THE RWTH LARGE VOCABULARY SPEECH RECOGNITION SYSTEM FOR SPONTANEOUS SPEECH

Stephan Kanthak, Sirko Molau, Achim Sixtus, Ralf Schlüter, and Hermann Ney
Lehrstuhl für Informatik VI, Department of Computer Science, RWTH Aachen, University of Technology,
52056 Aachen, Germany

ABSTRACT

This paper presents details of the RWTH large vocabulary continuous speech recognition system used in the VERBMOBIL spontaneous speech translation system. In particular, we report on methods for accelerating the search and algorithms for fast vocal tract normalization (VTN). We focus both on the improvements in word error rate and how to speed up the recognizer with only minimal loss in recognition accuracy. Implementation details and experimental results are given for the VERBMOBIL German development corpus dev99. The 24.6% word error rate of the baseline system is reduced to 22.8% using VTN. Decreasing the real-time factor by a factor of 5 resulted in only a small degradation in recognition performance of 2% relative on average. Furthermore, we study incremental methods for reducing the response time of the online speech recognizer and an efficient method to reduce the density of word graphs.

1. Introduction

This paper describes the RWTH large vocabulary continuous speech recognition (LVCSR) system used in the research demonstrator of the VERBMOBIL spontaneous speech-to-speech translation system [1]. The RWTH LVCSR system is based on a Viterbi search using linear discriminant analysis, continuous Gaussian mixture densities, a globally pooled variance vector, 3-state left-to-right Hidden Markov models, phonetically tied within-word triphone models, and a trigram language model [4].

The goal of the VERBMOBIL project is to develop a four times real-time speaker-independent speech-to-speech translation system for the domains scheduling appointments and hotel/travel reservations. Since the system consists of many different modules running simultaneously, the speech recognition needs to run in real-time. In order to counteract the problems of speaker-independent recognition, we also study speaker-adaptation techniques based on vocal tract normalization. In this paper we report recognition results and real-time factors on the VERBMOBIL German spontaneous speech development corpus dev99.

In order to reduce the overall response time, some VERBMOBIL translation modules are designed to process partial input data. Therefore, we study methods for reducing the response time of the recognizer and to incrementally report recognition results. Finally, we describe an efficient compression technique for word graphs used to keep the computational effort for the subsequent VERBMOBIL modules as low as possible.

2. Efficient Real-Time Recognition

Although the RWTH speech recognition system is capable of performing integrated across-word model search [10], in this paper we only apply within-word triphone models for speed purposes. In order to achieve real-time recognition, we use the following three acceleration techniques: language model look-ahead, phoneme look-ahead, and fast likelihood calculation of continuous mixture densities.

2.1. Language Model Look-Ahead

For more efficient hypothesis pruning during beam search, we distribute the language model probabilities over the tree using language model look-ahead [5]. We enhance this look-ahead by incorporating the language model score of the most probable successor word into the accumulated score *before* starting a new tree. This makes language model pruning more efficient.

2.2. Phoneme Look-Ahead

The idea of phoneme look-ahead is to anticipate the probability of future acoustic vectors for a phoneme arc before starting it in detailed search. To estimate the acoustic probabilities, we use simplified context independent models and perform an additional pruning step before starting the phoneme arcs. This method is described in detail in [6] and results in an overall speedup by a factor of about 2.

2.3. Fast Likelihood Calculation

Without acceleration, the calculations of likelihoods of the continuous mixture densities make up more than 80% of the time required for the whole recognition process. We obtain considerable improvements in speed by quantizing the components of the acoustic means and input vectors in order to calculate the vector distances using parallelizing SIMD (single instruction multiple data) instructions (e.g. MMX on Intel Pentium, VIS on Sun Sparc). As reported in [2], this leads to an overall speedup of the recognizer by more than a factor of 3 without any loss in recognition performance. Additionally, we accelerate the likelihood calculations by using projection search and the preselection VQ method described in [8].

Using all these methods, we achieve an overall speedup by a factor 10 at the cost of only a minor increase in word error rate.

