Unsupervised Training of Acoustic Models for Large Vocabulary Continuous Speech Recognition

Frank Wessel and Hermann Ney, *Member, IEEE*

***Abstract*—For large vocabulary continuous speech recognition systems, the amount of acoustic training data is of crucial importance. In the past, large amounts of speech were thus recorded from various sources and had to be transcribed manually. It is thus desirable to train a recognizer with as little manually transcribed acoustic data as possible. Since untranscribed speech is available in various forms nowadays, the unsupervised training of a speech recognizer on recognized transcriptions is studied in this paper. A low-cost recognizer trained with between one and six h of manually transcribed speech is used to recognize 72 h of untranscribed acoustic data. These transcriptions are then used in combination with a confidence measure to train an improved recognizer. The effect of the confidence measure which is used to detect possible recognition errors is studied systematically. Finally, the unsupervised training is applied iteratively. Starting with only one h of transcribed acoustic data, a recognition system is trained fully automatically. With this iterative training procedure, the word error rates are reduced from 71.3% to 38.3% on the Broadcast News'96 evaluation test set and from 65.6% to 29.3% on the Broadcast News'98 evaluation test set. In comparison with an optimized system trained with the manually generated transcriptions of the complete 72 h training corpus, the word error rates increase by 14.3% relative and 18.6% relative, respectively.**

***Index Terms*—Confidence measures, speech recognition, unsupervised training, word graph, word posterior probabilities.**

I. Introduction

THE BUILDING of a recognizer for a new language, a new domain, or different acoustic conditions usually requires the recording and transcription of large amounts of speech data which are then used to train the acoustic models. In contrast to the early days of speech recognition, large collections of speech data are available these days. Radio and television broadcasts can easily be recorded from different sources including the internet. Unfortunately, most of the acoustic material comes without a detailed transcription and has to be transcribed manually. One possible way to reduce the manual effort is to use an already existing speech recognizer trained previously on a different database to transcribe new data automatically, or to transcribe only a few hours of the new acoustic material manually, to build a recognizer with these data, and to use this recognizer to recognize larger quantities of untranscribed training data. If the recognizer was able to identify those portions of the new data where the automatic transcription is sufficiently accurate, these data could be used to train the acoustic models despite possible recognition errors in the transcriptions. Another advantage of this approach is that the recognizer lexicon need not be extended to cover all words or pronunciation variants of words occurring in the acoustic training corpus.

In [1], the authors investigate such a scenario for the CallHome corpus. They use a recognizer trained on three h of speech to transcribe 25 h of untranscribed data. A combination of several confidence features, cf. [2], is used to tag the recognized words and only those words whose confidence exceeds a certain threshold are used to train the acoustic models. The authors report only a very small reduction of the word error rate. In [3], a speech recognition system trained with 30 min of manually transcribed German broadcast news is used to recognize larger amounts of untranscribed speech in combination with confidence measures. In comparison with the initial system, the word error rates are reduced substantially. In comparison with a system trained on the correct transcriptions of the complete training corpus, the word error rate increases by 13.7% relative, from 30.6% to 34.8%. Unfortunately, this method is not applied to a publicly available corpus. In [4], [5], and [6], closed captions are used instead of confidence measures to filter the recognized transcriptions. Only those parts of the speech data base are used for training where the closed captions and the recognized transcription are in agreement. For training, the TDT-2 corpus is used. Recognition results are reported for the Broadcast News'99 evaluation test set in [5]. In comparison with a speech recognition system trained on correct, manually generated transcriptions, the word error rate increases by 9.5% relative from 18.0% to 19.8%. The filtering with closed captions imposes a strong constraint on the recognition process since the transcription is already very close to what was really said. This approach is very effective but obviously, it cannot be used if no closed captions are available.

In the following sections, a general procedure for the *unsupervised training* of acoustic models will be derived which can be used to rapidly prototype speech recognition system for new languages or domains with only very little manual expenditure [7]. This procedure does not rely on closed captions and can be used for speech corpora with no additional information about the data contained in the acoustic training corpus.

II. Description of the Training Procedure

Unsupervised in the context of training refers to the training of acoustic models with no human supervision, e.g., only very

Manuscript received August 21, 2002; revised October 28, 2003. This work was supported in part by the Philips Research Laboratories Aachen and in part by the European Commission in the framework of the CORETEX Project under Grant IST-11876. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Shrikanth Narayanan.

F. Wessel was with the Lehrstuhl für Informatik VI, University of Technology (RWTH), 52056 Aachen, Germany. He is now with Vodafone D2 GmbH, Office Products and Corporate Solutions, 40547 Düsseldorf, Germany (e-mail: frank.wessel@gmx.de; frank.wessel@vodafone.com).

H. Ney is with the Lehrstuhl für Informatik VI, University of Technology (RWTH), 52056 Aachen, Germany (e-mail: ney@cs.rwth-aachen.de).

