UNSUPERVISED TRAINING OF A SPEECH RECOGNIZER USING TV BROADCASTS

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 without transcriptions.
The experiments were carried out with TV newscasts, that were recorded using a satellite receiver and a simple MPEG coding hardware. The newscasts were automatically segmented into segments of similar acoustic background condition. This material is inexpensive and can be made available in large quantities, but there are no transcriptions available.
We develop a training scheme, where a recognizer is bootstrapped using very little transcribed data and is improved using new, untranscribed speech. We show that it is necessary to use a confidence measure to judge the initial transcriptions of the recognizer before using them. Higher improvements can be achieved if the number of parameters in the system is increased when more data becomes available. We show, that the beneficial effect of unsupervised training is not compensated by MLLR adaptation on the hypothesis. In a final experiment, the effect of untranscribed data is compared with the effect of transcribed speech. Using the described methods, we found that the untranscribed data gives roughly one third of the improvement of the transcribed material.

1. INTRODUCTION

Porting a speech recognition system to a new language requires the existence of a large training corpus of transcribed data. In many cases, such corpora do not exist. The transcription of large amounts of data, however, takes a long time and is rather expensive. On the other hand, untranscribed speech data can be made available quickly and in large quantities, e.g. from radio or TV broadcast recordings. The question is, whether untranscribed data can be used to enhance a speech recognizer that has been trained on very little manually transcribed bootstrapping data. The ultimate goal is a system that can improve its speech recognition performance just by watching TV.

Recently, this topic has received some attention [6] [3]. [6] focused on using captions to improve speech recognition performance, where the captions can be gathered automatically but are imperfect and not properly aligned to the acoustic data. Although there are captions available for a part of our data, we do not make use of this information but focus on the case where there is only a large collection of untranscribed data and a small set of transcribed training material available.

[3] investigated the problem of unsupervised training for Spanish data. The confidence measure employed in this work was simulated using the transcriptions in order to increase the amount of training material tagged as useable. Gains in terms of word error rate could be achieved which focused on the recognition of speakers that were present in the untranscribed training material.

The remainder of this paper is structured as follows. First, we describe the View4You system and the View4You database that was used for all experiments. Then, a short description of our speech recognition system is given. In the experimental section, we give details of all experiments with untranscribed data and report our results.

2. THE VIEW4YOU SYSTEM

The View4You project is a cooperation between the Interactive Systems Labs and the Carnegie Mellon Universities Informedia group [5]. 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 automatic transcriptions are stored in a database.
The user of the system can give queries in natural language, e.g. 'Tell me everything about the peace talks 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 [1].

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 exist specific differences between the US news shows used by the ARPA broadcast news evaluations 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', and that 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 ARPAs 'F0' condition), and **distorted** means everything else (and would mostly be tagged F4 or F2).

For our experiments, only a set of 12 transcribed news shows totaling 3 hours of speech was available. 8 shows (approx. 2 hours of speech) were used for training, 2 shows for test, and 2 were reserved as additional cross validation data.

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 covariances. 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 45 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 approximately 5%.

The recognizer was trained on two hours of speech (8 news broadcasts). Using this limited training material, we trained several systems that were clustered to different numbers of polyphones. We found, that the error rate dropped with increasing number of parameters to a maximum at about 15 frames per gaussian. If less than 15 frames were used to estimate a gaussian, the error rate rose again. Therefore, we chose this number of parameters (15 frames of data per gaussian) for our baseline recognizer. For more details on this experiment and the recognizer setup, see [1].

The decoder computes its hypothesis in a three-pass strategy. Using the intermediate or final recognition results, VTL normalization [8] and MLLR adaptation [7] can be performed.

The resulting system was tested on two complete news shows of 15 minutes each, assuming perfect (manual) segmentation. The baseline results, using vocal tract length normalization but no MLLR, are shown in table 1.

show (date)	Anchor	non-anchor	total
30/03	20.2%	41.0%	30.6%
13/04	22.7%	44.5%	33.6%
total	21.5%	42.8%	32.2%

Table 1. Baseline word error rates

3. EXPERIMENTAL

For our experiments with untranscribed data, we collected 10 additional news shows (dated between April 3, 1997 and April 12, 1997). This news shows were automatically segmented (for details see [1]) and segment hypothesis were computed. For the computation of the segment hypotheses, we used our baseline system described above.

