THE RWTH LARGE VOCABULARY CONTINUOUS SPEECH RECOGNITION SYSTEM

H. Ney, L. Welling, S. Ortmanns, K. Beulen, F. Wessel

Lehrstuhl für Informatik VI, RWTH Aachen – University of Technology, D-52056 Aachen, Germany

ABSTRACT

In this paper, we present an overview of the RWTH Aachen large vocabulary continuous speech recognizer. The recognizer is based on continuous density hidden Markov models and a time-synchronous left-to-right beam search strategy. Experimental results on the ARPA Wall Street Journal (WSJ) corpus verify the effects of several system components, namely linear discriminant analysis, vocal tract normalization, pronunciation lexicon and cross-word triphones, on the recognition performance.

1. INTRODUCTION

This paper describes the large vocabulary continuous speech recognizer developed at RWTH Aachen. The recognizer employs a cepstrum front-end that includes linear discriminant analysis (LDA) and vocal tract normalization (VTN). For acoustic modelling, continuous density hidden Markov models (CDHMM) and decision-tree based state tying are used. A time-synchronous left-to-right beam search strategy in combination with a tree-organized pronunciation lexicon is used for decoding. Word graphs are generated using the word pair approximation. Methods for fast evaluation of emission probabilities provide a significant reduction of the computational effort for distance calculations. Variants of this baseline system are currently in use in several projects in which our group is involved, e.g. Verbmobil [4] and Arise (EU project LE3-4229).

In this paper, we will give an overview of the system and report on the effect of several system components on the recognition performance. The paper is organized as follows. Section 2 describes the acoustic front-end of the recognizer. Section 3 is on the acoustic modelling. The search procedure is described in Section 4. Experimental results on the ARPA Wall Street Journal database are reported in Section 5. A summary is given in Section 6.

2. ACOUSTIC FRONT-END

In the following, we describe the acoustic front-end for speech sampled at 16 kHz [14]. Every 10 ms, a Hamming window is applied to preemphasised 25-ms segments and a 1024-point fast Fourier transform is performed. The magnitude spectrum is warped according to the mel scale. The obtained spectral magnitudes are integrated within 20 triangular filters arranged on the mel-frequency scale. The mid-frequency of filter n is $n/2 \cdot 270.48$ and the bandwidth is 270.48 for all filters. The filter output is the logarithm of the sum of the weighted spectral magnitudes. A decorrelation by a discrete cosine transform is performed. 16 mel-frequency cepstral coefficients (MFCC) are computed from 20 filter outputs. Augmenting the 16 cepstrum coefficients by 16 first-order linear regression coefficients and 1 second-order coefficient gives a vector with a dimension of 33. The linear regression coefficients are calculated over a window covering 5 neighboring cepstrum vectors. To suppress channel distortions, a mean normalisation is carried out.

Finally, a LDA [5] with classes defined as states is applied. 3 successive 33-dimensional vectors from times $t-1$, t and $t+1$ are adjoined to form a large input vector with a dimension of 99. A gender-independent transformation matrix reduces the dimension of this vector from 99 to 33.

The system also contains a component for vocal tract normalization (VTN) [6]. The training procedure using VTN is as follows. Intermediate models with a small number of densities per state are estimated from the unwarped features of all training speakers by maximum likelihood (ML) training. For each training speaker, a warp scale is chosen as the scale for which the training data of this speaker achieve the greatest likelihood, given the transcriptions and the intermediate models. The final acoustic models to be used for recognition are trained on the warped utterances by ML training. In recognition, we use a preliminary transcription of the test sentence for the warp scale selection.

3. ACOUSTIC MODELLING

Triphone acoustic models are represented by continuous density hidden Markov models. Each model contains 3 segments with 2 states each. Forward, skip and loop transitions between the states are allowed. The states of a segment share the same emission probability distribution, which are also shared among the segments of different triphones by using state tying, as will be explained later. The emission probabilities are represented by Gaussian or Laplacian mixture densities. Covariances are modeled by a single diagonal matrix pooled over all mixtures. The number of component densities per mixture differs between the states and is automatically adjusted during the training phase, as will be described later. The phoneme inventory consists of 44 phonemes including one phoneme for silence with a single state.

3.1. Maximum Likelihood Training

The parameters of the emission probabilities are trained using the maximum likelihood criterion together with several approximations: Only the best state sequence is used (Viterbi approximation). For the calculation of the emission probabilities during time alignment, the sum over all component densities of a mixture is approximated by the maximum. For parameter estimation, each observation vector is assigned to the density of a mixture which gives the highest probability. The transition probabilities are not trained but set to a constant value which depends only on the type of the transition.

