THE KARLSRUHE-VERBMOBIL SPEECH RECOGNITION ENGINE

Michael Finke *Petra Geutner* *Hermann Hild* *Thomas Kemp* *Klaus Ries* *Martin Westphal*

Interactive Systems Laboratories
University of Karlsruhe, Germany
Carnegie Mellon University, USA

ABSTRACT

Verbmobil, a German research project, aims at machine translation of spontaneous speech input. The ultimate goal is the development of a portable machine translator that will allow people to negotiate in their native language. Within this project the University of Karlsruhe has developed a speech recognition engine that has been evaluated on a yearly basis during the project and shows very promising speech recognition word accuracy results on large vocabulary spontaneous speech. In this paper we will introduce the Janus Speech Recognition Toolkit underlying the speech recognizer. The main new contributions to the acoustic modeling part of our 1996 evaluation system – speaker normalization, channel normalization and polyphonic clustering – will be discussed and evaluated. Besides the acoustic models we delineate the different language models used in our evaluation system: Word trigram models interpolated with class based models and a separate spelling language model were applied. As a result of using the toolkit and integrating all these parts into the recognition engine the word error rate on the German Spontaneous Scheduling Task (GSST) could be decreased from 30% word error rate in 1995 to 13.8% in 1996.

1. INTRODUCTION

Verbmobil is a long-term research project aimed at machine translation of spontaneous speech input. The ultimate goal of the Verbmobil project is the development of a portable machine translator which is capable of assisting business people from different countries to negotiate with each other in their native language. As first language to start with a German speech recognition system was to be built. In order to get a representative mix of different German dialects, data has been collected at four different sites in Germany. The domain of the system is restricted to scheduling of meetings, but no artificial restrictions are placed upon the speakers and their speaking style. Therefore, spontaneous phenomena like noise, stuttering, restarts and non-grammatical sentences occur. All of these phenomena have to be dealt with in both the speech recognition and the machine translation component of Verbmobil.

In this paper we will give an overview of the Karlsruhe-Verbmobil Speech Recognition Engine – a large vocabulary spontaneous speech recognizer developed to be used as speech recognition component in the Verbmobil speech-to-speech translation project.

2. JANUS RECOGNITION TOOLKIT

The Karlsruhe-Verbmobil Speech Recognition Engine is based on the Janus Speech Recognition Toolkit (JRTk) developed at the Interactive Systems Laboratories in Karlsruhe and at Carnegie Mellon University in Pittsburgh. This toolkit implements a new object-oriented approach. A flexible Tcl/Tk script based environment allows building state-of-the-art multimodal recognizers – this includes speech, handwriting and gesture recognition. Unlike other toolkits Janus is not a set of libraries and precompiled modules but a programmable shell with transparent, yet very efficient objects.

Ranging from mixture of gaussian hidden markov models and hybrid neural network-HMM recognition approaches to hierarchical mixture of experts models a large variety of recognition approaches is addressed in our group. For all those approaches objects are available within the toolkit that serve as building blocks for applications. The underlying data structures of all these objects can be inspected and modified at script level. This makes Janus an easy-to-use testbed for new research ideas. It also offers a great flexibility allowing rapid prototyping. The Tk component adds a graphical user interface to the recognition toolkit thereby simplifying setting up and running demos. The toolkit passed its first test with the Janus Switchboard recognizer which was top ranking in DARPA's spring 96 LVCSR evaluation and currently has a state-of-the-art error rate of 36% [4, 10].

3. ACOUSTIC MODELING

Currently, approximately 32 hours of labelled spontaneous speech training material is available for training the acoustic models of our speech recognition engine. We will discuss in greater detail the preprocessing steps and the polyphonic modeling approach, because they can be considered the major new contributions to the 1996 system in terms of word accuracy.

3.1. Preprocessing

From the short time spectral analysis of the 16 kHz sampled audio recordings we derive a 16ms wide power spectrum that is calculated every 10ms. Based on these spectral features the further preprocessing steps can be summarized as *speaker normalization*, *channel normalization*, and *speech feature extraction*.

3.1.1. Speaker Normalization

One major source of interspeaker variability in automatic continuous speech recognition is the variation in vocal tract shape among speakers. Andreou et al [1] proposed a set of maximum likelihood speaker normalization procedures to explicitly compensate for these variations. Based on the observation that the position of the spectral formants peaks for utterances are inversely proportional to the length of the vocal tract, these procedures reduced speaker dependent variations between formant frequencies through a simple linear warping of the frequency axis. As a consequence a *speaker normalization* step was introduced into our preprocessing.

