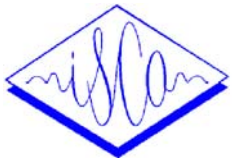

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

UNSUPERVISED TRAINING OF A SPEECH RECOGNIZER: RECENT EXPERIMENTS

Thomas Kemp *Alex Waibel*
Interactive Systems Laboratories, ILKD
University of Karlsruhe
76128 Karlsruhe, Germany

ABSTRACT

Current speech recognition systems require large amounts of transcribed data for parameter estimation. The transcription, however, is tedious and expensive. In this work we describe our experiments which are aimed at training a speech recognizer with only a minimal amount (30 minutes) of transcriptions and a large portion (50 hours) of untranscribed data. A recognizer is bootstrapped on the transcribed part of the data and initial transcripts are generated with it for the remainder (the untranscribed part). Using a lattice-based confidence measure, the recognition errors are (partially) detected and the remainder of the hypotheses is used for training. Using this scheme, the word error rate on a broadcast news speech recognition task dropped from more than 32.0% to 21.4%. In a cheating experiment we show, that this performance cannot be significantly improved by improving the measure of confidence. By combining the unsupervisedly trained system with our currently best recognizer which is trained on 15.5 hours of transcribed data, an additional error reduction of 5% relative (as compared to the system trained in a standard fashion) is possible.

1. INTRODUCTION

Current speech recognition systems require large amounts of transcribed data for parameter estimation. The transcription process, however, is tedious and expensive. An automatic procedure capable of training a speech recognizer on untranscribed data would therefore be very desirable, and has been the focus of some recent research ([1], [2]). The principle idea of the algorithm used in both [1] and [2] is as follows.

With the transcribed portion of the data, a bootstrap recognizer is built, which is used to generate transcripts of the untranscribed training material. To exclude the erroneous words from these transcripts, a measure of confidence is applied. In the last step, a new recognizer is trained on the remainder of the hypothesis words.

In the experiments described in this paper, we train the initial recognizer only on 30 minutes of transcribed data. Results of experiments with different amounts of untranscribed training data are given.

In our earlier work [2] only a very small amount of transcriptions was available. Currently, however, significantly more material is transcribed. We made use of this additional transcriptions in two ways in order to evaluate the effectiveness of the unsupervised training procedure.

First, we simulated a perfect confidence measure which was able to find 100% of the errors in the hypotheses without discarding a single correctly recognized word. The results achieved in this cheating experiment provide a test of the quality of the confidence measure, and an upper bound for the efficiency of the unsupervised training algorithm.

Second, we trained a speech recognizer in the traditional way using the transcriptions. The results of this experiment are compared to both the results achieved with the cheated and the 'real' confidence measure.

2. THE VIEW4YOU SYSTEM

The **View4You** project is a cooperation between the Interactive Systems Labs and Carnegie Mellon University's Informedia group [3]. It aims at the automatic generation of a searchable multilingual video database. In the prototype system, German and Serbocroatian TV news shows are recorded daily and stored as MPEG compressed files. Using the acoustic signal, a segmenter chops the newscasts into acoustically homogeneous segments ranging from several seconds to few minutes in length. A speech recognition system generates transcriptions for the segments. The segmentation information and the transcriptions are stored in a database.
The user of the system can enter queries in natural language, e.g. 'Tell me everything about the peace negotiations between Mr Netanyahu and Mr Arafat'. Using the speech recognizer's transcriptions in the multimedia database, an information retrieval component computes a ranked order of relevant segments, which are displayed to the user. By clicking on a segment, an MPEG-player is activated that plays the corresponding video segment.
For more details on the **View4You** system, see [4].

2.1. The View4You broadcast news database

For our experiments we used the German part of the View4You database, which has been collected at the University of Karlsruhe. A standard German news program (called 'Tagesschau') is recorded daily and stored as MPEG-1 compressed file with a total bit rate of 1.2 MBit/s and an audio bandwidth of 192 kbit/s, using layer 2 compression and a sampling rate of 44.1 kHz. The audio data is then downsampled to 16 kHz and stored. For the training and the test data, the audio signal is manually segmented and transcribed. The segmentation is done according to the acoustic condition of the audio signal. Therefore, each segment contains either clean speech from the anchor speaker, or speech with all kinds of background noise, like battlefield noise, street noise, other speakers in the background, speech over telephone lines, etc.

There are large differences between the US news shows used by the ARPA broadcast news evaluations [5] and the

'Tagesschau' newscast. We tried to segment the 'Tagesschau' using the same so-called F-conditions used by ARPA, but found that three out of 7 different F-conditions (F1, F5 and FX) are virtually nonexistent in the 'Tagesschau'. Most of the data would be categorized into one of two other F-conditions. Therefore, we decided to use only two classes, **clean** and **distorted**, where **clean** means the anchor speaker portion of the data (and can be identified with ARPA's F0 condition), and **distorted** means everything else (and would mostly be tagged F4 or F2).

