Multilingual Speech Recognition

Alex Waibel, Hagen Soltau, Tanja Schultz, Thomas Schaaf, and Florian Metze

Institut für Logik, Komplexität und Deduktionssysteme, Universität Karlsruhe, Germany

Abstract. The speech-to-speech translation system Verbmobil requires a multilingual setting. This consists of recognition engines in the three languages German, English and Japanese that run in one common framework together with a language identification component which is able to switch between these recognizers. This article describes the challenges of multilingual speech recognition and presents different solutions to the problem of the automatic language identification task. The combination of the described components results in a flexible and user-friendly multilingual spoken dialog system.

1 Introduction

Verbmobil is a multilingual speech-to-speech translation system, which opens a large spectrum of potential applications for international communication and cooperation. With the available languages German, English, and Japanese, Verbmobil covers widespread languages with more than 700 million native speakers in the world. Given the intrinsic multilinguality of Verbmobil the users would intuitively expect that the system accepts each of the three languages as input. This feature requires the ability of Verbmobil to automatically identify the language which is spoken by the user and provide him or her with the corresponding recognition engine. In the following sections we discuss such a multilingual setting which consists of recognition engines in the three Verbmobil languages that run in one common framework together with a language identification component which is able to switch between these recognizers. We first describe the recognition engines in the common multilingual framework and then present several solutions to the language identification problem. Incorporating these components into Verbmobil results in a very flexible and user-friendly multilingual spoken dialog system.

2 Language Differences

In this sections, we first discuss differences between the Verbmobil languages and highlight the resulting challenges of multilingual speech recognition, i.e. combining the different features into one framework. Recognition results in these languages are presented and compared to each other. Language differences that affect meaning or interpretation are beyond the scope of this paper.

2.1 Scripts and Fonts

Many different character types are used in the world's languages. Writing systems fall into two major categories: ideographic and the phonologic. In ideographic scripts, the characters reflect the meaning rather than the pronunciation of a word. Phonological scripts can be further divided into syllable-based scripts where each grapheme reflects one syllable, and alphabetic scripts where graphemes correspond roughly to one phoneme. German and English are using both the alphabetic latin script, whereas the Japanese writing system is one of the most complicated in the world. It uses the ideographic script *kanji* with about 7000 commonly used characters and two syllable-based alphabetic scripts *kana*, one for foreign words *katagana* and one for native words *hiragana*.

Phonologic scripts are often easier to handle than ideographic scripts in the speech recognition framework, as in many cases rule-based grapheme-to-phoneme tools can be used to generate the pronunciation dictionary needed to guide recognition, while this is usually not possible for ideographic scripts. However, among the languages using alphabetic scripts, the grapheme-to-phoneme relationship varies considerably. It ranges from a nearly one-to-one relationship such as for some Slavic languages up to languages like English that require complex rules and have many exceptions.

2.2 Romanization and Segmentation

English has a natural segmentation into words that can conveniently be used as dictionary units for speech recognition. The words are long enough to differ from each other in a sufficient number of phonemes, but short enough to be able to cover most material with a reasonable number of different word forms that occur frequently. This is important for the statistical analysis required by the automatic learning processes that modern speech recognition systems rely on.

Unfortunately Japanese lacks an adequate segmentation. Here sentences are written in strings of characters without any spacing. Taking these character strings as dictionary units for speech recognition would be not feasible. In order to determine appropriate dictionary units, the transcribed speech data has to be segmented manually or by morphological analysis programs. In the Japanese Verbmobil database all words are basically segmented into morphological units *bunsetsu*, using the Japanese morphological analyzer CHASEN (Matsumoto, 1997). The resulting segmentation is then error-checked by human experts (Kurematsu et al., 1999) following the definition of word units provided by CHASEN and by the "Daijirin" dictionary (Matsumura, 1985).

2.3 Prosodic Structure

Across the world's languages, the prosodic structure of words varies considerably. Japanese is a pitch accent language where pitch contrasts are drawn between polysyllabic words. English and German belong to the stress languages where individual syllables in a polysyllabic word are stressed. Both are lexical stress languages,

where the stress position is fixed for each word but varies across words. Prosody is not directly incorporated into the speech recognition process but is attached to the recognizers output afterwards and plays an important role in the Verbmobil system as described by Batliner et al. in **??**.