3. Vocal Tract Normalization (VTN)

The idea of VTN is to remove speaker dependencies caused by variations in the length of the vocal tract by scaling the frequency axis. In this section, we shortly describe the principles of VTN. Then, we discuss a text-independent Gaussian mixture based approach for fast warping factor estimation (*fast VTN*). Finally, we propose a method for incremental estimation of warping factors.

3.1. VTN Principles

The baseline VTN algorithm has been described in detail in [12]. In speaker-adaptive training, the warping factor $\hat{\alpha_i}$ for each training speaker i is obtained by exhaustive search in the range $0.88, \ldots, 1.12$ with step size 0.02:

$$\hat{\alpha_i} = \arg\max_{\alpha} Pr(X_i^\alpha | \mu, W_i).$$

In this equation, the HMM emission probability $Pr(X_i^\alpha | \mu, W_i)$ is computed by a Viterbi alignment of the warped acoustic vectors X_i^α for the HMM parameter set μ and the spoken word sequence W_i.

So-called normalized references are obtained by training the parameters of the emission distributions on warped acoustic vectors, i.e. the vectors are normalized using the optimal speaker dependent warping factors [12].

In recognition a similar procedure is applied. Since the spoken utterance W_i is unknown, a preliminary transcription $\hat{W_i}$ is obtained in a first recognition pass with no vocal tract normalization. Then the factor $\hat{\alpha_i}$ is determined which maximizes the emission probability using $\hat{W_i}$. Finally, the acoustic vectors are warped using $\hat{\alpha_i}$, and a second recognition pass is performed using the normalized references. Recognition tests have shown that a significant performance improvement is obtained by this two-pass method. However, the computation time is almost doubled. Using $\hat{W_i}$ instead of the actually spoken, but unknown transcription W_i does not degrade the recognition performance even if the preliminary transcription has a large word error rate in the order of 20 to 30%.

3.2. Fast VTN

For real-time performance, a two-pass recognition approach is virtually prohibitive. To estimate the warping factor of the test speaker without a preliminary recognition pass, [3] and [12] suggested a text-independent method using Gaussian mixture models. The approach used here relies on a separate emission distribution $Pr(X|\lambda_\alpha)$ for *each* warping factor α, where X denotes the sequence of acoustic vectors and λ_α denotes the α-dependent distribution parameters. For each α, the parameters λ_α are trained only on those *un*normalized acoustic vectors which are assigned to this warping factor. Since the training data is distributed over different models, only a fraction of the corpus is available for each model. To increase the robustness we tie the variance over all Gaussian mixture models.

In order to determine the optimal $\hat{\alpha}$ for an observed sequence X of acoustic vectors during recognition, we use the following equation (see Figure 1):

$$\hat{\alpha} = \arg\max_{\alpha} Pr(X|\lambda_\alpha).$$

Using this approach, the vocal tract normalization is significantly accelerated. The signal analysis has to be carried out twice (unwarped and warped with $\hat{\alpha}$), and there is no need for two recognition passes. As can be seen in Section 4, fast VTN performs almost as well as the baseline two-pass VTN, but is twice as fast.

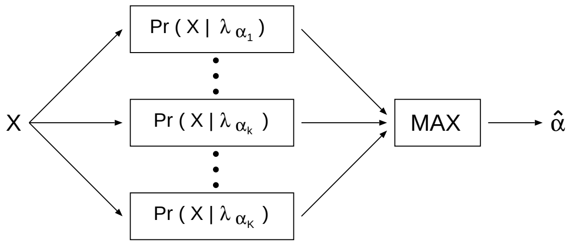


Figure 1: Mixture based warping factor approach using unwarped acoustic vectors X and one Gaussian mixture model λ_α for each warping factor.

3.3. Incremental Estimation of Warping Factors

To reduce the time delay introduced by VTN, we investigate a frame incremental warping factor estimation scheme for fast VTN. As shown in Figure 2, the signal analysis following FFT is always carried out twice. The acoustic vector warped with the currently best warping factor is used for recognition. The unwarped acoustic vector is immediately evaluated with the text-independent Gaussian mixture models and the scores $Pr(X|\lambda_\alpha), \alpha = 0.88, \ldots, 1.12$, are computed by accumulation over time. As long as not enough time frames have been collected (200 time frames, i.e. 2 seconds of speech), the warping factor is set to 1.0. After a few seconds, the best warping factor does not change anymore.

