UNSUPERVISED TRAINING OF ACOUSTIC MODELS FOR LARGE VOCABULARY CONTINUOUS SPEECH RECOGNITION

Frank Wessel and Hermann Ney

Lehrstuhl für Informatik VI, Computer Science Department,
RWTH Aachen – University of Technology, 52056 Aachen, Germany
{wessel, ney}@informatik.rwth-aachen.de

ABSTRACT

For speech recognition systems, the amount of acoustic training data is of crucial importance. In the past, large amounts of speech were thus recorded and transcribed manually for training. Since untranscribed speech is available in various forms these days, the unsupervised training of a speech recognizer on recognized transcriptions is studied in this paper. A low-cost recognizer trained with only one hour of manually transcribed speech is used to recognize 72 hours of untranscribed acoustic data. These transcriptions are then used in combination with confidence measures to train an improved recognizer. The effect of confidence measures which are used to detect possible recognition errors is studied systematically. Finally, the unsupervised training is applied iteratively. Using this method, the recognizer is trained with very little manual effort while loosing only 14.3% relative on the Broadcast News '96 and 18.6% relative on the Broadcast News '98 evaluation test sets.

1. INTRODUCTION

The personnel expenditure of transcribing audio data is very high and it is thus desirable to train a recognizer with as little hand-transcribed training material as possible. One way to reduce the manual effort is to use an existing recognizer to transcribe new data or to transcribe a few hours of new acoustic data manually, to train a recognizer with these data which is then used to generate transcriptions for larger quantities of untranscribed data, and to train the final recognizer on these recognized transcriptions.

In [1] the authors use a speech recognizer trained with 30 minutes of manually transcribed German broadcast news to recognize untranscribed speech data. The system is then retrained with these transcriptions in combination with confidence measures. In comparison with the initial system, the word error rates are reduced substantially. Unfortunately, the authors report a deterioration of the recognition performance on clean conditions. In comparison with a system trained with the correct transcriptions of the complete training corpus, the word error rate increases by 13.7% relative. In addition to the fact that the speech corpus is not publicly available, the amount of training material is very limited. The question remains whether the suggested confidence measure and the unsupervised training would perform equally well in a different scenario with more untranscribed training data available.

In [2] closed-captions of the broadcasts are used to determine which segments of speech should be used for unsupervised training. The recognized transcriptions and the closed-captions are aligned and only those segments are used for training, where the two transcriptions are in agreement. For training, the TDT-2 corpus is used. Recognition results are reported for the Broadcast News '99 evaluation corpus. Compared with a speech recognition system trained on correct transcriptions, the word error rate increases by 9.5% relative. Obviously, this approach cannot be used if no closed-captions are available.

This paper focuses on the unsupervised training of a recognizer on recognized transcriptions for a 72 hour subset of the Broadcast News '97 training corpus and on a systematic study of the effect of confidence measures. It is assumed that no closed-captions or other information about the corpus are available. A particular focus is on the effect of different amounts of manually transcribed speech which are used to bootstrap the initial recognizer needed to recognize the untranscribed training corpus. Based on the experimental results, a general procedure for the *unsupervised training* of a speech recognizer is derived which can be used to rapidly prototype a system for new languages or domains with very little manual effort.

2. TRAINING PROCEDURE

Given an initial alignment between the feature vectors and the Hidden Markov Model states, a phonetic classification and regression tree (CART) is computed. Based on this CART, a linear discriminant analysis (LDA) matrix is estimated. In order to reduce the effect of the CART and the LDA matrix which were used to compute the initial alignment, a second CART is estimated with the first LDA matrix and a second LDA matrix is then computed with this second CART. With the second CART and LDA matrix, the parameters of the acoustic models are then estimated (the Gaussian densities are split eight times). The resulting final time alignment is then used to repeat all of the above steps. The sequence of these steps is from now on referred to as the *standard training procedure* used during all experiments. In particular, no empirical training parameters were tuned so that the experimental results are compara-

This work was partly funded by the Philips Research Laboratories Aachen and by the European Commission in the framework of the CORETEX project under grant IST-11876.

ble. The unsupervised training procedure itself consists of several steps which include the above standard training:

1. The acoustic models are trained with the standard training procedure on small amounts of manually transcribed speech data.
2. These models are used to recognize a large untranscribed training corpus which is then used together with the recognized transcriptions to augment the manually transcribed part of the corpus.
3. The standard training is applied. The feature vectors aligned to words whose posterior probability [3, 4] is below a specified threshold are simply omitted. Feature vectors aligned to the silence model are omitted if one of the neighboring words is tagged as incorrect.

