EXPERIMENTS WITH LVCSR BASED LANGUAGE IDENTIFICATION

T.Schultz, I.Rogina, and A. Waibel

Interactive Systems Laboratories
University of Karlsruhe (Germany), Carnegie Mellon University (USA)
{tanja,rogina,waibel}@ira.uka.de

ABSTRACT

Automatic language identification is an important problem in building multilingual speech recognition and understanding systems. We have developed a front-end LID module based on LVCSR to identify English, German, and Spanish language for use in spontaneous speech-to-speech translation. We studied the constitution of different levels of knowledge to identify a language, i.e. the phonetic, phonotactic, lexical, and syntactic-semantic knowledge. A comparison of LID systems using different levels of these knowledge sources is presented. We showed that the incorporation of lexical and linguistic knowledge leads to a reduction of the language identification error by up to 50%.

1. INTRODUCTION

In recent years language identification (LID) has received renewed and increased interest as LVCSR technology is being applied to multiple languages. The arrival of multilingual databases like the OGI corpus [1], [2] and the Spontaneous Scheduling Task (SST) [3] enable us to compare different approaches. Language identification is closely related to speaker identification and speaker independent speech recognition. Most of the recent approaches to LID take advantage of units that are smaller than words such as phonemes [4], [5] or broad phoneme classes [6] for the identification process. Many LID systems are based on HMMs [4], [5], [7] or NNs [6],[10]. Some approaches add phonotactic information encoded as phoneme bigrams [5] or trigrams [7], [8]. Nevertheless, most approaches described by literature are restricted to phoneme-based knowledge sources; there is no study which uses word-based knowledge like a dictionary or a language model.

Although language identification can be done on the word recognition level, the phonetic recognition level is certainly very efficient as mentioned in [5]. Constructing dictionaries and word-based grammars for stand-alone LID systems involves greater computational requirements. On the other hand, in multilingual speech processing tasks, in which recognition is the objective, dictionaries, language models and other higher-level knowledge sources are already available. In some applications such as speech-to-speech translations e.g. JANUS the identification of the language could be employed as a front-end module to language-dependent LVCSR. Word level identification using higher linguistic knowledge can be integrated into the recognition process without requiring additional computational effort. Our goal is to show that the integration of word-based knowledge sources leads to improvements in the LID performance.

The paper is organized as follows. In the first section the multilingual database SST which is used for our experiments will be described. After that the influence of different channel conditions on the LID performance is analyzed. Thereafter, five systems using distinct levels of phonemic, phonologic and linguistic knowledge are presented and the performance is compared.

2. THE MULTILINGUAL SPONTANEOUS SCHEDULING TASK

A multilingual database of spontaneous human-to-human dialogs called the Spontaneous Scheduling Task (SST) has been collected at Carnegie Mellon

and Karlsruhe University over the last 20 months. In each session, two people are asked to schedule a meeting with their dialog partners. Constraints for the scenario, the calendar and the collection procedures of the data guarantee the comparability of the data recorded at different sites. The collection scenario and requirements are described in detail in [3] and [9]. The SST corpus currently consists of dialogs in the languages English, German, Spanish, and Korean spontaneously spoken by native speakers. The collection of Japanese dialogs has also begun recently. Table 1 summarizes the current status of data collection and the data used for training and testing our LID systems. Since the database is still growing not all the data were available at the beginning of our experiments.

English SST		
	dialogs	words
recorded	1984	505 K
transcribed	1826	460 K
used for training	117	38809
used for testing	20	4731
German SST		
	dialogs	words
recorded	734	158 K
transcribed	534	115 K
used for training	192	45034
used for testing	18	4107
Spanish SST		
	dialogs	words
recorded	340	79 K
transcribed	256	70 K
used for training	75	61382
used for testing	13	3740

Table 1: The database SST

A dialog is represented by so-called turns or utterances. We found that the number of turns per dialog was language-dependent. English and German dialogs contain on average 10 turns while Spanish dialogs contain on average 16 turns. Furthermore the length of a turn as can be seen in figure 1 is language specific. Table 1 shows that the number of spoken words in Spanish dialogs is much higher than in German and English. On the one hand this is due to the fact that Spanish turns tends to be longer as shown in figure 1 but on the other hand there are many more words per second for Spanish turns. For the experiments the dialogs are divided into a test and a training set of distinct speakers. The identification process is performed by presenting the complete turn to the system.

