CONFIDENCE MEASURE BASED LANGUAGE IDENTIFICATION

F. Metze, T. Kemp, T. Schaaf, T. Schultz, and H. Soltau

Interactive Systems Laboratories
University of Karlsruhe (Germany)
{*metze*|*kemp*|*tschaaf*|*tanja*|*soltau*}*@ira.uka.de*

ABSTRACT

In this paper we present a new application for confidence measures in spoken language processing. In today's computerized dialogue systems, language identification (LID) is typically achieved via dedicated modules. In our approach, LID is integrated into the speech recognizer, therefore profiting from high-level linguistic knowledge at very little extra cost. Our new approach is based on a word lattice based confidence measure [3], which was originally devised for unsupervised training. In this work, we show that the confidence based language identification algorithm outperforms conventional score based methods. Also, this method is less dependent on the acoustic characteristics of the transmission channel than score based methods. By introducing additional parameters, unknown languages can be rejected. The proposed method is compared to a score based approach on the Verbmobil database, a three language task.

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 multi-language problems. Current LID systems are based either on HMMs (e.g. [9], [7], [6]) or Neural Networks (e.g. [8]). In principle, models for each language, which are computed offline, are compared to the unknown speech sample and the best-fitting model determines the output of the LID module. Different model complexities have been evaluated: Phoneme models (e.g. [9], [7]), models for broad phoneme classes [8], or phoneme models with phonotactic bigram [7] or trigram [6] information. More recently, word models with language model, i.e. full LVCSR systems, have been proposed for language identification [11], [12], [4]. Although LVCSR based LID has shown very promising results, both the effort necessary to create LVCSR systems for every language and the computational requirements at run time are generally regarded as too high for most applications.

In many speech recognition tasks however, for example translation or dialogue systems, dictionaries, language models and other higher-level knowledge sources are already available. If LID could be integrated with the speech recognition process, it could use higher linguistic knowledge without 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. SPEECH RECOGNITION IN VERBMOBIL

VERBMOBIL[1] [13] is a multilingual speech-to-speech translation system in a travel arrangement domain. English, German and Japanese speakers can schedule a meeting and arrange a business trip in a dialogue session. As the speaking style of the dialogue partners is not restricted, spontaneous phenomena like stuttering, false starts and nongrammatical sentences as well as (background) noises occur.

Training data was recorded through close speaking microphones and cellular phones. In the demonstration system, speakers are free to share one input device or to switch between devices in the course of the dialogue. VERBMOBIL cuts every turn of input speech into shorter segments, which can then be processed by three monolingual speech recognizers, even before the completion of the turn. The demonstration system can run several speech recognizers in parallel. The LID module can therefore evaluate the output of several monolingual speech recognizers, but must work on an initial chunk of speech, as CPU time is needed for other system components such as semantic analysis, translation and speech synthesis, once an initial hypothesis on the language and content of the speech input is available. In order to keep the responsiveness of the system as high as possible, the length of the initial segment used for LID should be as short as possible.

Characteristics and performance of the recognizers used in this work are summarized in table 1.

Language	Vocab.	Training data	OOV rate	Trigram PP	Error rate
English	7k	32h	1.0%	47.3	29.0%
German	10k	57h	1.0%	93.4	23.0%
Japanese	2.8k	30h	2.6%	17.2	12.4%

Table 1: Characteristics and performance of the speech recognizers

3. THE CONFIDENCE MEASURE

In our experiments, we use the gamma confidence measure [3], which is basically an a-posteri word probability computed on a word lattice. The computation begins with the word lattice which is the output of our recognizer. The word lattice is interpreted as an HMM, with the nodes of the HMM being the words, and the links

[1] http://www.dfki.uni-sb.de/verbmobil/

of the HMM restricting the possible succession of words. The emission probabilities for the nodes are the (acoustic) scores of the words, and the state transition probability from one word node to the next is given by the (trigram) language model. With this interpretation, a forward-backward algorithm can be computed over the word lattice, which assigns a posterior probability to each of its nodes and links. The resulting posterior probabilities are used as the measure of confidence. In several experiments [2] [3], the *gamma* measure has outperformed all other single confidence measures.

4. EXPERIMENTS

4.1. Baseline

Table 2 summarizes the results from our previous experiments [4] on English and German. In all cases the performance increases when using lexical knowledge. Furthermore, tests including the language-dependent word grammars yield better results than those without linguistic knowledge. The word based systems outperformed the phoneme based systems.