In order to study the effect of the acoustic models used to recognize the training corpus, the recognition is performed with models trained on different amounts of manually transcribed data. In addition, the training of the acoustic models on recognized transcriptions is performed with different confidence tagging thresholds; i.e., different amounts of training material which are most presumably correct are retained in the training corpus. During all experiments, the language model was not changed since the main concern was to study the training of the acoustic models. The language model perplexity was 221.2 on the Broadcast News '97 training corpus, 218.1 on the Broadcast News '96 eval test set and 213.7 on the Broadcast News '98 eval test set.

3. BOOTSTRAPPING

In a first experiment, the effect of recognition errors in the recognized transcriptions was studied. The training corpus for all experiments was a 72 hour subset of the Broadcast News '97 training corpus. For all experiments, gender-independent models were used. No speaker-adaptive or normalization methods were applied. The testing corpus was the official Broadcast News '96 evaluation test set.

The training corpus was transcribed with a recognizer trained previously on the Broadcast News '96 training corpus with manual transcriptions (a word error rate of 33.1% was achieved on the Broadcast News '96 eval test set with this system). The Broadcast News '97 training corpus was then used with the recognized transcriptions to run the unsupervised training with different confidence thresholds. The first line of Table 1 shows the word error rate for the Broadcast News '96 testing corpus for the system trained with the manual transcriptions of the complete subset of the Broadcast News '97 training corpus. This word error rate is the baseline for all experiments. As the second line shows, the word error rate increases by only 1.8% relative if using the recognized transcriptions. This result is very surprising, since the word error rate of the transcriptions of the training corpus is 32.5%. Table 1 also shows the amount of training material and the number of Gaussian densities. As the experiment shows, the word error rate is not reduced with confidence measures. This result is attributed to two opposed effects: if the recognizer used to transcribe the training corpus is trained on large amounts of transcribed data, most of the incorrectly recognized words will be acoustically similar to the correct words. The negative

Table 1: Word error rates on the Broadcast News '96 evaluation corpus. The word error rate of the recognized transcriptions was 32.5% and the phoneme error rate 17.3%.

transcriptions	thresh.	corp. [h]	#dns	WER [%] eval '96
manual		72	273k	33.5
recognized		72	345k	34.1
recognized	0.3	66	333k	34.2
+	0.5	62	335k	34.1
confidence	0.7	56	318k	34.1
measures	0.9	47	294k	34.3

impact of these errors is thus rather small. In this scenario, confidence measures cannot improve the performance since they do not only exclude words which might be erroneous but also reduce the amount of training material. In order to validate these assumptions, the phoneme error rate (PER) on the training corpus was computed. Since the manually constructed reference transcriptions do not contain pronunciation variants, a phoneme graph of all possible pronunciation variants of the manual transcription was aligned with the phonetic transcription of the recognized sentence. The Levensthein alignment was computed with an algorithm implemented previously to compute the word graph error rate, cf. [5] for details. While the word error rate on the training corpus is 32.5%, the phoneme error rate is roughly half as high with 17.3%. In addition, most of the 9.2% of substituted phonemes are acoustically very similar.

The experiment shows that the word error rate hardly increases if the recognizer is trained on transcriptions generated with a well-tuned system. The question remains how a system trained with a few hours of speech would perform.

4. LOW-COST BOOTSTRAPPING

The scenario for the following experiments is as follows: it is assumed that the training corpus is untranscribed, but chopped into segments, and that no initial acoustic models, no initial CART, and no initial LDA matrix are available as in the previous section. In this scenario, it is straightforward to transcribe a small portion of the training corpus manually, to train a speech recognizer on it, and to use this system to generate transcriptions for the rest of the corpus. For the following experiments, the subset of the corpus which has to be transcribed manually is defined as the set of segments with a duration of less than a specified maximum. The motivation for this approach was that a linear time alignment of the features and the Hidden Markov Model states has to be used initially since no acoustic models are available. If the segments are too long, the resulting alignment path moves too far from the correct path between the features and the states. Assuming that the corpus subset was transcribed manually, the features and the Hidden Markov Model states are aligned linearly in order to estimate the parameters of single Gaussian density monophone Hidden Markov Models. These models are then used to compute an improved alignment and to start the unsupervised training. In the following, experimental results are

Table 2: Sizes of the different subsets of the 72 hour subset of the Broadcast News '97 training corpus and error rates achieved with the acoustic models trained on these subsets. PER denotes the phoneme error rate.

dur. [s]	corp. [h]	#dns	errors [%] on train '97 WER	PER	WER [%] eval '96
2	1.2	9k	67.5	44.0	71.3
4	3.1	23k	51.1	30.3	54.9
6	5.6	40k	43.6	24.7	47.5

presented for different maximum durations of the segments, i.e., for different sizes of the initial hand-transcribed subset of the corpus, and for different confidence thresholds during step three of the unsupervised training.