Each iteration of the training procedure consists of time alignment by dynamic programming followed by parameter estimation. In order to increase the acoustic resolution, a splitting step is carried out typically after every 6 iterations. If for a specific density the average log-likelihood score over all observations assigned to this density is larger than the average log-likelihood score over all densities, this density is split by replacing the mean vector by two small disturbed versions of it. As a result, the number of densities of a mixture depends on the actual distribution of the observation vectors assigned to a mixture.

This training procedure is sped up by two refinements: to cut down computation time for time alignment, a sort of beam search with a data-adaptive adjustment window is used. As a result, only about 15 states per time frame are evaluated. Also, a time alignment is performed typically only every 3 iterations, since time alignment paths from previous iterations can be used.

For starting from scratch, the training procedure can be initialized by a linear segmentation. Each training utterance is automatically segmented into 3 parts [3] which are silence at the beginning of the utterance, speech and silence at the end of the utterance. Then the frames belonging to the speech segment are linearly assigned to the corresponding states and the frames of the silence segments are assigned to the silence model.

3.2. State Tying

To calculate an appropriate state tying for the different corpora, simple Gaussian distributions are estimated for every triphone state using a precalculated segmentation of the acoustic data. These models are then used to calculate the tying. For the tying the system is able to use two different methods as described in [2, 16, 17]. The baseline method is decision tree based and works in a top-down fashion. It starts with all triphone states with the same central phoneme and the state number in one cluster and then splits these clusters using phonetic questions on the central phoneme and the context of the triphone states. The result is the desired number of state clusters and phonetic decision trees which assign the triphone states to their clusters. Unseen triphones are assigned to appropriate states using the phonetic decision trees so no additional backing-off models are needed. The second method is purely data-driven. The triphone states are sorted due to the central phoneme of the triphone and to the state number. Then the states are clustered together using a bottom-up-strategy until the desired number of states is reached. These states are then re-estimated with a higher acoustic resolution. For unseen triphones additional monophone models are trained which are used for backing-off. The advantage of this method is that no phonetic questions are needed e.g. in case of switching to a new corpus.

3.3. Cross-Word Models

A recent improvement to the system is the use of cross-word models [9]. Cross-word models enhance the acoustic modelling of a word boundary by fully incorporating the knowledge of the phonemes and the degree of coarticulation at this boundary. The implementation of cross-word models for our system is rather straightforward. During each training iteration, the lengths of the between word silences are estimated and then the appropiate triphones for the word boundaries are used according to a silence length threshold. No specific silence models or the like are used.

For recognition, we use a three-pass strategy. In the first pass, a word lattice is constructed using only within-word models. Using this lattice, in the second pass the n best sentences are calculated and in the third pass these n best sentences are rescored using the cross-word models. The between-word silence information of the first recognition pass is used to determine which triphone models have to be used at word boundaries.

3.4. Pronunciation Variants

Another important new feature of the system is the incorporation of pronunciation variants. These variants are manually transcribed by an expert listening to the respective training corpus. During training, an optional variant recognition pass is able to detect which variant was spoken in the training corpus. This variant recognizer uses simple monophone models with only 2-4 densities per mixture to keep the acoustic models from learning the pronunciation variants. For every sentence in the training corpus, it determines a pronunciation variant for the words in the sentence. This information can then be used to reestimate the monophone models of the variant recognizer which also improves the variant recognition and so on. After 2-3 iterations this procedure converges. Then a conventional acoustic training is performed which regards the precalculated information about the pronunciation variants.

During recognition, all variants are added to the tree lexicon, a mapping function identifies the pronunciation variant with the appropiate lexical word in the language model. No weighting of the variants of a word is used.

4. THE SEARCH METHOD

In this section, we describe the main characteristics of the baseline search approach used in the RWTH speech recognizer. The search method is based on a strictly time-

synchronous left-to-right beam search strategy connected with a tree-organized pronunciation lexicon (lexical prefix tree) [7]. The incorporation of a bigram or a more complex language model requires copies of the lexical prefix tree since the identity of a word is only known at a leaf of the tree. Thus, in the standard system we use word-conditioned copies of the lexical prefix tree, e.g. for a bigram language model the copies depend on the immediate predecessor word.

4.1. Standard Pruning Approach

The standard pruning approach which is used in the so-called word-conditioned lexical tree search method consists of three pruning steps: acoustic pruning, language model pruning and histogram pruning. Acoustic pruning (standard beam search) eliminates all state hypotheses with a score relatively worse to the best active state. Language model pruning is only applied to tree start-up hypotheses and works in a similar way as acoustic pruning. Histogram pruning confines the number of active states to a maximum number [13]. To further improve the efficiency of the pruning process look-ahead pruning techniques are used, e.g. language model look-ahead and phoneme look-ahead. In all experiments presented in this paper, we applied only a bigram language model look-ahead. This bigram language model look-ahead pruning is based on an on-demand computation of the factored bigram probabilities using a compressed language model look-ahead tree [11].