2.4 Morphology

Two major groups of languages can be distinguished when comparing their morphological properties: languages like English that show exceptionally simple morphological structure, and morphological rich languages like German and Japanese. German is a highly inflected language having a large number of distinct verb conjugations and noun declinations and beside that having lots of compounds. This results in rapid growth of the number of word-forms occurring in a given test. As a consequence, poor recognition results are achieved when using a certain set of word-forms as dictionary entries for speech recognition, and many new word forms are encountered in unseen speech material (Out-Of-Vocabulary words, OOV). For

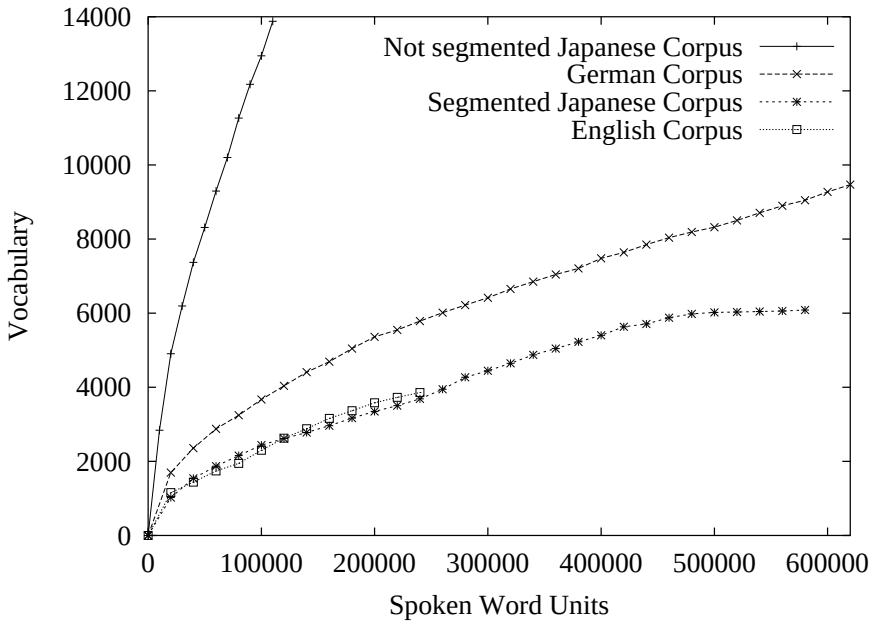

Figure 1. Vocabulary growth comparing German, English and Japanese

the Japanese language it would even not be feasible to use the original word forms as dictionary units, as discussed above. Figure 1 compares the vocabulary growth of the three Verbmobil languages and illustrate their differences. In case of Japanese the

vocabulary growth of segmented units is compared to that of not segmented units. It turns out that after segmentation the vocabulary growth in Japanese language is quite similar to English.

3 Speech Recognition

While the Verbmobil-I task covers spontaneous speech spoken in a very cooperative fashion, the data of the second phase is much more challenging in terms of speaking style, cross-talks and realistic spontaneous effects of conversational speech. Additionally the domain coverage was extended by the travel task, which nearly doubles the trigram perplexity on the testset.

Although it turns out that the Verbmobil languages are quite different, we developed recognition engines for spontaneously spoken speech in all three languages and manage to run them in one common framework. The development of all recognizers is based on our modular speech recognition toolkit (Schultz et al., 1997 and Finke et al., 1997). The observed differences in recognition accuracy are partly due to some of the above described language differences and to the corpus characteristics. The corresponding characteristics and numbers of the data and systems are summarized in Table 1.