For our experiments, a set of 64 transcribed news shows totaling 17 hours of speech was available. 4 of the shows were excluded as test and crossvalidation set.

2.2. The View4You speech recognizer

The speech recognizer of the **View4You** system is based on the JANUS-3 speech recognition toolkit. It uses fully continuous mixture gaussian densities based on decision-tree clustered context-dependent sub-triphones. All mixtures are chosen to have 30 gaussians, and the gaussians are modeled with diagonal covariance matrices. No parameter sharing of covariances or gaussians takes place. In the preprocessing stage, 13 mel-frequency cepstral coefficients, their deltas, and delta-deltas are computed. Mean and variance of the speech part of the signal are normalized. The 39-dimensional input vector is transformed by linear discriminant analysis (LDA) into one 16-dimensional feature vector. To capture the effects of the noise in the data, some noise phones (e.g. for breathing noise), were introduced.
The language model is a standard Kneser-Ney backoff trigram language model based on 102 million words worth of newspaper texts and radio broadcast transcriptions. The most frequent 60k words from the background corpus are used as vocabulary. Since German is an inflecting language with many compound nouns, the vocabulary coverage is relatively low. On the test set, the OOV (out-of-vocabulary) rate is 4.43%.

The decoder computes its hypothesis in a three-pass strategy. Using the intermediate recognition results, VTL normalization [8] and MLLR adaptation [10] are performed.

3. EXPERIMENTAL

In this work, different speech recognizers are compared to each other. In order to guarantee meaningful results of the comparisons, we devised a standard training procedure which was applied automatically to every system we trained. No additional manual hand-tuning took place. All systems, both the supervised and unsupervisedly trained, were trained with this standard training procedure, and the results were taken as-is without further processing. We feel that this procedure guarantees maximal neutrality towards the systems in question.

In the following paragraph, we describe our automatical training procedure in some more detail.

3.1. The training procedure

All systems use the same 60k dictionary, language model, phoneme set and the same preprocessing (13 VTLN-adapted, mel-cepstral mean-normalized coefficients, with their deltas and delta-deltas LDA-transformed to 16 final coefficients). The state alignment is pre-computed and stored in label files. It remains fixed throughout the training process. In the first step, a LDA matrix is computed using the context independent sub-triphones as classes. Initial Gaussian mixtures are generated using the k-means clustering algorithm, and are trained 3 iterations using viterbi training. Then, polyphonic decision trees are computed using a top-down clustering algorithm and a set of 90 phonetically motivated questions. The clustering procedure is terminated when the desired amount of different context dependent models is reached. In an earlier experiment [4] we determined, that at least 15 data samples are required to train one Gaussian. Therefore, the number of context dependent models is automatically chosen such as to have 15 data samples per Gaussian. For large amounts of training data, however, the maximum size of the model is restricted to 5000 triphone models (150,000 Gaussians) due to memory and speed limitations.

With the new triphone models as classes, a new LDA is computed, and new Gaussian mixture parameters are estimated with the k-means algorithm. 5 iterations of viterbi training yield the final acoustic models. In the recognition step, MLLR ([10]) mean adaptation is used. For this adaptation, a lattice-based confidence measure is applied to the hypothesis, so that words with low confidence are not used for adaptation.

For all experiments with unsupervised training, we used a two-stage approach, where an intermediate system was trained using 6 hours of untranscribed data. The state alignment and the hypotheses for this intermediate system were computed with the bootstrap recognizer. The intermediate system was then used to generate transcripts and state alignments for the final training on the whole untranscribed data set.

3.2. Baseline

For training our baseline system, we used 30 minutes of transcribed data from 2 newscasts recorded on November 25, 1996 and November 26, 1996. The baseline system was trained with the automatic procedure described above. The performance of the baseline system on the test set was 32% word error rate (cf table 1).

show (date)	Anchor	non-anchor	total
30/03	21.3%	39.6%	30.6%
13/04	22.7%	37.9%	33.6%
total	22.0%	38.75%	32.05%

Table 1. Baseline word error rates

We trained another system using all 60 transcribed newscasts (15.5 hours of speech). The result on the testset is given in table 2. This result can serve as an upper bound

show (date)	Anchor	non-anchor	total
30/03	12.8%	24.2%	19.6%
13/04	10.9%	24.9%	19.3%
total	11.85%	24.55%	19.5%

Table 2. Transcribed system

for any algorithm that makes use of untranscribed data.

3.3. Measure of Confidence

In all our experiments, we used the lattice-based 'gamma' confidence measure presented in [9]. On the independent test set, the reduction in relative cross-entropy achieved with this confidence measure was 26%. Figure 1 summarizes the performance of the confidence measure. The upper

curve shows the percentage of usable data over the threshold (correctly recognized words) that are marked as 'good' by the confidence tagger. This value should be as high as possible to make best use of the available data.