In Janus we implemented a maximum likelihood approach similar to [7], where the goal is to determine a frequency warping factor $\hat{\alpha}$ such that the warped speech signal fits best to the acoustic models. Let O_i^{α} be the acoustic observation vectors for utterance i warped by α based on a piecewise linear warping function as described in [9]. During both testing and training, the warp scale α is estimated by maximizing the likelihood of the utterance $P(O_i^{\alpha}|W_i)$, where W_i denotes the corresponding transcription of the utterance (for training this is the presumably correct transcription and for testing the hypothesis derived in a first search pass). Since it is very difficult to obtain a closed form solution for the optimal warping factor we used a grid search over a set of 12 different factors to determine the shape of the warping function. Experiments showed that referring to the likelihood of voiced frames only instead of computing the likelihood of all speech frames to find the best warping factor, reduced the word error rate by 5% relative.

Table 1 shows the performance of our baseline system and the system with vocal tract length normalization. VTLN reduces the error rate by 12%. Using the hypothesis of the first search pass instead of the correct transcription turned out to be almost as good as taking the reference to estimate the warping factor.

System		WER in %
no VTLN		21.7
VTLN	using reference	18.6
	using hypothesis	19.0

Table 1. Word error rate of the VTLN system.

Figure 1 shows the distribution of the optimal warping factors in the training set for all male and female speakers respectively. Warping factors above 1.0 correspond to frequency compression and those above 1.0 to frequency expansion.

Speaker variability may also be dealt with by defining speaker clusters and training different acoustic models for them. One way of clustering is to separate male and female speakers. With gender-dependent modeling we could achieve a 2% relative decrease in WER compared to the speaker independent non VTLN system assuming perfect gender detection. That means that gender-dependent modeling is outperformed by VTLN as it was observed on SWB, too [9]. One explanation is that the speaker clustering and subsequent training of independent acoustic models reduced the training data for each recognizer to about the half. The VTLN approach on the other hand aims at normalizing with respect to the speakers' vocal tract shape. With the vocal tract length normalization reducing the variability between speakers we can build more compact models and thus make more efficient use of the acoustic parameters.

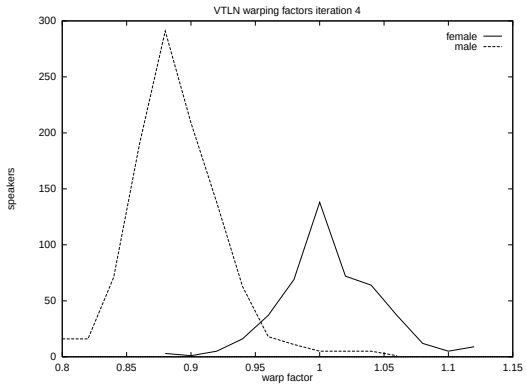

Figure 1. Distribution of warping factors.

3.1.2. Channel Normalization

Our *channel normalization* is a variant of the standard cepstral mean subtraction. The mean of a whole utterance that will be subtracted from each speech vector in the cepstral or log-spectral domain is a simple estimate of the acoustic channel. Since it contains not only channel distortions but also averaged speech (that should also be removed) the estimate depends on the silence-to-speech relation. Especially with utterances containing longer pauses – as it is the case in a spontaneous task like Verbmobil – we get a better and more consistent estimate when considering only speech frames to calculate the mean vector. In our Janus system this is done using a simple energy based speech detector. The "take all frames" method helped to reduce the word error rate by about 6% relative.

3.1.3. Speech Feature Extraction

For *speech feature extraction* we used a 30 dimensional melscale filterbank and derived 13 cepstral coefficients from it. The channel normalization technique is applied to this mel-cepstral feature stream. For the normalized coefficients we added the first and second order derivatives and reduced the dimension of the input space from 39 to 32 coefficients using linear discriminant analysis.

3.2. Polyphonic Clustering

Context-dependent acoustic models have been applied in speech recognition research for many years, and have proven to increase the recognition accuracy significantly. The most common approach is to use triphones. Recently, several speech recognition groups have started investigating the use of larger phonetic context windows when building acoustic models. We also make use of a larger context in our recognizer by allowing questions in the allophonic decision tree not only referring to the immediate neighboring phones but also to phones further away (for Verbmobil we used a con-

text of two instead of the context of one as in the triphone setup).

In a two stage decision tree based clustering approach the codebooks are clustered first and, based on the clustered codebooks, in a second step the distributions are clustered. For Verbmobil we ended up having 2500 codebooks and 10000 distributions. This clustering approach implementing a flexible parameter tying scheme gave us significant improvement across all tasks WSJ, SWB, and Spontaneous Scheduling Task, and across all languages involved (German i.e. Verbmobil, Spanish, English) [3].

4. LANGUAGE MODELING

In terms of language model training material the Verbmobil domain is a fairly small spontaneous speech corpus. As baseline we use a trigram backoff model with absolute discounting and non-linear interpolation. Like on the much larger Switchboard corpus long-range language models like cache models did not result in any WER reduction [10].