The second column of Table 2 gives the size of the corpus subsets for the different maximum segment lengths, given in the first column. These subsets were transcribed manually – the official transcriptions of the corpus were used to simulate this process – and the CART, the LDA matrix, and the acoustic models were trained with these transcriptions. Table 2 also shows the word error and phoneme error rates on the training corpus achieved with the different initial acoustic models. As the experiments show, the word error rates on the training corpus increased drastically in comparison with the system trained on manual transcriptions of the Broadcast News '97 training corpus. Despite these high word error rates, those parts of the training corpus which were not transcribed manually were used with the recognized transcriptions to augment the small portions of manually transcribed training material and the complete corpus was used to train new acoustic models.

Table 3 shows the results for the initial system trained on 1.2 hours of speech. As the experiment shows, the word error rate can be reduced from 71.3% to 49.4% using the recognized transcriptions for the untranscribed parts of the training corpus in combination with the manually generated transcriptions for the 1.2 hours of training data. The additional use of confidence measures reduces the word error rate to 44.0%. The total relative reduction is thus 38%. In contrast to the results discussed in [1], the error rates were reduced on all conditions. Although the word error rates on the testing corpus are very high, the experiments show the potential of the unsupervised training. The maximum possible reduction of the word error rate which could be achieved ideally with confidence measures is presented in the third line of Table 3. In this experiment, the feature vectors of all correctly recognized words were used to train the acoustic models whereas all others were omitted. This amounts to an ideal confidence measure which is able to detect recognition errors with 0% false acceptance and false rejection. As this additional experiment shows, the word error rate could be lowered by an additional 9% relative if a perfect confidence measure were available.

Table 4 shows the experimental results for the systems initially trained on 3.1 hours of speech. As the results show, the word error rate on the testing corpus was reduced from 54.9% to 41.7% with the system trained on the automatically generated transcriptions. Using the confidence

Table 3: Results using initial acoustic models trained with 1.2 hours of manually transcribed audio material.

type of transcription [1]	thresh.	corp. [h]	#dns	WER [%] eval '96
none		1.2	9k	71.3
recognized		72	440k	49.4
recognized + confidence measures	0.50	32	360k	46.1
	0.70	23	320k	45.2
	0.90	14	251k	44.0
	0.95	11	218k	45.5
correctly recognized		28	346k	40.0
manual		72	273k	33.5

Table 4: Results using initial acoustic models trained with 3.1 hours of manually transcribed audio material.

type of transcriptions [1]	thresh.	corp. [h]	#dns	WER [%] eval '96
none		3.1	23k	54.9
recognized		72	454k	41.7
recognized + confidence measures	0.50	45	417k	39.7
	0.70	35	393k	39.4
	0.90	25	351k	39.2
	0.95	21	327k	39.8
correctly recognized		40	402k	36.4
manual		72	273k	33.5

Table 5: Results using initial acoustic models trained with 5.6 hours of manually transcribed audio material.

type of transcriptions [1]	thresh.	corp. [h]	#dns	WER [%] eval '96
none		5.6	40k	47.5
recognized		72	460k	38.8
recognized + confidence measure	0.3	59	445k	38.4
	0.5	51	430k	37.6
	0.7	42	416k	36.8
	0.9	32	388k	37.4
correctly recognized		46	421k	35.3
manual		72	273k	33.5

measure, the word error rate decreased to 39.2%. The last set of experiments with 5.6 hours of manually transcribed data is summarized in Table 5. The word error rate was reduced from 47.5% to 38.8% using all of the 72 hours of training data. The use of the confidence measure further reduced the word error rate to 36.8%.

As the four tables clearly show, the relative reduction of the word error rate becomes the smaller, the larger the amount of initial hand-transcribed material is. The additional relative reduction which can be achieved with the

[1] "Type of transcription" refers only to the untranscribed parts of the training corpus which are used in addition to the small manually transcribed part. In case of "none", only the manually transcribed part was used.