Table 1. Speech Recognition: Characteristics of German, English, and Japanese

Item	German	English	Japanese
Training Data	62 hrs	32 hrs	39 hrs
Vocabulary (word forms)	10254	7965	3490
Pronunciation variants per word	1.15	1.20	1.08
Phoneme inventory	47	41	32
Acoustic Model (Quinphones)	3300	2250	2500
Language Model Corpus	670K	270K	580K
Trigram Testset Perplexity	77	47.3	17.3
Testset OOV-rate	1.0%	1.0%	2.6%

The German recognition engine is trained with roughly double the amount of speech data, since within the Verbmobil corpus the main focus was on the collection of German dialog data. The German vocabulary size is above 10.000 word forms which provides an excellent coverage of the scheduling and travel arrangement domain. Many English words are added to the English vocabulary list in order to get a corresponding coverage. Due to this fact the vocabulary size differs from the observed vocabulary growth in Figure 1. In case of Japanese most of the vocabulary words are spoken in the dialogs, its handy size results from the word segmentation.

English has the highest number of pronunciation variants ($1.20 \approx 2000$ additional variants) to cope with the large number of cross-word coarticulation effects

like for example in *gonna*, *wanna*, *gotta* and with the large variety of speakers dialects covered in the English part of the Verbmobil speech database. The context width for acoustic modeling is ± 2 for all languages with an underlying phoneme inventory which varies from compact (Japanese) to quite large (German). The number of quinphones, which we found to be highly language dependent, is chosen according to the likelihood on a cross validation set and systems complexity.

The English language models suffers from the small amount of training text data compared to Japanese and German and the high number of words added to the vocabulary list. This fact is reflected in the number of OOV-words which is relatively high for the English language. The low trigram testset perplexity of the Japanese language model is remarkable. On the one hand this is a result from the segmentation giving short and handy vocabulary units but on the other hand it turns out that Japanese speakers act in the spoken dialogs in a very disciplined way. Therefore the content and the used words in the dialogs stick closer to the given scenario. However the smaller vocabulary list results in higher OOV-rates. On the opposite site the German speakers are much more spontaneous which results in a large number of crosstalk events and OOV-words.

3.1 Acoustic Modeling and Training Algorithms

The Karlsruhe recognition engine has been improved over the years by incorporating state-of-the-art techniques and adding new features for acoustic and pronunciation modeling. In particular the following items gave significant improvements to the German Verbmobil system:

- Speaker incremental normalization
- Feature space adaptation
- Semi-tied covariances
- Context dependent noise modeling
- Filler models to detect unknown proper names
- Flexible transcription alignments
- Pronunciation modeling

One challenging task in speaker and channel normalization is the robust estimation of the adaptation parameters. For this purpose we implemented Cepstral Mean Subtraction (CMS), Vocal Tract Length Normalization (VTLN) and Feature Space Adaptation (FSA) in a speaker incremental fashion. A more powerful adaptation/ normalization technique than VTLN is the maximum likelihood linear regression (MLLR). Unfortunately these techniques are not feasible in a real time application. As a consequence of fast score computation (BBI, Gaussian selection) the model space can not be changed during decoding. Therefore we implemented a general linear feature based transformation technique. Both techniques, VTLN and FSA gave us substantial additive gains. In current state-of-the-art recognizer the emission probabilities are modeled by diagonal covariances since even with large speech corpora it is not possible to estimate full covariance matrices accurately. To overcome this problem we trained semi-tied covariances (Gales, 1997) based on broad phone

classes. Semi-tied covariances try to find optimal feature spaces for the broad phone classes in a maximum likelihood procedure. We combined the semi-tied covariance approach with the linear discriminant analysis (LDA).

The recognition performance highly depends on accurate training labels. Especially spoken speech with many spontaneous effects is hard to transcribe. The technique of flexible transcription alignments (Finke, 1997) allows us to cope with spontaneous effects. Another challenge of the Verbmobil task are the dialectal variations and effects like word fragments or cross-word coarticulation (multiwords). We addressed this problem by pronunciation learning algorithms which gave significant improvements. By context dependent modeling of human spontaneous effects (Hesitations, Breathing, etc.) we got further improvements. Due to the nature of human-to-human dialogs in the Verbmobil tasks the problem of unknown proper names has to be addressed. For this purpose we introduce filler models to detect these unknown proper names during the dialog.