Tests have shown that incremental warping factor estimation can be carried out without any delay at the cost of only little performance degradation in the first few seconds of a new speakers' utterance.

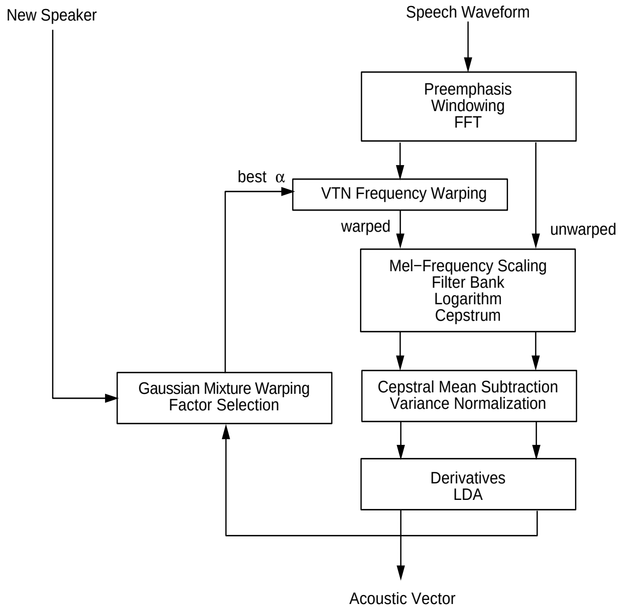


Figure 2: Signal analysis with fast VTN and incremental warping factor estimation without time delay.

4. Experimental Results

4.1. Test Condition

All experiments presented in this section are performed on the speaker independent VERBMOBIL German spontaneous speech task. The statistics of the training and testing set are summarized in Table 1. The baseline recognition system is characterized by:

- 16 cepstral coefficients with first derivatives and the second derivative of the energy, 10 ms frame shift;
- linear discriminant analysis on three adjacent vectors resulting in a 33-dimensional acoustic vector;
- 3-state HMM triphone models with skip, forward, and loop transitions;
- decision tree with 2,501 tied states including articulatory and non-articulatory noise and one silence state;
- gender independent Gaussian mixtures with a total of 170k to 240k densities and a globally pooled diagonal covariance matrix;
- recognition vocabulary of 10,810 words with additional 136 pronunciation variants in the pronunciation lexicon;
- trigram language model with a testing set perplexity of 62.0.

Table 1: Training (CD 1-41) and testing set (development set of the third integration 1999).

	training	testing
acoustic data	61.5h	1.6h
silence portion	13%	11%
# speakers	857	16
# sentences	36,015	1,081
# running words	701,512	14,662
# running phonemes	2,331,927	60,716
perplexity (trigram LM)	-	62.0

4.2. Recognition Experiments

In the following, we compare the search space (active states, arcs, trees after pruning), word error rate (WER) and real-time factor (RTF) of three systems on a 600 MHz Pentium III PC:

- recognition without speaker adaptation (baseline)
- recognition with fast VTN (fVTN)
- recognition with two-pass VTN (2VTN)

Note that for the unaccelerated recognition experiments we already use SIMD instructions to accelerate the distance calculation and the enhanced language model look-ahead as described in Section 2. Both methods do not affect the recognition accuracy. Therefore, they are usually used for efficient recognition even during development.

All recognition results are shown in Table 2. The word error rate can be reduced from 24.6% to 22.8% (7.3% relative) using two-pass VTN while increasing the RTF by a factor of 1.8. Using fast VTN the word error rate can be reduced from 24.6% to 23.0% (6.5% relative) while decreasing the RTF from 6.7 to 5.7. This proves that VTN allows for more efficient pruning.

As can be seen in the second part of Table 2, we accelerate the systems to almost real-time using all the methods described in Section 2. The acceleration parameters are chosen in such a way that the word error rate increases by at most 0.5% absolute. With this constraint the RTF of the baseline system can be reduced from 6.7 to 1.3, and the RTF of the fast VTN system from 5.7 to 1.2.