Table 6: Results for the eval '96 and the eval '98 test set and the repeated application of the unsupervised training. The confidence tagging threshold was 0.7 for all iterations. PER denotes the phoneme error rate.

iter.	corp. [h]	#dns	errors [%] on train '97 WER	PER	WER [%] on eval '96	'98
1	1.2	9k	67.5	44.0	71.3	65.5
2	23	320k	49.2	27.2	45.2	36.9
3	49	428k	44.6	24.1	40.9	32.0
4	55	441k	43.0	23.0	39.5	30.9
5	57	446k	41.9	22.3	39.1	30.1
6	58	447k	41.2	21.8	38.4	29.6
7	59	448k	-	-	38.3	29.3

confidence measure also decreases for increasing sizes of the initial hand-transcribed training corpus.

5. ITERATIVE UNSUPERVISED TRAINING

The speech recognition system described in Table 5 performs quite well in comparison with a word error rate of 33.5% achieved with the complete manually transcribed training corpus. Since the quality of the best system in Table 3 trained on recognized transcriptions is comparable with the system initially trained with 5.6 hours of manually transcribed training material, it is straight forward to repeat the process of unsupervised training with the best system in Table 3, i.e., to use these acoustic models to recognize the training corpus again and to train a new system with these recognized transcriptions.

The manually transcribed 1.2 hours of the training corpus were thus used to recognize the remaining 70.8 hours. Using these transcriptions, a new system was trained with the unsupervised training procedure in combination with the confidence measure. This system was then used to generate new transcriptions. The whole process was repeated seven times. The baseline word error rate for the eval '96 test set with the fully tuned speech recognition system is 33.4% and 24.7% for the eval '98 test set. As the results presented in Table 6 show, the iterative application of the unsupervised training procedure can be used to train a speech recognition system with very little manual effort. Instead of 72 hours, only one hour of speech was transcribed manually. In comparison with a system trained on the manual transcriptions of the complete training corpus, the word error rate increases by 14.3% relative on the eval '96 and by 18.6% relative on the eval '98 corpus.

6. CONCLUSIONS

An unsupervised procedure for the training of acoustic models was studied. Experimental results were presented for the Broadcast News '96 and the '98 evaluation corpora. Table 7 summarizes the effect of confidence measures during the initial recognition of the Broadcast News '97 training corpus with a low-cost recognizer. The experiments show that confidence measures can be used successfully to restrict the unsupervised training to those portions of the transcriptions where the words are most probably correct.

Table 7: Comparison of word error rates for the initial recognition of the training corpus.

type of transcription for the untranscribed part of the corpus	corp. [h]	WER [%] eval '96
none	1.2	71.3
recognized	72	49.4
recognized + confidence measures	14	44.0
correctly recognized	28	40.0

Table 8: Comparison of word error rates for the different training procedures. The second column gives the amount of data with manual transcriptions.

training procedure	corp. [h] man.	tot.	WER [%] on eval '96	'98
standard	1.2	1.2	71.3	65.5
unsupervised	1.2	14	44.0	35.8
iterative	1.2	59	38.3	29.3
standard	72	72	33.5	24.7

Table 8 shows the experimental results for the different training methods studied in this paper. Using the suggested method, the system was initialized with only one hour of manually transcribed acoustic training data and was improved iteratively. The final word error rate on the testing sets increased by 14.3% and 18.6% relative in comparison with a system trained on the manual transcriptions of the complete training corpus. Using this training procedure, the manual expenditure of transcribing speech data can be reduced drastically for new application scenarios.

7. REFERENCES

[1] T. Kemp and A. Waibel, "Unsupervised training of a speech recognizer: Recent experiments," in *Proceedings European Conference on Speech Communication and Technology*, Budapest, Hungary, Sept. 1999, vol. 5, pp. 2725–2728.

[2] L. Lamel, J. L. Gauvain, and G. Adda, "Lightly supervised acoustic model training," in *Automatic Speech Recognition Workshop*, Paris, France, Sept. 2000, pp. 150–154.

[3] F. Wessel, K. Macherey, and R. Schlüter, "Using word probabilities as confidence measures," in *Proceedings International Conference on Acoustics, Speech, and Signal Processing*, Seattle, WA, USA, 1998, vol. 1, pp. 225–228.

[4] F. Wessel, R. Schlüter, K. Macherey, and H. Ney, "Confidence measures for large vocabulary continuous speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 3, pp. 288–298, Mar. 2001.

[5] S. Ortmanns, H. Ney, and X. Aubert, "A word graph algorithm for large vocabulary continuous speech recognition," *Computer Speech and Language*, vol. 11, no. 1, pp. 43–72, Jan. 1997.