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. Building a language identification module for four languages we studied the influence of applying different levels of knowledge sources on a large vocabulary continuous speech recognition (LVCSR) approach, i.e. the phonetic, phonotactic, lexical, and syntactic-semantic knowledge. The resulting language identification (LID) module can identify spontaneous speech input and can be used as a front-end for our multilingual speech-to-speech translation system JANUS-II. A comparison of five LID systems showed that the incorporation of lexical and linguistic knowledge reduces the language identification error for the 2-language tests up to 50%. Based on these results we build a LID module for German, English, Spanish, and Japanese which yields 84% identification rate on the Spontaneous Scheduling Task (SST).

1. INTRODUCTION

In recent years language identification (LID) has received renewed and increased interest as large vocabulary continuous speech recognition (LVCSR) technology is being applied to multiple languages. Most of the recent approaches to LID take advantage of units that are smaller than words such as phonemes [1], [2] or broad phoneme classes [3] for the identification process. Some approaches add phonotactic information encoded as phoneme bigrams [2] or trigrams [4], [5], another approach was presented by [6], using a word-based recognizer. Nevertheless, most approaches for identifying languages are restricted to phoneme-based knowledge sources.

Knowing that the integration of a word-based lexicon and grammars leads to a large improvement in speech recognition systems, we focused our experiments on how such knowledge sources can improve a LID system. Constructing dictionaries and word-based grammars for stand-alone LID systems requires extra effort and LID requires more computational effort on word level than on phoneme level. Nevertheless, in multilingual speech processing tasks, in which recognition is the objective, dictionaries, language models and other higher-level knowledge sources are already available. In speech-to-speech translation applications like JANUS-II [7] the identification of the language could be employed as a front-end module to language-dependent LVCSR. Word level LID using higher linguistic knowledge can be integrated into the speech recognition process without requiring additional computational effort. Even for stand-alone LID systems it is interesting to know whether the additional effort for word-based systems with higher-level knowledge can be justified by better LID performance.

2. THE MULTILINGUAL DATABASE SST

To develop and test our LID system we used a multilingual database of spontaneous human-to-human dialogs called the Spontaneous Scheduling Task (SST). This database has been collected at Carnegie Mellon University (Pittsburgh, USA), Karlsruhe University (Germany), and at ATR International (Japan) over the last two years [8].

Languages	utterances	hours
English	7644	6.9
German	12292	30.5
Spanish	5730	10.7
Japanese	3311	8.0

Table 1: The Spontaneous Scheduling Task SST

The SST corpus currently consists of English, German, Spanish, Japanese and Korean dialogs, spontaneously spoken by native speakers. Table 1 summarizes the currently available data. For the experiments the German, English, Spanish and Japanese dialogs are divided into a test and a training set of distinct speakers. The identification process is performed by presenting a complete utterance to the system.

3. OVERALL SYSTEM STRUCTURE

There are several kinds of architectures for LID systems. An *integrated* architecture consists of a single global recognition system which is language-independent as described in [3]. One drawback is the increasing ambiguity when adding languages to be identified to the system.

In *parallel* architectures, for each language to be identified a language-dependent system is trained, language identification is performed by running all systems in parallel. Each system decodes the utterance with the language-dependent system to determine the best hypothesis and the language belonging to the system with the best score (or highest likelihood) is hypothesized. This kind of structure is used in [2], [5] and [1]. One problem with this approach is, that the scores of the language dependent system cannot be compared without prior normalization.

Therefore, we use a parallel architecture but instead of choosing a normalization we combine the outputs of the different language-dependent systems with a multilayer perceptron to decide on the language spoken as shown in figure 1.

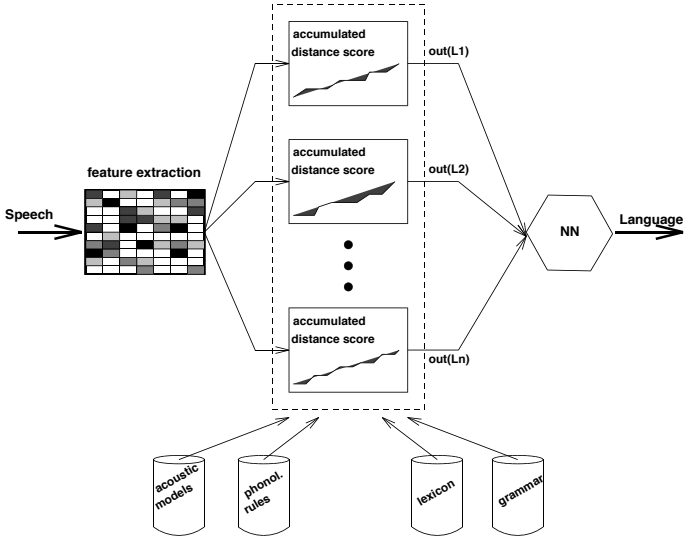

Figure 1: parallel architecture of our LID system

4. EXPERIMENTAL LID-SYSTEMS

To investigate the benefit of different knowledge sources for LID we constructed five systems applying various degrees of knowledge and applied the resulting systems to language pairs [9].