4.1. Function and Content Words

In order to introduce longer-term dependencies than conventional trigrams, some linguistic constraints were introduced into our language models. The notion of function and content words [5] was used in order to predict the next word not only based on the last word pair, but also on the last function/content word pair. An improvement of 0.4% WA abolute was achieved.

4.2. Interpolation and Class-Based Models

The Verbmobil domain contains 300.000 words with a vocabulary of 6000 words, i.e. trigram backoff models are potentially not well trained. To make up for the lack of training data, word-dependent linear interpolation of the baseline language model with models built on different corpora was used. Also class-based trigram models [6] were applied where each word is assigned to exactly one class. We achieved a word error reduction of 0.3% absolute by interpolating the baseline with a class-based Verbmobil model and a model built on a large German newspaper corpus (FAZ).

4.3. Domain adaptation and phrase models

Due to changes to the recording scenario within the course of the project there was a small domain shift in the collected data. It seems that the unigram distribution is most influenced while the conditional class probabilities $p(c_i|c_{i-1}c_{i-2})$ remain stable. So the idea is to adapt a language model to a new target corpus as:

$$\hat{p}(w_i|w_{i-1}w_{i-2}) \quad = \quad \hat{p}(w_i|c_i) \cdot p(c_i|c_{i-1}c_{i-2})$$

where $\hat{p}(w|c)$ is estimated on the target corpus and $p(c_i|c_{i-1}c_{i-2})$ on a corpus sufficiently similar to it. The classes were found by an adaptive clustering algorithm, a variant of [6] that minimizes the perplexity of the adapted bigram model.

We also achieved a word accuracy improvement around 0.5% absolute using phrases of words as the base unit of language modeling [8] on an earlier version of the system without retraining acoustics. Since the Verbmobil evaluation conditions did not allow phrases in the lattices we didn't apply it. In the final evaluation model the domain adaption technique was for conservative reasons just applied to adapt the newspaper corpus to Verbmobil.

Model	PP	WA
Standard Trigram	42.26	78.8
Standard Class Trigram	40.40	78.8
Class Trigram, $p(w\|c)$ adapted	40.33	79.2
same, but adaptive clustering	40.16	79.5
+ Std. Trigram on small corpus	38.61	
+ Std. Class Trigram on small corpus	38.39	
+ Std. Trigram	38.29	79.8
same, but no adaptive clustering	38.75	79.7
same, but $p(w\|c)$ not adapted	38.80	79.6

Table 2. Language model experiments.

4.4. Integrating Spelling Sequences

A further difficulty in the Verbmobil corpus is the presence of spelling sequences. If a language model is directly computed from the text corpora over the recognition vocabulary $V = V_W \cup V_L$ of words $V_W = \{w_1, \ldots, w_N\}$ and letters $V_L = \{A, \ldots, Z\} = \{L_1, \ldots, L_M\}$, transitions both within as well as into or out of the letter sequences will be poorly modeled due to the small number (a few hundred) of available spelling examples.

To allow for a more robust recognition, we can assume a letter sequence and its embedding text independent and collapse all letter sequences to an equivalence class `LS`. A new language model LM_W is then computed over the vocabulary $V_W \cup \{\texttt{LS}\}$, resulting in transitions $P_W(\texttt{LS}|w_i)$ and $P_W(w_j|\texttt{LS})$ instead of the less robust estimates for $P(L_j|w_i)$ and $P(w_j|L_i)$. The letter bigrams $P_L(L_j|L_i)$ within a sequence can be modeled by a separate, independent "sublanguage model" LM_L. The final language model LM recombines LM_W and LM_L. For example, $P(L_j|w_i) = P_W(\texttt{LS}|\texttt{w}_\texttt{i}) \cdot \texttt{P}_\texttt{L}(\texttt{L}_\texttt{j}|\texttt{<s>})$.

Depending on the task, LM_L can be computed from a separate text source or an equal distribution can be assumed. In addition, a duration model can be implemented as illustrated in figure 2, where a length of one, two or more letters is explicitly modeled with probabilities $\frac{1}{4}$, $\frac{3}{4}\frac{2}{3} = \frac{1}{2}$, and $\sum_{i=0}^{\infty} \frac{3}{4}\frac{1}{3}(\frac{1}{4})^i\frac{3}{4} = \frac{1}{4}$, respectively.

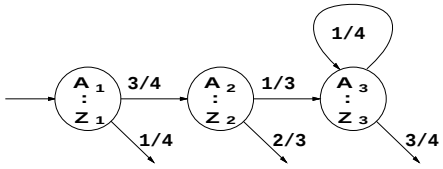

Figure 2. Duration modeling for letter sequences.