All the described techniques result in error rate reductions which add up to substantial improvement of our evaluation system. Like in all former Verbmobil evaluations the Karlsruhe speech recognizer achieved best performance in the final official Verbmobil evaluation.

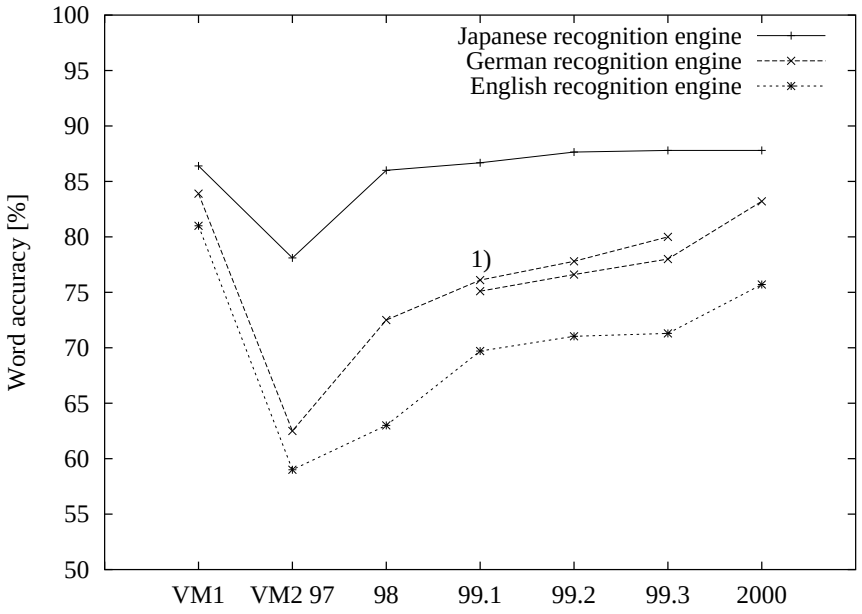

Figure 2. Speech recognition performance in all Verbmobil languages
1) = Internal evaluation set is changed from ”dev98” to ”dev99”

3.2 Systems Performances

Due to the expansion of the Verbmobil-II task to a more spontaneous conversational speaking style and to a larger domain in the beginning of the second phase we expected a much lower recognition accuracy compared to Verbmobil-I. From Figure 2 which shows the word recognition accuracy achieved for the three languages it can be seen that this assumption of performance degradation in the beginning of the second phase (VM2 '97) can be observed in all three languages.

Figure 2 illustrates the development of the systems in all three languages over the last years. The final numbers of our last (internal) evaluation set shows, that the performance degradation could nearly be compensated by improving the systems over the last years. In the framework of Verbmobil we are now able to recognize much more spontaneously spoken speech in all three languages with word accuracies between 76% and 89%. When comparing the systems across languages the discussed language differences should be taken into account. The Japanese gives the highest accuracies which can on the one side explained by its restricted phonotactics and compact phonetic inventory which makes Japanese acoustically easy to recognizer but must be on the other hand extrapolated from the low perplexity and the small vocabulary list compared to German and English.

4 Language Identification

Given the recognition engines in three languages a Language IDentification (LID) component provides the Verbmobil system with the ability to directly identify the user's language, therefore greatly improves the translation system's flexibility and user-friendliness.

Several approaches to the LID task have been investigated (Zissman and Berkling, 1999), although rapid LID on the first few seconds of an utterance, as is needed for a dialogue system, is still a challenging task. We therefore implemented and tested two different approaches to the LID problem:

- Score-based: Using a multilingual or several monolingual recognizers, the score (a number describing the "fit" between the acoustic evidence and the stored models established by the search) for each language is computed. The best score determines the language.
- Confidence-based: A confidence value is assigned to each recognizers output by some appropriate method. The highest confidence determines the language of the utterance.

