on a four language task the multilingual speech recognizer achieves nearly the same accuracy as using monolingual speech recognizers and error free language identification. Additionally a novel language identification module is presented which can be used for preselection of a subset of languages or for setting the apriori probabilities of the languages inside the multilingual speech recognizer. It achieves 76 percent accuracy on a 13 language task.

1 Introduction

Within the Copernicus project SQEL [1] the monolingual German dialogue [2] system was extended for working in a multilingual environment. The SQEL - demonstrator can handle four different languages and domains: German (German train connections), Slovak (Slovak train connections), Czech (Czech train connections) and Slovenian (European flights).

The currently implemented demonstrator uses a multilingual speech recognizer, where the trained monolingual recognizers of the languages to be recognized, the lexicon, and the language models are combined into a single multilingual recognizer [3]. Only allowing transitions between the words from one language, each hypothesized word chain only contains words from one language and language identification is an implicit by-product of the speech recognizer. The demonstrator starts up in German with a German opening phrase and the user is free to use any of the implemented languages. The language is identified implicitly by the multilingual word recognizer which passes to the linguistic processor both the identity of the spoken language and the best matching word chain. The recognition results of the multilingual recognizer are almost as good as when we run the monolingual version of the language that was spoken. Once the language has been identified by the word recognizer, it is associated with the corresponding domain, which calls the appropriate database and task parameters.

When there are a lot of different languages to work with the application of the multilingual word recognizer is not practical. With the help of a separate module for language identification, the number of languages can be reduced inside the multilingual speech recognizer or special actions can be performed when observing languages not handled by the recognizer. The novel concept of language identification presented here is based on stochastic segment models, a generalizations of HMMs, where every state consumes a variable length of observations.

2 Multilingual Speech Recognition

When developing a multilingual dialogue system one of the major parts is the recognition process of the spoken queries. Inside the SQEL system this is done by a multilingual speech recognizer. One method to perform multilingual speech recognition is to run all existent recognizers in parallel and choose the most probable word chain. To reduce the computational load, we build only one recognizer that contains the words from all languages in its dictionary.

The basis for our multilingual speech recognition system are monolingual speech recognizers. We use semi-continuous HMMs for acoustic and bigrams for linguistic modeling. The monolingual recognizers are trained in the ISADORA environment which uses polyphones with maximum context as subword units [4]. The construction of the multilingual speech recognizer is as follows:

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 multilingual 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 equation 1.
$$P(\text{word}_{language_i}|\text{word}_{language_j}) = 0 \quad \text{for } i \neq j. \tag{1}$$
4. Build a special silence category for language specific silence models which allows transition to and from every language, so the language can be switched by inserting pauses.

To reflect the quality of the acoustic models in the different languages an additional apriori value for each language was introduced.

Theoretically using the standard beam search in forward decoding after a few seconds there will be only word chains in the spoken language. The effect is, that the number of words inside the active vocabulary will be the same as using monolingual recognizers.

3 Experiments

Our approach for multilingual speech recognition was evaluated on the four languages of the SQEL project German, Slovenian, Slovak and Czech. Due to our special silence category the recognized word chain can contain words from different languages. For evaluating the accuracy the language of this word chain is determined by counting the number of words of every language and choose the language with the most words. Every word of the wrong language will be deleted from the recognized word chain, so there is a deletion error for every word in the wrong language.

Recognizer	**Recognition rates (word accuracy)**				**RTF**
	Slovenian	Slovak	Czech	German	
Mono Slovenian	88 % (90 %)				1
Mono Slovak		88 % (88 %)			1
Mono Czech			84 % (83 %)		1.3
Mono German				90 % (91 %)	1.2
Multi	83 % (87 %)	86 % (85 %)	84 % (83 %)	84 % (86 %)	2.5

Table 1. Recognition rates and real time factors (RTF) using monolingual and multilingual speech recognizers on all sentences of the SQEL test database and in brackets the recognition rates on all sentences longer than 5 words.

Test set	**Slovenian**	**Slovak**	**Czech**	**German**
All sentences	97 %	90 %	92 %	90 %
Sentences longer than 5 words	97 %	96 %	97 %	96 %

Table 2. Language identification rate using the multilingual recognizer for all sentences and for sentences longer than 5 words.

As shown in table 1 the monolingual speech recognizer is still superior to the multilingual one, because of the language identification failures. These failures occur especially with short sentences, where not enough time for a robust discrimination between languages is available, as it can be seen in table 2.