Base:	Phonemes		Words	
Method:	phonotactics		language model	
	no	yes	no	yes
Error-rate	9.8%	9.0%	8.6%	6.7%

Table 2: Performance of different score based LID methods

4.2. Data

The tests described in this work were conducted on the VM database.[2] Different parts of the database and the names we use throughout this paper to refer to them are described in table 3.

Name	Lang.	Channels	# utts.	Remarks
E	Eng.	1, 2	504	Ch. identical with G
e	Eng.	3, 4, 5, 6	504	Ch. different from E
e'	Eng.	3, 4	224	Subset of e, parts not recorded in studio
e"	Eng.	5, 6	280	Subset of e
G	Ger.	1, 2	467	Ch. identical with E
j	Jap.	7, 8	500	Test-set for jap. rec.

Table 3: The different parts of the VERBMOBIL database referred to in this paper

The English utterances E share the same channels with the German utterances G. For evaluation purposes all parts were divided equally into a development-set and a test-set. Turn length varied between 1.8s and 32.2s, with an average length of 7.9s.

[2]For more information contact: Bavarian archive for Speech Signals, http://www.phonetik.uni-muenchen.de/Bas/BasHomeeng.html

Score based "Best-of" LID	E–G	E–j	G–j	Overall (trilingual)
Error-rate	10.1%	1.0%	1.0%	7.2%

Table 4: Error-rates for LID using a score based "Best-of" classificator

4.3. Score based LID

Word based speech recognizers minimize the score associated with a path through the word lattice. The ratio of a specific utterance's score per unit time to an average score computed over the development-set, gives a measure for how good that utterance fits this recognizer's acoustic and language models. Figure 1 shows these normalized scores the English and German recognizer produce for their respective best hypothesis on the 1471 utterances of our data.

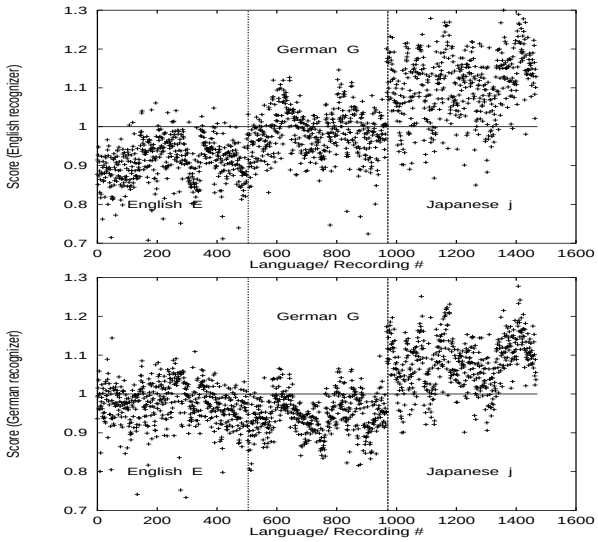

Figure 1: Normalized scores from the English (top) and German (bottom) recognizer for English E (left), German G (middle), and Japanese j (right) utterances

Score based LID can be performed by assigning each utterance (or turn) to the language whose recognizer has produced the best (lowest) normalized score. The error-rates we achieved when using this "Best-of" approach are shown in table 4.

The error-rate when discriminating Japanese from English and German is one order of magnitude lower than when discriminating English from German, it is therefore necessary to scrutinize the dependency of this type of LID on channel properties.[3]

We therefore replaced the 504 English turns E with 504 other utterances e taken from the same domain. The resulting scores are shown in figure 2. The English scores now show a large cloud of scores, corresponding with different recording conditions: the utterances e' have been collected in several different rooms. Generally, the English scores increased for both the English and German

[3]Not only was Japanese recorded at a different site, but it was also stored on a DAT-tape prior to cutting and labeling, which was not the case with the other data.