4.1 Score-based LID

Score-based LID is straight forward and well proven (Muthusamy et al., 1993), as the score is the value speech recognizers try to optimize during the recognition process. Our experiments on English and German data using both phoneme- and

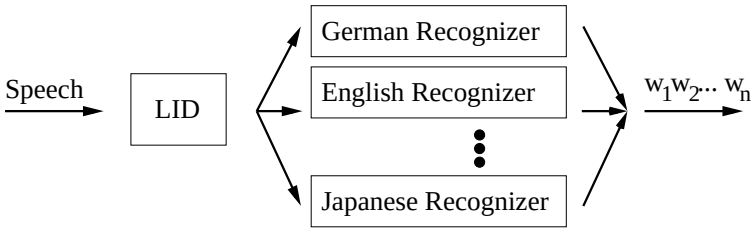

Figure 3. Front-end Language IDentification (LID)

word-based recognizers gave the error rates shown in Table 2. In both cases, the performance increased when using higher lexical knowledge and language-dependent word grammars as could be shown in (Schultz et al., 1996). The word based systems outperformed the phoneme based systems. In the Verbmobil system, these higher-level knowledge sources are readily available, so a word-based approach using language models can easily be implemented. Such an LID component is illustrated in Figure 3.

Table 2. Performance of different score-based LID methods

Based on	Phonemes		Words	
	w/o phonotactics	with phonotactics	w/o language model	with language model
Error-rate	9.8%	9.0%	8.6%	6.7%

Score-based approaches however suffer from channel effects, which means that the absolute value of a score is highly dependent on the acoustic characteristics of the transmission channel (Schultz et al., 1996). The LID problem then becomes inseparable from the channel identification problem, which means that the recognizer that produces the best score is not the one that was trained on the right language, but the one that was trained with acoustic data that most resembles the current test-data. This effect has been observed on the Verbmobil corpus, which contains English, German and Japanese data from several channels as well as channel-identical English and German data. For these experiments, we used the recognizers shown above.

Table 3. Error-rates for LID using a score-based classifier

Score based LID	E–G	E–J	G–J	Overall (trilingual)
Error-rate	10.1%	1.0%	1.0%	7.2%

Table 3 shows the results for LID on channel-identical English and German Data compared to Japanese data from a different channel[1].

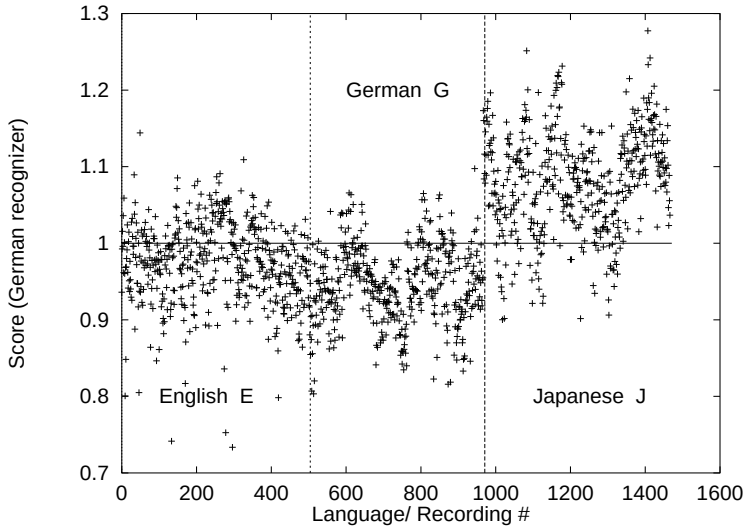

Figure 4. Normalized scores from the German recognizer for English, German, and Japanese utterances

Channel effects are responsible for the large discrepancy in error-rates and average score for the various languages, as can be seen by the imbalanced error rates in Table 3 and from another experiment in which we replaced the English data E by data $e = e' \cup e''$, which was also taken from the Verbmobil corpus, but which had been recorded under various *different* acoustic conditions. These results are shown in Table 4.

The error-rates for the subsets e' and e", corresponding to different channels, differs significantly and the overall error rate is higher than before. Also, the scores for the English data vary over a wider range than before. When used for LID or similar purposes, scores are usually normalized in order to compensate for the characteristics of a specific acoustic channel. A slow change in acoustic properties (e.g. the speaker moves, background noise changes , etc.) is usually hard to detect, so that a renormalization is difficult to achieve in most practical applications. The normalized scores for the first experiment are shown in Figure 4 (lower score means better match).