For a dialogue system only the first sentence of a dialogue must be multilingual. So the language identified with the first sentence will be used for the whole dialogue and a monolingual speech recognizer can be used for every following sentence. When evaluating the mono and the multilingual speech recognizer on utterances with more than 5 words there are only slight differences between the word accuracies of the mono and the multilingual system but the language identification rates are much more higher. The real time factor for the multilingual system is more than two times higher than using monolingual recognizers with given language identification but nearly twice as fast as using 4 monolingual recognizers in parallel. The reasons are, that at the beginning all possible languages are inside the beam and that for every language all codebook densities have to be calculated, regardless if they are inside the active vocabulary or not.

4 Language Identification

When thinking of applications where more than 4 language have to be distinguished the concept of the multilingual speech recognizer becomes unacceptable due to the high computational costs. One method is to make a preselection of a small subset of most possible languages by using an implicit language identification module, which classifies the given utterance according to its language. Also, the result of the language identification modules can be used to set the apriori information of the languages used inside the multilingual speech recognizer.

The basic technology for discriminating languages is to extract phonemes by language specific phoneme recognizers and to score the phone sequences by stochastic language models [5].

In our approach the classification of an observation $\mathbf{X}$ is done according to

$$\mathcal{LS}^* = \underset{\mathcal{LS}_j}{\operatorname{argmax}} P(\mathcal{LS}_j|\mathbf{X}) = \frac{P(\mathbf{X}|\mathcal{LS}_j)P(\mathcal{LS}_j)}{P(\mathbf{X})} \tag{2}$$

The idea is that speech is a sequence of unknown segments s_j like phonemes which could be expressed as

$$\begin{aligned} P(\mathbf{X}|\mathcal{LS}_j) &= \sum_{\mathbf{S}} P_{\mathcal{LS}_i}(\mathbf{S}) P_{\mathcal{LS}_i}(\mathbf{X}|\mathbf{S}) \\ &= \sum_{\mathbf{S}} P_{\mathcal{LS}_i}(\mathbf{S}) \prod_{j=1}^{|S|} P(\mathbf{x}_{s_j}, \ldots, \mathbf{x}_{s_{j+1}-1}|\mathbf{x}_0, \ldots \mathbf{x}_{s_j-1}, s_j) \end{aligned} \tag{3}$$

$P_{\mathcal{LS}_i}(S)$ represents the phonotactic model, $P(\mathbf{x}_{s_j}, \ldots, \mathbf{x}_{s_{j+1}-1}|\mathbf{x}_0, \ldots \mathbf{x}_{s_j-1}, s_j)$ the probability for observing sequence $\mathbf{x}_{s_j}, \ldots, \mathbf{x}_{s_{j+1}-1}$ within the segment s_j, which can approximated by

$$P(\mathbf{x}_{s_j}, \ldots, \mathbf{x}_{s_{j+1}-1}|\mathbf{x}_0, \ldots \mathbf{x}_{s_j-1}, s_j) \approx P(\mathbf{x}_{s_j}, \ldots, \mathbf{x}_{s_{j+1}-1}|s_j). \tag{4}$$

Corresponding to hidden Markov models we can formulate this as

$$\begin{aligned} P_{\mathcal{LS}_i}(\mathbf{X}) &= \sum_{\mathbf{S}} P_{\mathcal{LS}_i}(\mathbf{S}) \prod_{j=1}^{|S|} P_{\mathcal{LS}_i}(\mathbf{x}_{s_j}, \ldots, \mathbf{x}_{s_{j+1}-1}|s_j) \\ &= \sum_{s_1, s_2, \ldots, s_{|S|}} \alpha_{s_1} \cdot \alpha_{s_1,s_2} \cdot \alpha_{s_1,s_2,s_3} \cdots \alpha_{s_1,s_2,\ldots,s_g} \cdot \\ &\qquad \prod_{l=g+1}^{\mathbf{S}} \alpha_{s_{l-g},\ldots,s_l} \prod_{j=0}^{|\mathbf{S}|} P_{\mathcal{LS}_i}(\mathbf{x}_{s_j}, \ldots, \mathbf{x}_{s_{j+1}-1}|s_j) \end{aligned} \tag{5}$$

where $\alpha_{s_{l-g},\ldots,s_l}$ represents the conditional probability $P(s_l|s_{l-g}, \ldots, s_{l-1})$. The parameter g describes the order of statistical dependency, setting $g = 1$ will result in a normal HMM model. Every state within this model consumes not only one observation but a variable length of observations.