System 1: PnoPT is a recognizer with phoneme-based acoustic modeling. For each language a system with context-independent phonemes which are modeled by CDHMMs with 50 mixture Gaussians was built. 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 different non-speech events as described in [10].

System 2: PwithPT is similar to PnoPT but in addition phonotactics i.e. a phoneme bigram is applied. This phonological knowledge is integrated into the search procedure as shown in [2]. The phoneme accuracy for the PwithPT system 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.

System 3: WnoLM is a word-based recognizer including a pronounciation lexicon which contains the rules for concatenating phonemes to build words. The phoneme models are similar to PnoPT except that generalized triphones are used to model coarticulation effects.

System 4: WwithLM is similar to the WnoLM system but with integrated word bigrams as a form of linguistic knowledge. This is our JANUS-II system used for speech recognition. 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.

System 5: WpostLM is a two-stage process, i.e. in opposite to the WwithLM system the language model is not integrated into the search process. In the first step WnoLM is performed to the test utterances. In the second step a scoring routine is applied to the given first best hypotheses to compute the language model probability $p(w_1, w_2, \cdots, w_n|L)$. The language belonging to the utterance with the highest likelihood is hypothesized. The basic idea is that the language model of the correct language matches best to the first best hypothesis.

5. EXPERIMENTS

Using the five experimental systems we analysed the effect of different levels of knowledge on language identification performance and whether the additional effort of building a word-based LID systems, is justified.

5.1. First Experiments

In earlier experiments we used German data recorded at Karlsruhe and English data recorded at CMU (to get native speakers). The CMU data are collected in a noisy office environment while the data collected at Karlsruhe are very clean. We found that testing under different channel conditions overestimate the language identification performance significantly [9]. To further avoid such influences on our LID-results, we collected additional German data under noisy conditions at CMU and English data under clean conditions at Karlsruhe. We than performed for each 2-language ID system two different tests, one under noisy (CMU) and one under clean (Karlsruhe) conditions as shown in figure 2.

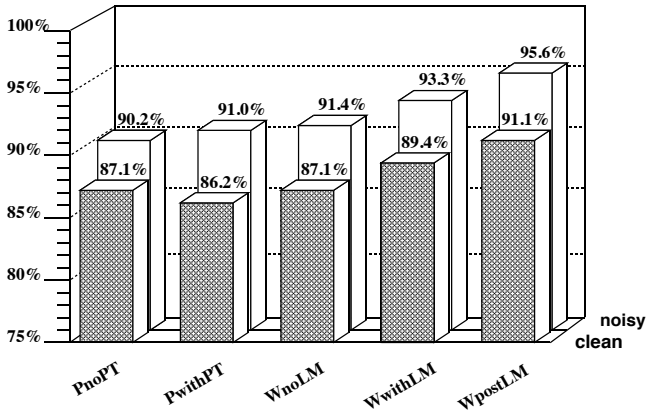

Figure 2: Comparison of the five systems

The front row in figure 2 shows the percent correct LID on clean data, the back row shows the results for the noisy conditions. In all cases the word-based systems outperformed the phoneme-based systems. Applying higher knowledge sources improved the language identification.

Additional to the results above we found that 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 2 are for those hypotheses which contain more than 3 words. Figure 3 shows the tests on English and German input in which we examined how the performance improves as the minimal number of words increases. When the number of given words is increased to 6 words, the system identification error is reduced by 5% for data in clean environment and by 20% for data recorded in the noisier environment.