A more stable solution is therefore needed for the Verbmobil system, which is required to function in a variety of environments and without complicated adjustment procedures.

[1] Channel-identical data was not available for Japanese.

Table 4. Channel dependency of error-rates for score-based LID between German and English.

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%

4.2 Confidence-based LID

The confidence measure *gamma* attaches a confidence to every word in the word graph. To arrive at a single value for the whole utterance, we calculate a phrase confidence using the word confidences from the recognizers best hypothesis (Metze et al., 2000).

Gamma is basically an a-posteriori word probability computed on a word-lattice. To calculate it, the word lattice is interpreted as an HMM, with the nodes of the HMM being the words and the links 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 (Schaaf and Kemp, 1997 and Kemp and Schaaf, 1997), the *gamma* measure has shown very good performance.

Figure 5 shows the average word confidence assigned to the channel identical utterances E and G by the English and German recognizer. The corresponding error rate is given in Table 5. The number of overall errors is reduced by 10% when compared to the score based method and the distribution of error-rates for the three bilingual subtasks is better balanced, indicating less channel dependence.

Table 5. Error rates for LID using the confidence based classifier

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%

To save resources, it is also possible to discern three languages by using two recognizers (we used English and German running in parallel) and hypothesize the third language, if the output of both recognizers (score or confidence) falls below a certain threshold. Using this approach, we were able to reach the LID rates shown in Table 6.

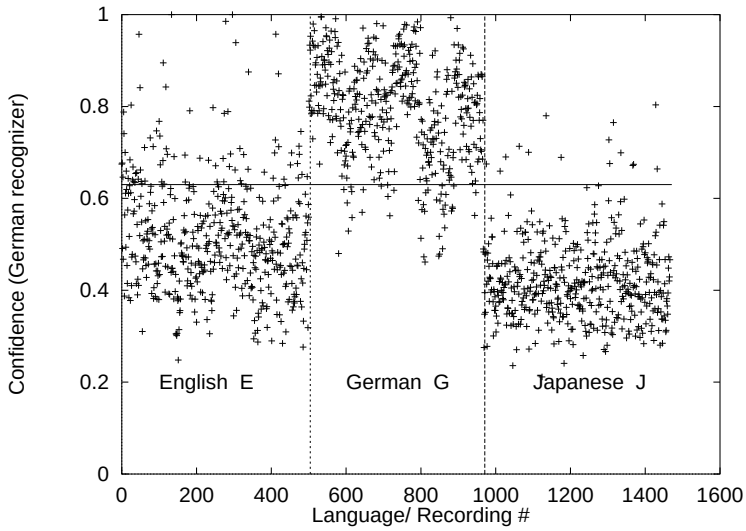

Figure 5. Average word confidence assigned to the English, German, and Japanese utterances of the VM database by the German recognizer. English and German share the same channels. The threshold used in our experiments is also indicated.

Table 6. Trilingual LID using only two recognizers and thresholds. English and German data share the same channel

Error-rate	Recognizer Pair		
Trilingual. LID	Eng./ Ger.	Eng./ Jap.	Ger./ Jap.
Score	15.3%	35.3%	40.5%
Confidence	8.0%	13.5%	15.5%

To calculate the phrase confidence, we used the arithmetic mean of the word confidences. In order to further improve the performance of the system, we did not count confidences attached to noises and noise-like words and of the remaining words only used the words with the highest confidence. Their fraction was derived from the recognizers word accuracy. We then reached the performance shown in Table 7.

In the Verbmobil demonstration system, only the first three seconds of speech from each utterance can be evaluated, in order to guarantee rapid system response. In this case, it also improved the performance, not to take into account the last word from the recognizer output, as it is probably incorrect due to the segmentation at this point. Using the first three seconds of speech in each turn to identify a language, the Verbmobil languages German, English and Japanese can therefore be identified with an accuracy of more than 87% using two recognizers only. The score-based approach gave error rates of more than 25% on that task while additionally suffering from the channel identification problem described above.