3.1. Training on the recognition result

In a first experiment, we computed MLLR adaptation on the hypotheses of the unknown data. With this adapted system, we ran a recognition run on the test data, but the error rate was increased by 1.8% absolute due to the errors in the automatic transcriptions. Therefore, we decided to use a measure of confidence (MOC) tagger to identify correctly recognized words in the hypotheses and to focus the training on these words.

3.2. Measure of confidence

To tag all words in the hypotheses with a a-posteriori probability of being correctly recognized, we used the lattice-based 'gamma' confidence measure. This is basically a forward-backward algorithm computed on the word lattice produced by the speech recognizer. The gamma confidence measure is described in detail in [4].

We ran our confidence measure estimation on the output of our baseline system on the untranscribed data. To evaluate the performance of the confidence measure, we also computed confidence scores on our test data, where transcriptions are available. The results in terms of PRC and RCL of the correct (C) words in the hypothesis and the recognition errors (E) are given in table 2. At a threshold of 0.9, our confidence measure can identify correctly recognized words with an accuracy of 90% and will find 57% of all correctly recognized words. At a threshold of 0.5, close to 90% of all correctly recognized words are identified, but only 82% of all words tagged as correct are really correct.
The performance of a confidence tagger is frequently given

in terms of relative cross entropy S. The S-score of our system on our test set was 0.261.

threshold	PRC (C)	RCL (C)	PRC (E)	RCL (E)
0.5	0.82	0.89	0.74	0.60
0.9	0.90	0.57	0.50	0.88

Table 2. PRC and RCL

3.3. Unsupervised training with confidence measure

Using our baseline system, we computed viterbi state alignments for all hypotheses on the untranscribed data set. All words with a confidence of less than 0.5 were then excluded from further processing by setting their state occupation probability to zero. We also computed state alignments (label files) for the transcribed 8 shows of our training set. The resulting set of label files, both from the transcribed and the untranscribed portion of the data, was used to bootstrap and train a new system. This system was then tested on our testset. The result is shown in table 3. With the use of a confidence measure, the word error rate on the field speech segments drops by 5% relative; however, there is a slight increase in word error rate on the anchor speaker segments.

Condition	baseline	unsupervised trained
anchor	21.5	21.7
field speech	42.8	40.6
total	32.2	31.2
improvement	-	3.1%

Table 3. Word error rate of unsupervised trained system

3.4. Increasing the number of parameters

Since the total amount of training data has increased, it should be possible to increase the number of parameters in our system. Therefore, we trained another system, where the number of polyphones was chosen such that this system again used 15 frames of data to estimate one mean vector. The result of this experiment is shown in table 4. There is a word error rate reduction of 9.3% for the field speech, and even a slight improvement for the anchor speaker segments.

Condition	baseline	unsupervised trained
anchor	21.5%	21.1%
field speech	42.8%	39.7%
total	32.2%	30.4%
improvement	-	5.5%

Table 4. Word error rate of unsupervised trained system with structure adaptation

3.5. Linear confidence measures

In the experiments described above, the confidence measure score was only used to decide whether a word is assumed to be correct (1.0) or not (0.0). Therefore, a word with a confidence score of 0.45 would be discarded, but a word with a confidence score of 0.55 would be used in the training. To investigate the effect of a linear scaling, we replaced the state likelihoods of all label files with the confidence score of the corresponding word. In effect, this means that a word with a confidence of 1.0 is used normally in the training, but the data frames aligned to a word with confidence 0.5 are weighted only half.

Using this scheme, we trained a new system with 15 frames/parameter, as described above. The result is shown in table 5. There is a significant increase in word error rate compared to the 'digital' use of the confidence measure.