4.2. Fast Likelihood Calculation Method

A fast likelihood calculation method is used to reduce the computational effort of the mixture densities calculations in the system [12]. The method combines the preselection VQ method with the so-called projection search algorithm. This methods works as follows: In a first step, the VQ method is used to get a coarse preselection of the prototype vectors of the densities. Then, the selected prototype vectors can by further confined by considering only vectors which are located inside a 'hypercube' centered at a given acoustic observation vector. These vectors are then evaluated in the log-likelihood computation procedure. By this fast likelihood calculation, the real time factor of the recognition system described (20k vocabulary) is reduced from 10 to 2.5 on an ALPHA 5000 PC (SpecInt'95: 15.4).

4.3. Word Graph Generation

The concept of the word conditioned tree search method for determining the single best sentence can be extended to produce high quality word graphs [1, 10]. The advantage of a word graph is that the computationally expensive acoustic recognition task can be decoupled from the application of a complex language model in a subsequent post-processing step or for computing the n-best sentence hypotheses. The generation of word graphs is based on the word pair approximation which fits directly into the word conditioned tree search method using a bigram language model. In addition, pruning is used to reduce the number of word arcs in the word graph. This so-called word graph pruning works in a straight-forward beam search strategy.

Table 1: Effect of LDA on the word error rates for Gaussian and Laplacian densities (WSJ0 5k Nov.'92 dev/eval test sets: 10/8 speakers, 410/330 sentences, 6779/5353 spoken words; bigram language model with a perplexity of 107; gender-dependent models; deletions (del), insertions (ins) and word error rate (WER) in %).

Models	LDA	#dens (m+f).	del – ins	WER
Gaussians	no	68k+70k	1.5 – 0.7	7.6
	yes	68k+82k	1.3 – 0.8	6.9
Laplacians	no	61k+58k	1.4 – 0.8	8.0
	yes	75k+86k	1.3 – 0.7	7.1

Table 2: Effect of VTN (WSJ0 5k 92 dev/eval, bigram, gender-independent Gaussians, LDA).

VTN	#dens	del – ins	WER
no	122k	1.4 – 0.7	7.0
yes	140k	1.2 – 0.6	6.1

4.4. Language Modelling

Our language models are based on the frequency of unigrams, bigrams and trigrams in a training corpus. The event counts are efficiently stored [15] and are used in a multi-level smoothing approach, using 'absolute discounting' with an interpolation between the different language model levels [8]. All language model parameters are estimated using Leaving-One-Out.

5. EXPERIMENTAL RESULTS

We will now present the results of a series of experiments that were conducted to optimize the recognition performance of our system. All experiments were carried out on the ARPA Wall Street Journal (WSJ) corpus. Training was done on the WSJ0 84-speaker corpus and testing on the WSJ0 Nov. '92 development and evaluation test sets. The recognition lexicon contained 4986 words, for the tests using pronunciation variants 668 variants were added. For the state tying the number of 23509 triphone states was reduced to 2001 (including silence) by the decision tree based method using 88 phonetic questions. Recognition was done with a bigram language model with a perplexity of 107 on the development and evaluation data.

A characteristic of the acoustic front-end is the combination of LDA and cepstrum. We tested the performance of this combination for both Gaussian and Laplacian densities. Table 1 summarizes the results. As can be seen from the table, LDA reduces the word error rate by approximately 10% for both Gaussian and Laplacian density functions.

Results for VTN applied in training and recognition are given in Table 2. In this experiment, we used GI models since VTN should discard gender-specific variations from the training data and beneficially exploit the larger training database. The table shows a reduction in the word error rate from 7.0% to 6.1% due to VTN.

Table 3: Effect of cross-word triphones (WSJ0 5k 92 dev/eval, bigram, gender-dependent Laplacians, LDA).

cross-word triphones	#dens (m+f).	del – ins	WER
no	86k+75k	1.3 – 0.7	7.1
yes	73k+66k	1.0 – 0.8	6.4

Table 4: Effect of pronunciation lexicon (WSJ0 5k 92 dev/eval, bigram, gender-dependent Laplacians, LDA).

lexicon	pro. var.	#dens (m+f).	del – ins	WER
baseline	no	86k+75k	1.3 – 0.7	7.1
improved	”	92k+74k	1.4 – 0.6	6.9
”	rec	92k+74k	1.4 – 0.5	6.5

The results for cross-word triphones are shown in Table 3. Here the initial number of triphone states was 51118 due to the additional cross-word triphones. These were also reduced to 2001 tied states by the decision tree method. The error rate of the baseline method is reduced by 10% relative. Table 4 contains the results for different recognition lexicons. Line “baseline” shows the results for our baseline lexicon. For the “improved” lexicon several corrections of the word transcriptions were made. These improvements gave us a reduction of about 3% in word error rate. An additional improvement of about 6% was achieved by adding 688 pronunciation variants, about 10% of the overall lexical size. These pronunciation variants were used only during recognition, for the training we employed only the canonical pronunciations of the words. However, we found no further improvement on the WSJ0 5k task by using pronunciation variants also in training