Table 2: Recognition results.

system	search space			word errors [%]		
	states	arcs	trees	del - ins	WER	RTF
baseline	4173	2202	55	5.3 - 3.8	24.6	6.7
+ 2VTN	3670	1785	48	4.8 - 3.5	22.8	12.1
+ fVTN	3710	1806	49	4.8 - 3.6	23.0	5.7
accelerated	1839	827	19	5.8 - 3.9	25.1	1.3
+ fVTN	1771	767	17	5.5 - 3.5	23.5	1.2

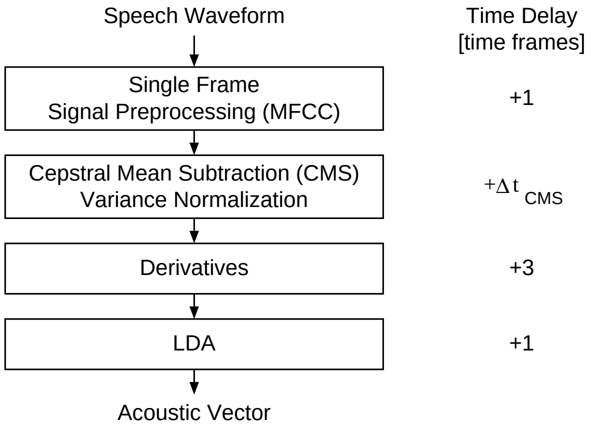
5. Incremental Processing for Online Recognition

In this section, we focus on integrating the recognizer into the online environment with subsequent processing steps of the VERBMOBIL system. The time delay introduced by the recognizer should be as small as possible since all postprocessing steps depend on the recognition output. To this purpose, we incorporate the following concepts:

- reducing the time delay of the recognizer by performing cepstral mean subtraction not on the whole sentence but on a running window with constant length;
- reducing the response time of the system by incrementally preparing recognition results.

The constraint which all these concepts should satisfy is not to increase the word error rate.

5.1. Recognition Preprocessing



Speech Waveform
Time Delay [time frames]
Single Frame Signal Preprocessing (MFCC)
+1
Cepstral Mean Subtraction (CMS) Variance Normalization
+Δt CMS
Derivatives
+3
LDA
+1
Acoustic Vector

Figure 3: Signal processing steps and associated time delays.

The time delay introduced by the signal analysis and feature extraction of the recognizer is shown in detail in Figure 3. We consider two types of cepstral mean subtraction: the cepstral mean can be either computed on the whole sentence or on a running window. As we will see, the minimum window length Δt_{CMS} of the cepstral mean subtraction is of the order of several hundred 10-ms time frames. Therefore, the cepstral mean subtraction contributes most to the overall time delay.

In several experiments we found that for the German VERBMOBIL recognition task a symmetric running window of at least 2-seconds length results in exactly the same word error rate as sentence-wise normalization while reducing the time delay to 1 second (Figure 4 top). By using an asymmetric window of the same length, the delay can be further reduced to 250 ms (Figure 4 bottom). This is possible, since the time delay depends only on the future portion of the window.