The lower curve shows the number of recognition errors that are (erroneously) marked as 'good' by the confidence tagger. This value should be as low as possible to avoid training on bad targets.

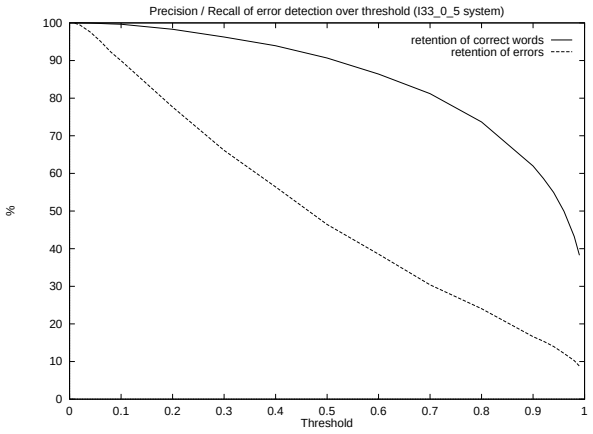

Figure 1. Retention of correct and errorful hypothesis words over threshold (intermediate system)

Obviously, it is not possible to optimize both the retention of correctly recognized words and the rejection of the recognition errors. Therefore, we trained four different recognizers with four values of the threshold on 120 news shows each (approximately 32 hours of speech). The results are summarized in table 3. Only results of the noisy part of the news shows are given, because the results on the anchor speaker part differ only insignificantly in this experiment, ranging between 22.4% and 23.3% word error rate.

Threshold	training data used (hours)	word error rate (non-anchor)
0	32	29.15%
0.2	30	28.65%
0.5	26	27.30%
0.9	19	28.25%

Table 3. Error rates for different operating points of the confidence tagger

The lowest word error rate is reached at a threshold of 0.5. Therefore, this value was chosen in all other experiments.

With a threshold of 0.9, only such words are used for training that have been robustly recognized, i.e. words where the language and acoustic modelling was good. Such words, however, do not add substantial new information to the recognizer as they reflect the current modelling: the system only learns what it already knows. With a lower threshold, words are added to the training that are poorly modelled, and these words increase the modelling capacity of the system; however, more errors are added to the training set, which in turn lowers the performance.

To evaluate the effect of the errors in our measure of confidence, we performed a cheating experiment. In this experiment, we trained two speech recognizers on our training data without using the transcriptions. For the first recognizer, we used our measure of confidence with a threshold of 0.5. For the second recognizer, we used the transcriptions to simulate a perfect confidence measure by tagging each correctly recognized word with 1 and each recognition error with 0. Both systems were trained starting from the baseline system and using the automatic training procedure. The results are summarized in table 4.

System	Error rate (non-anchor)
baseline	38.75%
real confidence measure	28.50%
cheated confidence measure	27.35%
transcriptions	24.55%

Table 4. Effect of a perfect confidence measure

Note that for this experiment the amount of training data was 15.5 hours, which is lower than for the results given in table 3. We could not use the full amount of data as we had no transcriptions available for a large part of it.

3.4. Usage of more data

After collecting another 84 news shows (21 hours of data), unsupervised training was performed again with the enlarged training set (53 hours). However, we were not able to significantly improve on the result achieved with only 32 hours of data. To make use of the additional data, we therefore evaluated a different approach.

We trained a new system on 60 new shows (15 hours of data) and combined this system with the one trained on 32 hours using a ROVER-like [7] scheme. For this, we computed hypotheses on the test set with both systems, and applied the measure of confidence. The output of the two recognizers was aligned against each other. If the two hypotheses disagreed with a given word, the one with the higher confidence score was selected as the final output. Using this scheme, the error rate dropped to 26.4% on the non-anchor portion of the newscasts and 20.6% on the whole testset.

3.5. Improving the performance of the supervised trained system

The same ROVER-like scheme was used to improve the system that was trained using the transcriptions. The output of this system was combined with the output of a system that was trained using 15.5 hours of transcribed data plus 32 hours of untranscribed data. Although the two systems performed at a comparable total word error rate, the errors were made at different locations. Using the confidence measure, we could achieve almost 1% absolute error reduction as compared to our topline recognizer. The results are summarized in table 5.