Condition	baseline	unsupervised trained
anchor	21.5%	21.2%
field speech	42.8%	41.2%
total	32.2%	31.4%
improvement	-	2.5%

Table 5. Word error rate using linear MOC

3.6. Unsupervised MLLR on the test data

It is well known, that maximum likelihood linear regression (MLLR, [7]) adaptation using the hypothesis of a first recognition pass significantly reduces the word error rate for most speech recognition tasks. In our experience, we frequently found that an improvement achieved on a system without MLLR did not carry over to the MLLR-adapted system. Therefore, we ran both our baseline recognizer and the best unsupervised trained system, using a two-pass recognition, where the acoustic models were adapted using the hypothesis of the first recognition pass. The results are summarized in table 6. We found, that the gains achieved by unsupervised training are reduced, but not cancelled out by MLLR adaptation.

Condition	baseline	unsupervised trained
anchor	21.5%	20.9%
field speech	39.7%	38.2%
total	30.6%	29.5%
improvement	-	3.5%

Table 6. Word error rates with MLLR

3.7. Supervised vs. unsupervised training

It is interesting to compare unsupervised training with 'standard' training (i.e., with manually generated transcriptions). Since we did not have transcriptions for the 10 news shows used for unsupervised training, we designed a new experiment, where we assumed that the size of the training database was only 2 news shows or 30 minutes of speech. We trained a recognizer with optimal number of parameters on the 2 transcribed news shows. This recognizer performed at 36.9% word error rate on our testset, when both vocal tract length and MLLR adaptation were used.

We computed hypotheses for the other 6 news shows of our training set using this system. Note that this data can be regarded as unseen since the system has not been trained on it. The hypotheses were confidence annotated as described above. Then, we trained a new system, discarding every

word with less then 0.5 confidence and optimal number of parameters. The performance of this system, when using both vocal tract length and MLLR adaptation, is shown in table 7.

Condition	baseline	unsupervised	supervised
anchor	27.9%	26.1%	21.5%
field speech	45.9%	43.5%	39.7%
total	36.9%	34.8%	30.6%
improvement	-	5.6%	17.1%

Table 7. Result starting with 30 minutes training

Training without transcriptions is capable to give approximately one third of the improvement of training with transcriptions.

In this experiment, the improvement using unsupervised training is larger than for the baseline system trained on 8 news shows. This is probably due to the larger relative increase in the size of the training corpus (2 shows transcribed plus 6 shows untranscribed, vs. 8 shows transcribed plus 10 shows untranscribed).

4. CONCLUSION

In this work we have shown, that it is possible to reduce the word error rate of a speech recognition system bootstrapped on very little training data with the use of untranscribed additional data. For 30 minutes of initial training material, untranscribed data yielded about one third of the gain of transcribed data. We found that the use of a good measure of confidence tagger is mandatory. Our experiments have shown, however, that roughly one half of the gain achieved is due to the increase in the number of parameters which is possible when more data becomes available. Since it is usually impossible to increase the amount of recognizer parameters beyond a certain level given by computational and memory constraints, we expect our method to become less effective for larger amounts of additional data.

5. ACKNOWLEDGEMENTS

This work was carried out at the Interactive Systems Labs, Karlsruhe. The authors would like to thank all members of the Interactive Systems Labs for helpful discussions and support. We also thank Manfred Weber for providing the manual transcriptions for the news shows.
The views and conclusions contained in this document are those of the authors.

REFERENCES

[1] T. Kemp, P. Geutner, M. Schmidt, B. Tomaz, M. Weber, M. Westphal, A. Waibel, *The interactive systems labs View4You video indexing system*, elsewhere in these proceedings

[2] T. Kemp, A. Waibel, *Reducing the OOV rate in broadcast news speech recognition*, elsewhere in these proceedings

[3] G. Zavaliagkos, T. Colthurst, *Utilizing untranscribed training data to improve performance*, in Proc. of the broadcast news transcription and understanding workshop, Landsdowne Conference Resort, VA, February 8-11, 1998, Morgan Kaufman, ISBN 1-55860-564-9, pp. 301 ff

[4] T. Kemp, T. Schaaf, *Estimating confidence using word lattices*, in Proc. EUROSPEECH 97, Rhodos, Greece, September 1997

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

[6] P. Placeway, J. Lafferty: *Cheating with imperfect transcripts*, in Proc. ICSLP 96, Philadelphia, September 1996

[7] 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

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