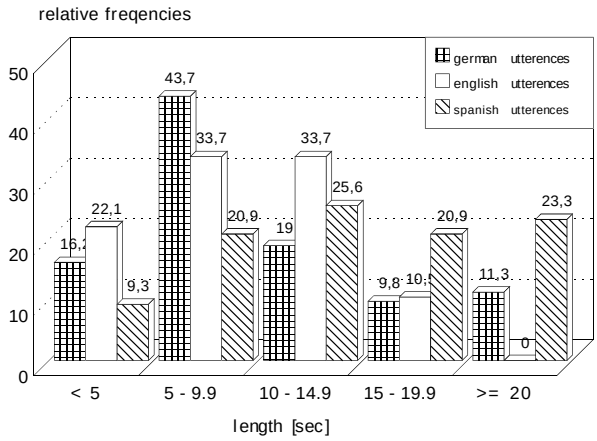

Figure 1: language specific length of the test utterances

3. THE SYSTEM STRUCTURE

There are several kinds of architectures for LID systems. In the first architecture called parallel architecture, for each language that is to be identified a language-dependent system is trained. Language identification of the incoming test turn is performed by running all systems in parallel. Each system decodes the turn with the language-dependent models and calculates the likelihood or the distance score (depending on the decoding algorithms) to determine the best hypothesis for that turn. The language belonging to the system with the highest likelihood or the minimal distance score is hypothesized. This kind of structure is used in [5], [8] and [4] for example. The second architecture called integrated structure consists of a single global recognition system which is language-independent as described in [10] and [6]. One drawback of the integrated stucture is

the increasing ambiguity when adding languages to be identified to the system. We use a parallel architecture as shown in figure 2.

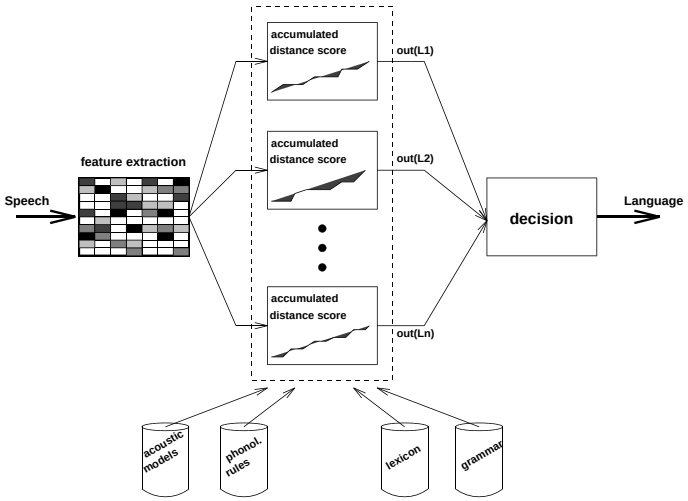

Figure 2: parallel architecture of our LID system

4. CROSS-CHANNEL CONTROL

As mentioned before, many constraints are applied to the data collection procedure to guarantee the comparability of the recorded data. Both of the sites at which we collect our data use the same closed speaker microphones, the same hardware to digitize the speech input, and the same scenario and calendar. To control for possible channel variations or different environmental noise we have recorded additional cross-channel data. We collect German input speech of native speakers in the U.S. under the same conditions as the English input from U.S. and similarly English input speech of English native speakers in Germany under the same conditions as the German input is collected. It is obvious that even slight differences in channel conditions might have considerable influence on the LID results. Many studies in the past do not take into account that channel variations leads to erroneous results. To demonstrate the effect of different channel conditions to our system we performed channel-dependent language identification tests in addition to our cross-channel experiments. The tests as shown in figure 3 and figure 4 are based on the phoneme recognizer system PnoPT which is described in the next section. Figure 3 shows the results of the channel-dependent test by plotting the score calculated by the German recognizer against the score calculated by the English recognizer for each given test turn. As can be seen the languages German and English are easy to separate from each other. For the channel-independent test as shown in figure 4 for the data recorded in Karlsruhe the identification problem becomes much harder.