One of the big problems of this complex modeling structure is the initialization. To restrict the number of parameters inside the system we allow only segments with the length of 1 which we call the *Codebook approach.* So equation (5) can be simplified to

$$P_{\mathcal{LS}_i}(\mathbf{X}) \approx \sum_{\mathbf{S}} P_{\mathcal{LS}_i}(\mathbf{S}) \prod_{j=1}^{n} P_{\mathcal{LS}_i}(\mathbf{x}_j|s_j). \qquad (6)$$

When $P_{\mathcal{LS}_i}(\mathbf{S})$ is just a unigram model, this can be interpreted as a Gaussian mixture model.

5 Experiments

Due to the fact that our SQEL database is very restricted in domain and number of words we choose another corpus for our language identification experiments which was collected by the Federal Institute for Language Engineering (Bundessprachenamt) in Germany. This corpus was collected via television and radio and contains 13 different languages (German, English, Arabian, Chinese, Italian, Dutch, Polish, Portuguese, Rumanian, Russian, Swedish, Spanish and Hungarian). Contrary to the SQEL database the recording conditions are very variable, the domain is not restricted and spontaneous and read speech are mixed.

For training we used about half an hour of speech per language with 100 different speakers. For test we used about half an hour to 2 hours of speech per language with unseen speakers.

Our first experiment uses the generalized HMM as given in equation (5). For initialization we trained a codebook with 64 classes on the first cepstral coefficients from three consecutive frames. Using this codebook the training material was automatically transcribed and the transcribed corpus was used to train one HMM for every codebook class. All HMMs are combined in the generalized HMM with $g = 1$ and retrained on the training material.

Approach	**RR**	**KRR**
Generalized HMM approach	76 %	74 %
Codebook approach	73 %	76 %

Table 3. Recognition rates (RR) and class wise averaged recognition rates (KRR) for language identification using the *generalized HMM approach* (with $g = 1$) and the *Codebook approach* on 30 second signals.

In table 3 it becomes obvious that this model performs quite good on this difficult task. But one has to keep in mind the high complexity of this algorithm. A suboptimal but fast algorithm is the following (*Codebook approach*):

1. Train the Gaussian mixture model of equation (6)
2. Extract the most possible codebook sequence using only acoustic probabilities $P_{\mathcal{LS}_j}(x_j|s_j)$
3. Rescore the sentence using a codebook bigram $P_{\mathcal{LS}_j}(s_j|s_{j-1})$.

As can seen in table 3 the recognition rate of this very fast approach are nearly the same as using the generalized HMM.

6 Conclusion and Further Work

We presented two strategies for multilingual speech recognition in the context of multilingual information retrieval dialogues. The first approach is to combine the monolingual recognizers to one recognizer. By forcing word transitions to stay within one language, the system identifies the language and decodes the utterance simultaneously. Since the beam search eliminates partial hypotheses with bad scores, the size of the search space approaches that of the monolingual recognizers. Thus, the delay caused by increased vocabulary size is small. With the additional language identification module a subset of language can be chosen which should be activated in the multilingual speech recognizer. Additionally probability measures coming from the language classifier can be used inside the multilingual speech recognizer to weight the a-priori probability of the languages.

In the future we plan to use common codebooks for every language inside the multilingual speech recognizer and to combine both language identification and multilingual speech recognition.

References

1. *Proc. of the 2nd SQEL Workshop on Multi-Lingual Information Retrieval Dialogs*, Pilsen, April 1997. University of West Bohemia.
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. S. Harbeck, E. Nöth, and H. Niemann. Multilingual Speech Recognition. In *Proc. of the 2nd SQEL Workshop on Multi-Lingual Information Retrieval Dialogs* [1], pages 9–15.
4. E.G. Schukat-Talamazzini, T. Kuhn, and H. Niemann. Speech Recognition for Spoken Dialogue Systems. In H. Niemann, R. De Mori, and G. Hanrieder, editors, *Progress and Prospects of Speech Research and Technology: Proc. of the CRIM / FORWISS Workshop*, PAI 1, pages 110–120, Sankt Augustin, September 1994. Infix.
5. M. Zissman. Comparison of four approaches to automatic language identification of telephone speech. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 4:31–44, 1996.

This article was processed using the LaTeX macro package with LLNCS style