For the final evaluation, we used a length model with five states which discouraged one- and two-letter sequences in order to avoid false letter insertions. As most of the spellings in the training material were four-letter acronyms, the length model was adjusted to encourage sequences of this length. All bigrams in LM_L were considered equally likely. The trigrams found in the training texts were added to LM_L to account for the repetitive occurrence of some

of the acronyms. With this measures, the recognition rate measured on the letter sequences improved from 89% to 92% on our development test set, which resulted in an relative overall improvement of 1%.

5. EVALUATION

To assess and evaluate the performance of each of the components of Verbmobil, evaluations are conducted on a yearly basis. The evaluations are run by the University of Braunschweig, and the test data is chosen independently by LMU in Munich. Evaluation rules asked one mandatory test on exactly the same conditions from every participant, a suite of other test conditions were optional on a voluntary basis. In the 1996 mandatory test, the language model was constrained to a bigram with a test set perplexity of about 54, the training material was restricted to the official Verbmobil database, and the vocabulary size was 5300 words.

343 utterances (43 minutes of speech) were chosen as test set. Approximately one half of them originated from female speakers. The speakers came from different locations throughout Germany thereby providing a representative mixture of German dialects. The trigram test set perplexity was about 40. All types of spontaneous speech effects like noise, restarts, hesitations, etc. were present in the testing material.

There were four other sites participating in the evaluation in 1996: Daimler-Benz, the universities of Munich, Erlangen, and Hamburg. The evaluation results are given in table 3. Only one optional test (where no limits were imposed to the algorithms and databases used for recognition) was done by more than one institution. Therefore, we only report the results of the mandatory test and the unrestricted optional one.

Site	Error rate (mandatory test)	Error rate (optional test)
Daimler-Benz	21.7%	-
Univ. Erlangen	-	24.5%
Univ. Hamburg	20.0%	-
JANUS	16.1%	13.2%
Univ. Munich	25.2%	-

Table 3. Results of 1996 Verbmobil speech recognition evaluation.

6. CONCLUSION

As can be seen in table 4, steady progress in performance has been achieved in the Verbmobil system during the last 3 years. In this paper we have described several techniques, including improved acoustic modeling and better language models, which were able to reduce the word error rate to less than 50% compared to the 1995 result.

7. ACKNOWLEDGEMENTS

This research was partly funded by grant 413-4001-01IV101S3 from the German Ministry of Science and Technology (BMBF) as a part of the Verbmobil project. The JANUS project was supported in part by the Advanced Research Project Agency and the US Department of Defense. The authors wish to thank all members of the Interactive Systems Labs for useful discussions and active support.

Year	Error Rate
1994	54.2%
1995	30.0%
1996	13.8%

Table 4. Error rates of JANUS over the last three years.

REFERENCES

[1] A. Andreou, T. Kamm, and J. Cohen. Experiments in Vocal Tract Normalization. In *Proceedings of the CAIP Workshop: Frontiers in Speech Recognition II*, 1994.

[2] Ellen Eide and Herbert Gish. A Parametric Approach to Vocal Tract Length Normalization. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Atlanta, 1996. IEEE.

[3] Michael Finke and Ivica Rogina. Wide Context Acoustic Modeling in Read vs. Spontaneous Speech. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Munich, Germany, 1997. IEEE.

[4] Michael Finke, Torsten Zeppenfeld, Martin Maier, Laura Mayfield, Klaus Ries, Puming Zhan, John Lafferty, and Alex Waibel. Switchboard April 1996 Evaluation Report. In *Proceedings of LVCSR Hub 5 Workshop*, April 1996.

[5] Petra Geutner. Introducing Linguistic Constraints into Statistical Language Modeling. In *Proceedings of the 1996 International Conference on Spoken Language Processing (ICSLP)*, Philadelphia, Pennsylvania, pages 402–405, October 1996.

[6] Reinhard Kneser and Herman Ney. Improved Clustering Techniques for Class-Based Statistical Language Modeling. In *Eurospeech*, Berlin, Germany, 1993.

[7] Li Lee and Richard C. Rose. Speaker Normalization using Efficient Frequency Warping Procedures. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 353–356, Atlanta, 1996. IEEE.

[8] Klaus Ries, Finn Dag Buø, and Alex Waibel. Class phrase models for language modelling. In *International Conference on Spoken Language Processing*, Philadelphia, USA, 1996.

[9] Steven Wegmann, Don McAllaster, Jeremy Orloff, and Barbara Peskin. Speaker Normalization on Conversational Telephone Speech. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 339–341, Atlanta, 1996. IEEE.

[10] Torsten Zeppenfeld, Michael Finke, Klaus Ries, Martin Westphal, and Alex Waibel. Recognition of Conversational Telephone Speech using the JANUS Speech Engine. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Munich, Germany, 1997. IEEE.