6. SUMMARY

In this paper, we have described the RWTH Aachen large vocabulary continuous speech recognizer. The system is based on continuous density hidden Markov models and a time-synchronous left-to-right beam search strategy. It employs state-of-the-art techniques such as LDA, decision tree based state-tying, speaker normalization by VTN, cross-word models, pronunciation variants, fast likelihood calculation and word graph rescoring. Experiments on the ARPA WSJ corpus showed significant performance improvements due to speaker normalization and cross-word models.

7. REFERENCES

[1] X. Aubert, H. Ney, “Large Vocabulary Continuous Speech Recognition using Word Graphs,“ *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 49-52, Detroit, MI, May 1995.

[2] K. Beulen, E. Bransch, H. Ney, “State Tying for Context Dependent Phoneme Models,” *Proc. Fifth European Conference on Speech Communication and Technology*, pp. 1179-1182, Rhodes, Greece, September 1997.

[3] J. S. Bridle, N. C. Sedgewick, “A method for segmenting acoustic patterns, with applications to automatic speech recognition,” *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 656-659, Hartford, CN, May 1977.

[4] T. Bub, W. Wahlster, A. Waibel, “Verbmobil: The Combination of Deep and Shallow Processing for Spontaneous Speech Translation,” *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 71-74, Munich, Germany, April 1997.

[5] M. J. Hunt, C. Lefèbvre, “A comparison of several acoustic representations for speech recognition with degraded and undegraded speech,” *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 2.2-2.2, Glasgow, Great Britain, May 1989.

[6] L. Lee, R. Rose, “Speaker normalization using efficient frequency warping procedures,” *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 353-356, Atlanta, GA, May 1996.

[7] H. Ney, R. Haeb-Umbach, B.-H. Tran, M. Oerder, “Improvements in Beam Search for 10000-Word Continuous Speech Recognition,” *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 13-16, San Francisco, CA, March 1992.

[8] H. Ney, S. Martin, F. Wessel, “Statistical Language Modeling Using Leaving-One-Out,” in *Corpus-Based Methods in Language and Speech Processing*, S. Young, G. Bloothooft (eds.), pp. 174-207, Kluwer Academic Publishers, The Netherlands, 1997.

[9] J. J. Odell, “The Use of Context in Large Vocabulary Speech Recognition,” Ph.D. Thesis, Cambridge University, Cambridge, March 1995.

[10] S. Ortmanns, H. Ney, X. Aubert, “A Word Graph Algorithm for Large Vocabulary Continuous Speech Recognition,” *Computer, Speech and Language*, Vol. 11, No. 1, pp. 43-72, January 1997.

[11] S. Ortmanns, A. Eiden, H. Ney, N. Coenen, “Look-Ahead Techniques for Fast Beam Search,” *Proc. IEEE Int. onf. on Acoustics, Speech and Signal Processing*, pp. 1783-1786, Munich, Germany, April 1997.

[12] S. Ortmanns, H. Ney, T. Firzlaff, “Fast Likelihood Computation Methods for Continuous Mixture Densities in Large Vocabulary Speech Recognition,” *Proc. Fifth European Conference on Speech Communication and Technology*, pp. 143-146, Rhodes, Greece, September 1997.

[13] V. Steinbiss, B.-H. Tran, H. Ney, “Improvements in Beam Search,” *Proc. Int. Conf. on Spoken Language Processing*, pp. 2143-2146, Yokohama, Japan, September 1994.

[14] L. Welling, N. Haberland, H. Ney, “Acoustic Front-End Optimization for Large Vocabulary Speech Recognition,” *Proc. Fifth European Conference on Speech Communication and Technology*, pp. 2099-2102, Rhodes, Greece, September 1997.

[15] F. Wessel, S. Ortmanns, H. Ney, “Implementation of Word Based Statistical Language Models,” *Proc. SQEL Workshop on Multi-Lingual Information Retrieval Dialogs*, pp. 55-59, Pilsen, Czech Republic, April 1997.

[16] S. J. Young, P.C. Woodland, “The Use of State Tying in Continuous Speech Recognition,” *Proc. European Conference on Speech Communication and Technology*, pp. 2203-2206, Berlin, Germany, September 1993.

[17] S. J. Young, J. J. Odell, P. C. Woodland, “Tree-Based State Tying for High Accuracy Acoustic Modelling,” *Proc. ARPA Human Language Technology Workshop*, pp. 405-410, Plainsboro, NY, March 1994.