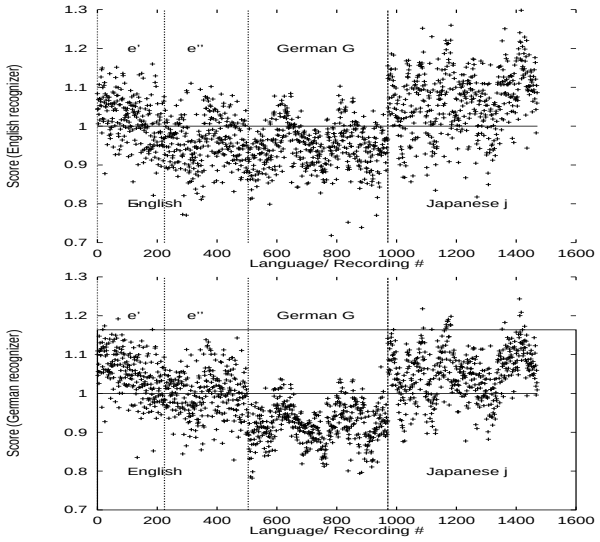

Figure 2: Normalized scores from the English (top) and German (bottom) recognizer for multi-channel English (left), German (middle), and Japanese (right) utterances

Score based LID	e–G	e'–G	e"–G
ER with renormalization	13.1%	14.7%	11.8%
ER w/o renormalization	15.3%	17.9%	13.2%

Table 5: Channel dependency of error-rates for score based LID between German and English. The scores are shown in figure 2

recognizer. Although the domain is identical for all utterances, the intra-class variance of scores due to channel effects has the same order of magnitude as the inter-class differences. It is therefore not surprising, that LID error-rates for English (e, e' and e") and German increase in this case.

Table 5 gives the LID error-rates with and without recalculation of the normalization factor. In practical applications, the LID often is not aware of changes in the input channels and can therefore not adapt to the new situation.

4.4. Confidence based LID

The *gamma* confidence measure attaches a confidence to every word in the word graph. To arrive at a single confidence value for a whole utterance, we calculated the arithmetic mean of all words of the best hypothesis. It is therefore not necessary to introduce further factors or constants.

Figure 3 shows the average word confidence assigned to the channel identical utterances E and G by the English and German recognizer. The corresponding "Best-of" error-rate is given in table 6. The number of overall errors is reduced by 10% as compared to the score based method and the distribution of error-rates for the three bilingual subtasks is better balanced, indicating less channel dependence.

Inspecting figure 3, it seems feasible to distinguish English, German and Japanese using only two recognizers by the following "Threshold" decision rule:

Corpus	E–G	E–j	G–j	Overall
Error-rate	4.9%	4.4%	3.3%	6.4%

Corpus	e–G	e'–G	e"–G	e–j	G–j	Overall
Error-rate	1.9%	2.9%	1.1%	1.2%	3.3%	4.0%

Table 6: Error-rates for LID using the confidence based classificator

Decision rule	$C_1 < T_1$	$C_1 > T_1$
$C_2 < T_2$	3	1
$C_2 > T_2$	E	$argmax(C_1, C_2)$

where C_X denotes the confidence of recognizer X's output and T_X a threshold for that recognizer. The confidence based classificator in this case does of course not identify the third language as such, but it rejects a language that does not match phonetic and/ or linguistic models of any recognizer. The threshold values, which were computed on the development-set, are shown as horizontal bars in figure 3.

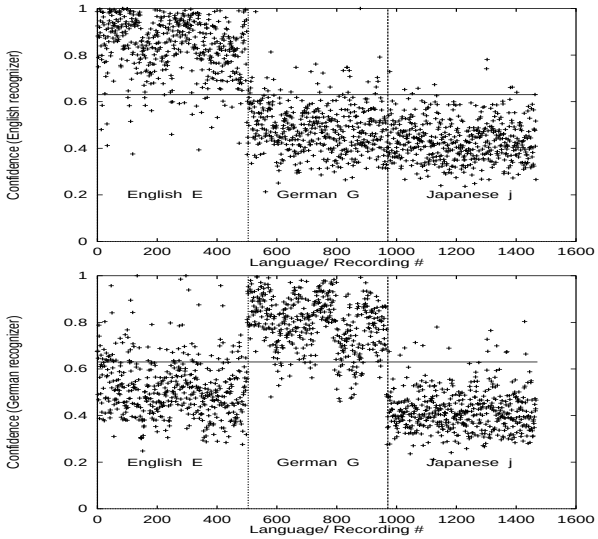

Figure 3: Average word confidence assigned to the English E (left), German G (middle), and Japanese j (right) utterances of the VM database by the English (top) and German (bottom) recognizer. English and German share the same channels

Table 7 summarizes the results. Using the confidence measure, the error-rate on the three language task using only two recognizers is lower than using the score based LID on two languages alone.