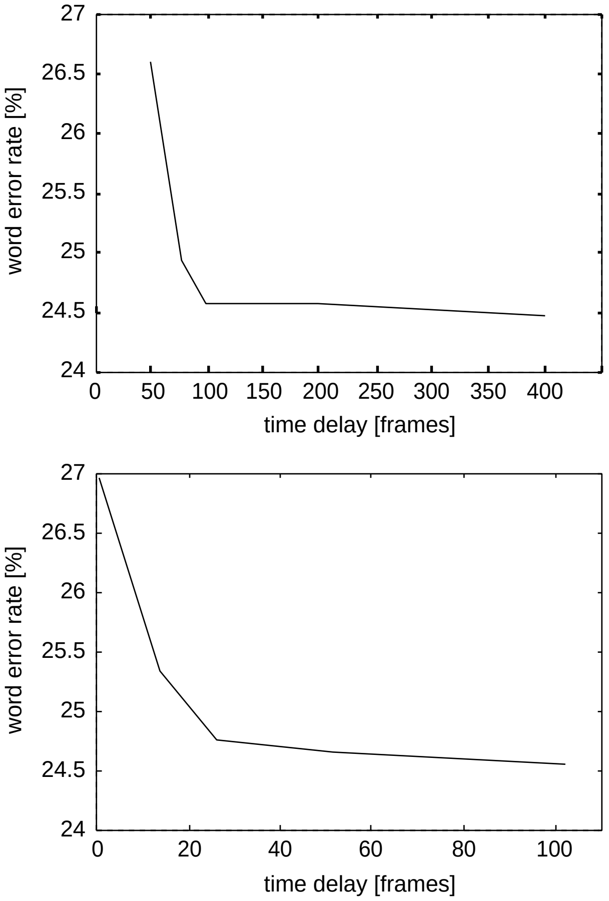


Figure 4: Word error rate as a function of time delay Δt_{CMS} for cepstral mean subtraction using symmetric (top) and asymmetric (bottom) windows. In the symmetric case, the time delay is half of the window length. In the asymmetric case, the window length was kept fixed at 200 time frames.

5.2. Incremental Recognition Output

The output of the two main recognition results, the single best word sequence and the word graph, is usually delayed until the end of the utterance. Nevertheless, subsequent modules could already start working if there were partial results available from the recognizer. The approach to output partial first best word sequences and word graphs is based on search space decisions and was inspired by [11].

The idea is to trace back the currently active state hypotheses in order to determine a single word end node shared by all paths. Once we have found such a node we can already report the first best word sequence up to this node because in the future it will not change anymore.

This partial traceback is performed only after a sufficiently long silence segment has been recognized.

Since the paths of the currently active hypotheses share a single node in the resulting word graph, too, partial word graphs can be generated in a similar fashion. Using this method it is guaranteed that the single best word sequence is still in the word graph.

6. Forward-Backward Word Graph Pruning

To keep the computational effort within the subsequent processing steps of the VERBMOBIL system as low as possible, it is desirable to generate compact word graphs with low word graph density (the total number of word graph arcs divided by the number of spoken words) and low graph error rate (the word error rate of the sentence hypothesis matching best the spoken sentence).

For generating compact word graphs we apply the *forward-backward pruning* described in detail in [9]. This pruning technique is based on the score of the global best path through the entire word graph. For each arc of the graph representing a word hypothesis $(w;\tau,t)$, with word identity w, starting time $\tau+1$ and ending time t, we compute the overall score $Q(w;\tau,t)$ of the best path passing through the arc. Only arcs with scores close to the global best path are kept.

This pruning method is based on the paradigm of the forward-backward algorithm since the following two quantities are required to compute $Q(w;\tau,t)$:

1. The forward score, which is defined as the overall score of the best partial word sequence that starts at the first time frame of the utterance and ends at time t in word arc $(w;\tau,t)$;

2. The backward score, which is defined as the overall score of best partial word sequence that starts at the time frame $\tau+1$ in word arc $(w;\tau,t)$ and ends at the last time frame of the utterance.

These two quantities are computed by traversing the word graph from left to right (forward pass) and from right to left (backward pass).

As can be seen in Figure 5, at a given GER forward-backward pruning leads to significant smaller word graphs than the conventional forward pruning algorithm, where only the forward score is used as pruning criterion [8]. Additionally, when using forward pruning, the first best word sequence might be pruned from the word graph while it is guaranteed to remain in the word graph when using forward-backward pruning. That can be observed in Figure 5: The WER of the first best word sequence of the unpruned word graph is 23%. For very low values of the WGD the value of the GER increases beyond this value in the case of forward pruning, while in the case of forward-backward pruning it remains the same.