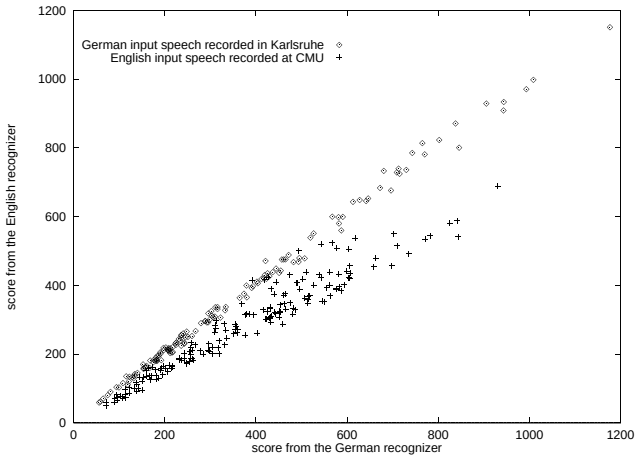

Figure 3: channel-dependent experiments with PnoLM

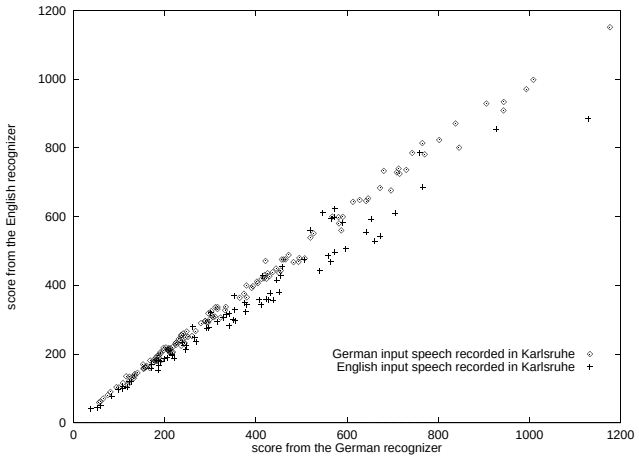

Figure 4: cross-channel experiments with PnoLM

5. LID USING DIFFERENT KNOWLEDGE SOURCES

For each language we constructed five systems applying different levels of knowledge.

5.1. System PnoPT

PnoPT is a recognizer with phoneme-based acoustic modeling. For each language a system with context-independent phonemes which are modeled by SCHMMs with 50 tied mixture weights was build. For the German language we used a set of 46 phonemes, for English 54 phonemes and for Spanish 48 phonemes. The phoneme sets include special noise models to model human and nonhuman noises as described in [11].

5.2. System PwithPT

PwithPT is similar to PnoPT but in addition phonotactics i.e. a phoneme bigram was applied. This phonological knowledge is integrated into the search procedure as presented in [5]. The phoneme accuracy for our PwithPT was 49.6% for German input, 48.3% for English and 46.9% for Spanish speech which is comparable to the performance of other spontaneous spoken speech systems.

The identification process with the system PnoPT is restricted to the short-term acoustic differences between languages, i.e. the use of different phoneme sets and the different realizations of some phonemes in distinct languages. An example for the first is the phoneme /ch/ in the German word *ich* which has no English counterpart. An example for the latter is the phoneme /r/ which has different realizations in English and German. The system PwithPT integrates phonotactic knowledge. For example the transition from /s/ to /w/ in the English word *switch* is not allowed in German.

5.3. System WnoLM

WnoLM is a word-based recognizer including a word dictionary which contains the rules for concatenating phonemes to make words. The phoneme models are similar to PnoPT except that generalized triphones are used to model coarticulation effects. The German dictionary contains 2077 words, the English 1073 and the Spanish dictionary has 2781 entries. The search is implemented as a word-dependent N-best algorithm.