4.5. Performance on short segments

To evaluate the performance of the LID methods on short segments, we tested the algorithms on the first three seconds of each turn. The exact starting position of this three second segment of speech was calculated by a power based segmenter. The results of this experiment are shown in table 8 and 9.

Error-rate e–G–j	Recognizer Pair		
	Eng./ Ger.	Eng./ Jap.	Ger./ Jap.
Score	15.3%	35.3%	40.5%
Confidence	8.0%	13.5%	15.5%

Table 7: Trilingual LID using only two recognizers and thresholds. English and German data share the channel

Error-rate	E-G	E-j	G-j	Overall
Score	7.2%	3.0%	1.2%	6.9%
Confidence	4.9%	5.4%	5.4%	8.2%

Table 8: LID on a 3s chunk from the start of each turn for different language pairs. As was discussed in section 4.3, score based values for Japanese are in fact largely to channel identification

It is interesting to note that the performance on the e–G subtask with two recognizers does actually improve when using only the initial three seconds. Looking at the recognizer output, we attribute this to a greatly reduced language perplexity in our task at the beginning of each turn[4], leading to a significantly lower language model score per frame for the correct language.

5. CONCLUSION

Confidence measure based LID was shown to outperform traditional score based language identification methods with respect to both classification error and robustness against channel influences on a three language task.

Using three recognizers it was possible to distinguish three languages, two of which share the same input channel, with an error-rate of 6.4%, compared to 7.2% for the score based approach. On the two channel-identical languages, the confidence based LID reached an error-rate of 4.9%, compared to 10.1% for the score based LID. If the data of one language was replaced by data that had been recorded under several different conditions, the score based LID's performance deteriorated to 14.9%, because channel influences on the scores are bigger than language influences. The confidence based approach however improved to an error-rate of 4.0%, without the need to recalculate parameters on account of the changing input channel.

Future research will be directed towards the relation of the baseline word error-rate of the underlying speech recognizer to the LID's error rate and the behaviour of the confidence measure if speed-ups such as beams or Look-Ahead systems are used aggressively. Also, the influence of language models and domain mismatches as opposed to channel mismatches will be investigated.

[4]For example, over 50% of the English hypothesizes start with one of the following 11 words: I, I'm, #NIB# (Coughing, ...), hi, good, okay, hello, my, all, so, can.

Error-rate E–G–j	Recognizer Pair		
	Eng./ Ger.	Eng./ Jap.	Ger./ Jap.
Score	33.5%	24.7%	31.3%
Confidence	12.9%	16.0%	17.3%

Table 9: Trilingual LID using two recognizers and thresholds on a three second chunk from the start of each turn

6. ACKNOWLEDGMENTS

This work is partly funded by grant 413-4001-01IV101S3 from the German Ministry of Science and Technology (BMBF) as a part of the VERBMOBIL project. The authors wish to thank all members of the Interactive Systems Laboratories for useful discussions and active support.

7. REFERENCES

[1] M.A. Zissman and K.M. Berkling: *Automatic Language Identification*, Proc. Multilingual Interoperability in Speech Technology, pp. 93-101, Leusden 1999.

[2] T. Schaaf, T. Kemp: *Confidence measures for spontaneous speech*, Proc. ICASSP-97, Vol 2, pp. 875 ff., Munich, Germany, 1997

[3] T. Kemp, T. Schaaf: *Estimating confidence using word lattices*, Proc. Eurospeech 97, Vol 2, pp. 827 ff., Rhodes, Greece, 1997

[4] T. Schultz et al.: *LVCSR-based Language Identification*, Proc. ICASSP, pp. 781-784, Altlanta 1996.

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

[6] A.A. Reyes, T. Seino, and S. Nakagawa: *Three Language Identification Methods based on HMMs.* Proc. ICSLP, pp. 1895-1898, 1994.

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

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

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

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

[11] S. Mendoza et al.: *Automatic language identification using large vocabulary continuous speech recognition.* Proc ICASSP, pp. 785-788, Atlanta, 1996.

[12] J.L. Hieronymus and S. Kadambe: *Robust spoken language identification using large vocabulary speech recognition.* Proc ICASSP, pp. 1111- 1114, 1997.

[13] R. Karger and W. Wahlster: *Multilinguale Verarbeitung von Spontansprache* KI 4/1997 (in German).