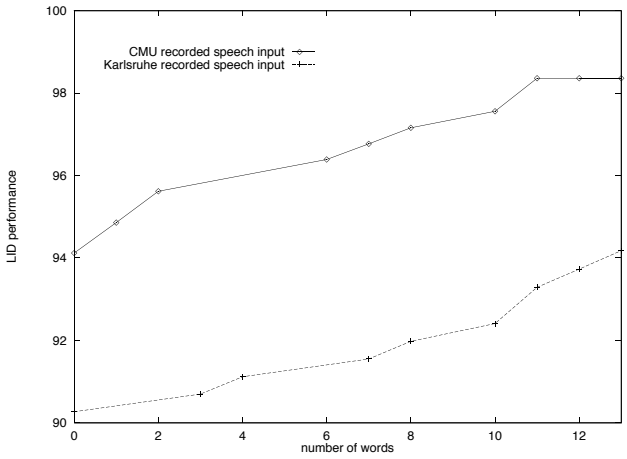

Figure 3: WpostLM depending on the number of words

5.2. 3-language Test

Table 2 summarizes the results from our experiments on English (E), German (G) and Spanish (S) based on data recorded under noisy environmental conditions. The identification of the two languages English and Spanish seems to be easier than German vs English, a fact often mentioned in other studies.

System	G - E	G - S	S - E
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 for German, English, and Spanish

In all cases the performance increases when using lexical knowledge. Furthermore, tests including the language-dependent word grammars outperform the results of those without linguistic knowledge. The word-based systems outperformed the phoneme-based systems significantly. The more knowledge is incorporated in the word-based LID system, the better the performance.

Additionally we performed a test on all 3 languages German, English, and Spanish. Given that we did not want to perform an extra postprocessing step (as for WpostLM), we choose the WwithLM system, which is best for speech recognition, and still one of the best LID systems. WwithLM gave 86% language identification rate on the 3-language test.

5.3. Final System

Finally we built two 4-language systems to identify German, English, Spanish and Japanese. For these final systems we used the new recognizer [7] which were improved in the meantime by e.g. incorporating trigrams into the decoder and better phoneme models for the German recognizer. Therefore we called the new LID systems Pwith3PT and Wwith3LM respectively. The table 3 summarizes the recognition performance and the language identification rate of Pwith3PT and Wwith3LM. Again the word-based system outperforms the phoneme-based system and gave 84% identification rate on the 4-language test.

Language	Pwith3PT Phoneme Accuracy	Wwith3LM Word Accuracy
German	53.1%	69.0%
English	56.1%	69.6%
Spanish	52.0%	69.4%
Japanese	65.5%	70.0%
4-LID	82.6%	84.0%

Table 3: Performance for German, English, Spanish and Japanese

6. CONCLUSION

In this paper we showed how applying different levels of knowledge sources to LVCSR-based LID systems can lead to significant improvements of performance. A comparison of five LID systems, using different levels of knowledge sources, showed that the incorporation of lexical and linguistic knowledge gave up to 50% improvements for the 2-language identification tests. The word-based systems outperformed the phoneme-based systems significantly. The more knowledge is incorporated in the word-based LID system, the better the performance. We want to point out that the recording conditions for different languages have to be similar to get significant LID results. Being aware of this problem we recorded additional data and performed experiments on channel normalisation. Based on the results for language pairs, we built a LID module for German, English, Spanish, and Japanese which gives an overall identification rate of 84% on the Spontaneous Scheduling Task (SST). This LID module is used as front-end for our JANUS-II multilingual speech-to-speech translation demo system.

7. ACKNOWLEDGEMENTS

The JANUS project is partly funded by grant 413-4001-01IV101S3 from the German Ministry of Science and Technology (BMBF) as a part of the VERBMOBIL project. We gratefully acknowledge support and cooperation with ATR Interpreting Telecommunication Laboratories and the University of Electro-Communications in Tokyo, Japan. The authors wish to thank all members of the Interactive Systems Laboratories, especially Tilo Sloboda, Puming Zhan and Torsten Zeppenfeld for usefull discusion and active support. We thank Hagen Soltau for training and testing the Neural Networks.

8. REFERENCES

[1] 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.

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

[3] 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.

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

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

[6] S. Lowe et al.: *Language Identification via Large Vocabulary Speaker Independent Continuous Speech Recognition* Personal Communication.

[7] A. Waibel, M. Finke, D. Gates, M. Gavalda, T. Kemp, A. Lavie, L. Levin, M. Maier, L. Mayfield, A. McNair, I. Rogina, K. Shima, T. Sloboda, M. Woszczyna, T. Zeppenfeld, and P. Zhan: *JANUS-II - Translation of Spontaneous Conversational Speech* to appear in ICASSP 96.

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

[9] T. Schultz, I. Rogina, and A. Waibel: *Experiments with LVCSR based Language Identification.* Proceedings of the SRS 1995, pp. 89-94.

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