5.4. System WwithLM

This is similar to the WnoLM system but with integrated word bigrams as a form of linguistic knowledge. WwithLM is our JANUS-2 recognizer engine. The word accuracy of the system used in the language identification experiments is 65.8% for German speech, 65.2% for the English input and 63.6% for Spanish speech. For the recognition of speech the integration of a word-based lexicon and grammars leads to very large improvement. We want to analyze how the integration of these knowledge sources helps for the language identification for LVCSR.

5.5. System WpostLM

In the system WwithLM the language model was integrated into the search process, so the identification of languages is a one-stage process. The WpostLM is a two-stage process. In the first step WnoLM is performed to the test turns. In the second step a language model scoring routine is applied to the given first best hypotheses. This routine computes the language model probabilities $p(w_1, w_2, \cdots, w_n|L)$ for a given turn. The language belonging to the turn with the best score is hypothetized. The basic idea is that the language model of the correct language matches best to the first best hypothesis.

Figure 5 shows the improvement reached by the different knowledge sources for the identification test between English and German. To show the effects of channel dependencies for each system we conducted 3 different tests. In the first row in front of the chart labeled with KA is the result of the test on data recorded at site Karlsruhe, the second row describes the test on data recorded at CMU, and the third row are the tests without channel control; that is, data for each language was collected on the different channel of the collecting site. As can be seen in figure 5 the

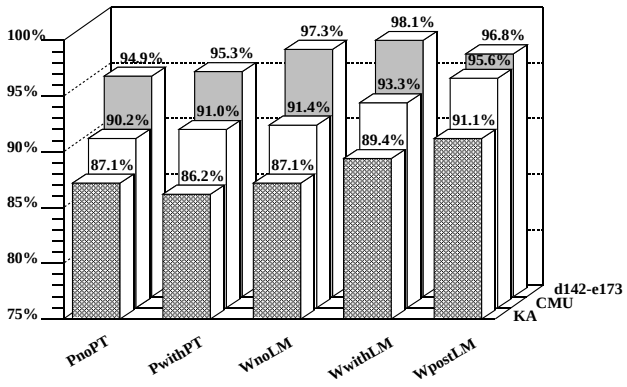

Figure 5: Comparison of the five systems

incorporation of knowledge sources improves the language identification accuracy significantly. For the cross-channel tests system WpostLM leads to best results. In all cases the performance increases when adding the dictionary. Furthermore, tests including the language-dependent word grammars outperform the results of those without linguistic knowledge. Testing under different channel conditions increases the performance significantly. Here the two-stage system WpostLM does not have any impact. Note however, that in practice for language identification systems channel dependencies are undesirable and removed by channel normalisation techniques.

System	G - E	G - S	S - E
data recorded at U.S.			
PnoPT	90.2%	70.2%	91.9%
PwithPT	91.0%	74.9%	89.9%
WnoLM	91.4%	82.1%	96.5%
WwithLM	93.3%	88.6%	97.7%
WpostLM	94.1%	95.2%	90.3%

Table 2: LID-performance for German, English and Spanish input

Table 2 summarizes results from our experiments with the three languages English (E), German (G) and Spanish (S). As can be seen the identification of the two languages English and Spanish seems to be easier than German vs. English, a fact often mentioned in other studies. For all language tests the incorporation of dictionary and language models leads to significant improvements. The WpostLM system is not effective in identification between Spanish and English. The hardest task seems to be the separation of German and Spanish. For this task the system WpostLM improves the performance drastically. The effectiveness of the WpostLM system depends on the number of words in an utterance. Since we are working with bigrams a sentence has to contain at least two words to benefit from the WpostLM system. Therefore the results given in the figure 5 are for those hypotheses which contain more than 3 words. Figure 6 shows the tests on English an German input in which we examined how the performance improves as the number of words increases.