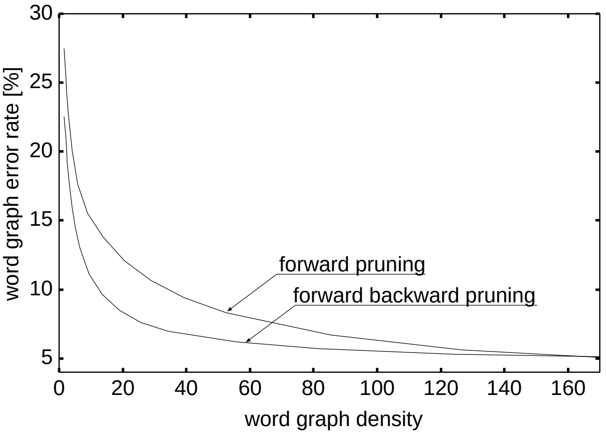


Figure 5: Comparison of forward- and forward-backward word graph pruning. The WER of the first best word sequence was 23.0%.

7. Summary

We have presented acceleration methods that result in a drastic speedup of the RWTH speech recognizer used in the VERBMOBIL speech-to-speech translation system. The increase in word error rate was only small. We have described the vocal tract normalization technique and proposed a fast text-independent method for warping factor estimation which can be used in real-time applications. Furthermore, we have presented incremental methods for reducing the response time of the recognizer. Finally, we have described the forward-backward word graph pruning technique which produces compact word graphs.

Acknowledgement. This research was partly funded by grant 01 IV 701 T4 from the German Ministry of Science and Technology (BMBF) as a part of the VERBMOBIL project. The views and conclusions contained in this document are those of the authors.

8. References

[1] T. Bub, J. Schwinn: VERBMOBIL: The Evolution of a Complex Large Speech-to-Speech Translation System, Proc. of the International Conference on Spoken Language Processing, pp. 2371-2374, Philadelphia, PA, October 1996.

[2] S. Kanthak, K. Schütz, H. Ney: Using SIMD Instructions for Fast Likelihood Calculation in LVCSR, Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 1531-1534, Istanbul, Turkey, June 2000.

[3] L. Lee, R. Rose: Speaker Normalization using Efficient Frequency Warping Procedures, Proc. of the IEEE International Conference on Acoustics,

Speech and Signal Processing, pp. 353-356, Atlanta, GA, May 1996.

[4] H. Ney, L. Welling, S. Ortmanns, K. Beulen, F. Wessel: The RWTH Large Vocabulary Continuous Speech Recognition System, Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 853-856, Seattle, WA, May 1998.

[5] S. Ortmanns, H. Ney, A. Eiden: Language-Model Look-Ahead for Large Vocabulary Speech Recognition, Proc. of the International Conference of Spoken Language Processing, pp. 2091-2094, Philadelphia, PA, October 1996.

[6] S. Ortmanns, H. Ney, A. Eiden, N. Coenen: Look-Ahead Techniques for Improved Beam Search, Proc. of the CRIM-FORWISS Workshop, pp. 10-22, Montreal, October 1996.

[7] S. Ortmanns, H. Ney, T. Firzlaff: Fast Likelihood Computation Methods for Continuous Mixture Densities in Large Vocabulary Speech Recognition, Proc. of the European Conference on Speech Communication and Technology, pp. 139-142, Rhodos, Greece, September 1997.

[8] S. Ortmanns, H. Ney, X. Aubert: A Word Graph Algorithm for Large Vocabulary Continuous Speech Recognition, Computer, Speech and Language, No. 1, pp. 43-72, January 1997.

[9] A. Sixtus, S. Ortmanns: High Quality Word Graphs Using Forward-Backward Pruning, Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing, Phoenix, AZ, pp. 593-596, March 1999.

[10] A. Sixtus, S. Molau, S. Kanthak, R. Schlüter, H. Ney: Recent Improvements of the RWTH Large Vocabulary Speech Recognition System on Spontaneous Speech, Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 1671-1674, Istanbul, Turkey, June 2000.

[11] J.C. Spohrer, P.F.Brown, P.H. Hochschild, J.K. Baker: Partial Traceback in Continuous Speech Recognition, Proc. of the International Conference on Cybernetics and Society, pp. 36-42, Cambridge, MA, October 1980.

[12] L. Welling, S. Kanthak, H. Ney: Improved Methods for Vocal Tract Normalization, Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 761-764, Phoenix, AZ, March 1999.