Table 7. Summary of the performance of the Verbmobil LID module

LID Error-rate	"Best-Of" decision rule				"Threshold" decision rule		
	Trilingual	E/G	E/J	G/J	Trilingual (E+G rec.)	E/G (E rec.)	E/G (G rec.)
Full utterance	3.1%	3.4%	1.1%	0.9%	5.9%	7.7%	6.0%
First 3 seconds	7.3%	6.1%	2.3%	3.4%	12.9%	14.9%	10.0%

Using this confidence-based approach developed during the Verbmobil project, it is possible to integrate language identification into the recognizer, therefore using high-level knowledge at no extra cost, reaching low overall error rate and establishing stability against changes of channel characteristics without the need to readjust parameters on the fly.

5 Summary and Conclusion

In this paper we described our final recognition engines which have been developed for the three languages German, English, and Japanese in the Verbmobil task. Combined with the language identification component the user is provided with a flexible and user-friendly multilingual spoken dialog system.

6 Acknowledgments

The authors gratefully acknowledge support and cooperation with ATR Interpreting Telecommunication Laboratories and the University of Electro-Communications in Tokyo. We wish to thank all members of our Verbmobil group at the Interactive Systems Laboratories.

References

Finke, M. and Waibel, A. (1997). Flexible Transcription Alignment. In: *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*. The Institute of Electrical and Electronics Engineers (IEEE), Signal Processing Society. 34–40.

Finke, M., Geutner, P., Hild, H., Kemp, T., Ries, K. and Westphal, M. (1997). The Karlsruhe-Verbmobil Speech Recognition Engine. In: *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*. The Institute of Electrical and Electronics Engineers (IEEE), Signal Processing Society. 83–86.

Gales, M. (1997). Semi-tied full-covariances matrices for Hidden Markov Models Available as *Technical Report, CUED/F-INFENG/TR 287, Cambridge University Engineering Department*,

Kemp, T., and Schaaf, T. (1997). Estimating confidence using word lattices. In: *Proceedings of the European Conference on Speech, Communication and Technology, Eurospeech*. European Speech Communication Association (ESCA). 827–830.

Kurematsu, A., Akegami, Y., Schultz, T., and Burger, S. (1999). Development of Data Collection and Transliteration of Japanese Spontaneous Database in the Travel Arrangement Task Domain. In: *International Workshop on East-Asian Language Resources and Evaluation (Oriental COCOSDA)*.

Matsumoto, Y. (1997). Japanese Morphological Analysis System: CHASEN. Available as *Information Science Technical Report NAIST-IS-TR97007*, Nara Institute of Science and Technology.

Matsumura, A. (1985). *Daijirin Dictionary*. Sanseido Publishing.

Metze,F., Kemp, T., Schaaf, T., Schultz, T., and Soltau, H. (2000). Confidence measure based Language Identification. In: *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*. The Institute of Electrical and Electronics Engineers (IEEE), Signal Processing Society. 1827–1830.

Muthusamy, Y., and Berkling, K. M., Arai, T., Cole, R. A., and Barnard, E. (1993). Comparison of Approaches to Automatic Language Identification using Telephone Speech. In: *Proceedings of the European Conference on Speech, Communication and Technology, Eurospeech*. European Speech Communication Association (ESCA). 1307–1310.

Schaaf, T., and Kemp, T. (1997). Confidence measures for spontaneous speech. In: *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*. The Institute of Electrical and Electronics Engineers (IEEE), Signal Processing Society. 875–878.

Schultz, T., Rogina, I., and Waibel, A. (1996). LVCSR-based Language Identification. In: *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*. The Institute of Electrical and Electronics Engineers (IEEE), Signal Processing Society. 781–784.

Schultz, T., Koll, D., and Waibel, A. (1997). Japanese LVCSR on the Spontaneous Scheduling Task with JANUS-3. In: *Proceedings of the European Conference on Speech, Communication and Technology, Eurospeech*. European Speech Communication Association (ESCA). 367-370.

Zissman, M. A., and Berkling, K. M. (1999). Automatic Language Identification. In: *Proceedings of the workshop on Multilingual Interoperability in Speech Technology*. European Speech Communication Association - NATO. 93–101.