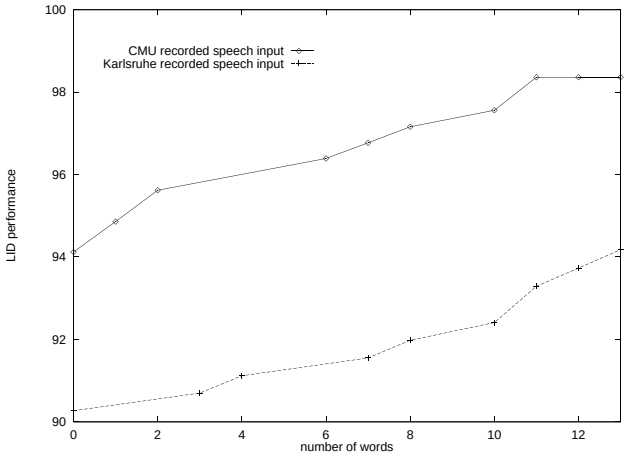

Figure 6: LID performance depending on the number of words

When the number of given words is increased to 6 words, the system identification error is reduced by 5% for data recorded in Karlsruhe and 20% for data recorded at CMU.

6. CONCLUSION

In this paper a new approach to language identification on LVCSR task is presented. We developed five systems with different levels of knowledge sources. Overall the incorporation of higher linguistic knowledge such as the dictionary and language models leads to a reduction in the identification error of about 50%.

7. ACKNOWLEDGEMENTS

The authors wish to thank all members of the Interactive Systems Laboratories, especially P. Geutner, T. Kemp, T. Sloboda, M. Woszczyna, Laura L. Mayfield, and Puming Zhan for useful discussions and active support.

8. REFERENCES

[1] Y.K. Muthusamy, R.A. Cole, and B.T. Oshika: *The OGI multi-language telephone speech corpus.* Proceedings of the ICSLP 1992.

[2] Y.K. Muthusamy, E. Barnard, and R.A. Cole: *Reviewing Automatic Language Identification.* IEEE Signal Processing Magazine, Vol. 11 No. 4 Oktober 1994, pp. 33-41.

[3] M. Woszczyna, N. Aoki-Waibel, F.D. Buø, N. Coccaro, K. Horiguchi, T. Kemp, A. Lavie, A. McNair, T. Polzin, I. Rogina, C.P. Rose, T. Schultz, B. Suhm, M. Tomita, A. Waibel: *JANUS 93: Towards Spontaneous Speech Translation.* Proceedings of the ICASSP 1994, volume 1, pp. 345-348.

[4] M.A. Zissmann and E. Singer: *Automatic Language Identification of Telephone Speech Messages using Phoneme Recognition and N-gram Modeling* Proceedings of the ICASSP 1993, volume 2, pp. 309-402.

[5] L.F. Lamel and J. Gauvain: *Identifying Non-linguistic Speech Features.* Proceedings of the Eurospeech 1993, volume 1, pp. 23-30.

[6] Y. Muthusamy, K. Berkling, T. Arai, R.A. Cole, and E. Barnard: *Comparison of Approaches to Automatic Language Identification using Telephone Speech* Proceedings of the Eurospeech 1993, pp. 1307-1310.

[7] A.A. Reyes, T. Seino, and S. Nakagawa *Three Language Identification Methods based on HMMs.* Proceedings of the ICSLP 1994, pp. 1895-1898.

[8] T.J. Hazen and V.W. Zue *Automatic Language Identification using a Segment-based Approach.* Proceedings of the Eurospeech 1993, pp. 1303-1306.

[9] B. Suhm, P. Geutner, T. Kemp, A. Lavie, L. Mayfield, A.E. McNair, I. Rogina, T. Schultz, T. Sloboda, W. Ward, M. Woszczyna, A. Waibel: *JANUS: Towards Multilingual Spoken Language Translation.* DARPA Speech and Natural Language Workshop 1994.

[10] K.M. Berkling, T. Arai, and E. Barnard: *Analysis of Phoneme-based Features for Language Identification* Proceedings of the ICASSP 1994, volume 1, pp. 289-292.

[11] T. Schultz, and I. Rogina: *Acoustic and Language Modeling of Human and Nonhuman Noises for Human-to-Human Spontaneous Speech Regognition* Proceedings of the ICASSP 1995, pp. 293-296.