System	WER (non-anchor)	WER (anchor)	WER total
transcriptions	24.55%	11.85%	19.47%
plus unsupervised (ROVER)	23.45%	11.30%	18.59%

Table 5. Combining supervised and unsupervised training

3.6. End-to-end evaluation

In the View4You system, the speech recognizer is used to generate the index of the video database. In order to evaluate the performance of unsupervised training, we ran an end-to-end evaluation using the speech recognizer that was trained on 51 hours of untranscribed data and 30 minutes of transcriptions. We used a set of ten questions which had been asked by naive users, like e.g. 'Is there anything about Benjamin Netanyahu?', or 'I would like to see reports about the visit of president Herzog in Japan'. The full set of (German) questions is given in [6]. The results in terms of precision and recall are summarized in table 6.

Index	PRC	RCL
transcriptions	0.78	0.69
unsupervised trained	0.75	0.66

Table 6. End-to-end performance

The results show, that the end to end performance of the **View4You** system that makes use of a speech recognizer which has been trained only on 30 minutes of transcriptions and 51 hours of untranscribed data is very close to transcription performance. We therefore conclude, that for video indexing the use of the unsupervised training algorithm is a suitable and inexpensive way to create a fully operational system.

3.7. Summary

Table 7 summarizes the results of our experiments. Using unsupervised training, the error rate could be reduced by 30% as compared to our baseline system trained on 30 minutes of speech.

System	Trainset size	transcribed?	WER
1	0.5 hrs	yes	32.1%
intermediate	6 hrs	no	24.17%
2	15.5 hrs	no	22.40%
3	32 hrs	no	21.42%
combined 2/3	15.5 hrs + 30 hrs	no	20.74%
4	15.5 hrs	yes	19.47%
5	15.5 hrs + 30 hrs	yes/no	19.88%
combined 4/5	15.5 hrs + 30 hrs	yes/no	18.59%

Table 7. Summary of results

4. CONCLUSIONS

We exploited a simple approach to unsupervised learning, where an initial hypothesis is generated by the bootstrap recognizer, some of the recognition errors are spotted by a confidence tagger and the remainder of the words is used for training.

Using only 30 minutes of transcriptions and 51 hours of untranscribed data, the word error rate of our broadcast news recognizer went down from 32% to 21.5% using this approach. We have shown, that the final speech recognizer achieves nearly transcription performance if used in our video indexing system **View4You**.

In another experiment we found that it is possible to make even better use of the untranscribed data by training more than one recognizer and combining the recognition results weighted by word confidence. Using this scheme, the word error rate dropped by another 0.9% to 20.6%. We were also able to reduce the error rate of our best recognizer, which was trained on 15.5 hours of transcribed data, from 19.5% to 18.6% by combining it with the output of our recognizer trained on untranscribed data.

5. ACKNOWLEDGEMENTS

The authors wish to thank all members of the Interactive Systems Labs for useful discussions and active support.

REFERENCES

[1] J. Billa, K. Ma, M. Siu, G. Zavaliagkos, *Acoustic modeling work at BBN*, in Proc. of the Hub-5 Conversational Speech Recognition workshop, NIST, Linthicum Heights, Maryland, November 1997

[2] T. Kemp, A. Waibel, *Unsupervised training of a speech recognizer using TV broadcasts*, in Proc. of ICSLP 98, Vol 5, pp. 2207 ff, Sydney, Australia, November 30 - December 4, 1998

[3] H. Wactlar, A. Hauptmann, M. Witbrock: *Informedia: news-on-demand experiments in speech recognition*, Proc. of ARPA SLT workshop, 1996.

[4] T. Kemp, P. Geutner, M. Schmidt, B. Tomaz, M. Weber, M. Westphal, A. Waibel, *The Interactive Systems Labs View4You video indexing system*, in Proc. of ICSLP 98, Vol 4, pp. 1639 ff, Sydney, Australia, November 30 - December 4, 1998

[5] National Institute of Standards (NIST), *Proceedings of the DARPA Broadcast News transcription and understanding workshop*, Lansdowne, VA, February 8-11, 1998

[6] T. Kemp, M. Weber, P. Geutner, J. Guertler, P. Scheytt, M. Schmidt, B. Tomaz, M. Westphal, A. Waibel, *Automatische Erstellung einer Video-Datenbank: das View4You-System*, in Proc. of the 4th Conference on Natural Language Processing KONVENS-98, Vol. 1, pp. 347 ff., Bonn, Germany, October 5-7, 1998

[7] J. G. Fiscus, *A post-processing system to yield reduced word error rates: recognizer output voting error reduction (ROVER)*, in Proc. of the Hub-5E conversational speech recognition workshop, ARPA, 1997

[8] P. Zhan, M. Westphal, *Speaker normalization based on frequency warping*, in Proc. ICASSP-97, Munich, April 1997

[9] T. Kemp, T. Schaaf, *Estimating confidence using word lattices*, in Proc. EUROSPEECH-97, Vol 2, pp. 827 ff, Rhodes, Greece, September 1997

[10] C.J. Legetter, P.C. Woodland: *Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models*, Computer Speech and Language **9** (1995), 171-185