Digital Object Identifier 10.1109/TSA.2004.838537

1063-6676/$20.00 © 2005 IEEE

compute initial time alignment between feature vectors and Hidden Markov Model states
compute phonetic *classification and regression tree* (CART) on basis of acoustic feature vectors and time alignment
compute *linear discriminant analysis* (LDA) matrix for resulting CART classes
compute new CART on basis of LDA-transformed feature vectors
compute new LDA matrix for resulting new CART classes
estimate parameters of acoustic models with last CART and LDA
compute new time alignment
repeat until performance is satisfactory

Fig. 1. RWTH standard training procedure.

estimate parameters of acoustic models with standard training procedure on small amounts of manually transcribed speech
use resulting acoustic models to recognize large untranscribed training corpus and to generate word graphs
compute word posterior probabilities on basis of word graphs in order to obtain confidence of recognized words
use previously untranscribed training corpus together with recognized transcriptions to augment small, manually transcribed corpus
use standard training procedure to train new acoustic models on complete corpus in combination with confidence measure

Fig. 2. RWTH unsupervised training procedure.

little effort is spent to bootstrap an initial speech recognition system which is used to transcribe speech data. In the following, the unsupervised training procedure and the necessary modifications of the training algorithms are presented in detail. The main building block of the unsupervised training procedure is the RWTH *standard training procedure* which is illustrated in Fig. 1. In the RTWH speech recognition sytem, context-dependent triphone hidden Markov models (HMMs) [8] in the so-called Bakis topology are used [9], [10]. Normally, two iterations are used. This standard training procedure is applied to all experiments without changes. In particular, no training parameters, e.g., number of LDA or CART classes, language model weights, are tuned in order to minimize the amount of manual intervention and to be able to compare the experimental results.

The unsupervised training procedure itself is illustrated in Fig. 2 and consists of several steps which include the above standard training procedure.

The application of confidence measures during the final step of the unsupervised training procedure is rather straightforward and is illustrated in Fig. 3. The recognized transcriptions of the training corpus are annotated with the standard RWTH word-graph-based word posterior probability, cf. [11], [12] for details,

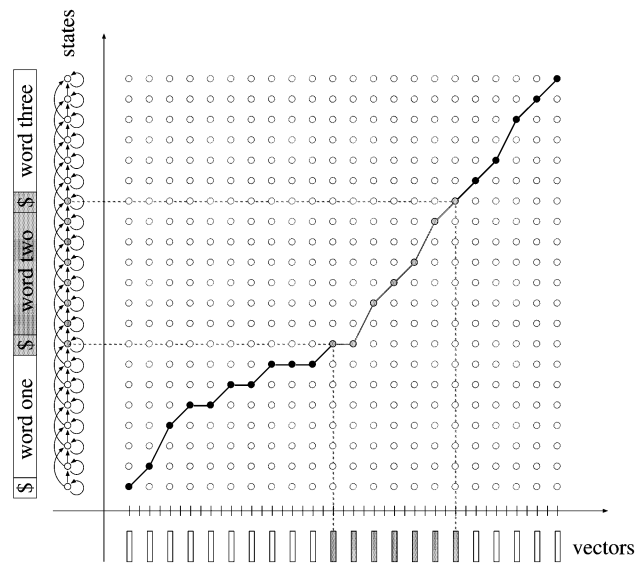

Fig. 3. Illustration of the unsupervised training. The grey color indicates words whose confidence is below a specified threshold. $ denotes the pause model.

e.g., the application of language model weights. This confidence measure is computed with a forward–backward algorithm on the basis of word-conditioned word graphs [13], [14]. All of the feature vectors corresponding to a word whose confidence is below a specified threshold are simply omitted for the estimation of the CART, of the LDA matrix, and of the acoustic models. In other words, the feature vectors of all presumably incorrect words are ignored. It is important to note that the feature vectors which are omitted cannot be determined in a preprocessing step. In order to determine these vectors, a particular time alignment between the HMM states and the acoustic feature vectors has to be used. Since this alignment changes during the training process, the word boundaries of the words also change. The set of vectors which is used to train the acoustic models is thus dynamic.

In the context of these training procedures, two different questions arise. First, the effect of the quality of the initial acoustic models (used to recognize the untranscribed training data) on the final recognition accuracy is unclear. In other words, the question has to be addressed what the impact of errors in the recognized transcriptions is. To this end, the recognition of the untranscribed training corpus is performed with different sets of acoustic models and the experimental results are compared. Second, the effect of the confidence measures which are used to reduce the number of incorrectly recognized words in the transcriptions on the final recognition accuracy has to be studied. For this purpose, the training of the final acoustic models is performed with different tagging thresholds, i.e., different amounts of training material which are most presumably correct are retained in the training corpus.

During all of the following experiments, the trigram language model which was trained on the official LDC language model training corpus for the HUB4 evaluation was left unchanged since the main concern of the following experiments was to study the training of acoustic models on recognized transcriptions. In addition, this experimental setup can easily be motivated since language model training material can be collected from the internet at very low cost even for new languages and

TABLE I
WORD ERROR RATES ON THE BROADCAST NEWS'96 EVALUATION CORPUS FOR DIFFERENT ACOUSTIC MODELS. THE WORD ERROR RATE OF THE RECOGNIZED TRANSCRIPTIONS WAS 32.5% AND THE PHONEME ERROR RATE 17.3%. #dns DENOTES THE NUMBER OF GAUSSIAN DENSITIES

		corp.		WER [%] on eval '96							
transcr.	thresh.	[h]	#dns	F0	F1	F2	F3	F4	F5	FX	total
manual		72	273k	24.1	31.2	41.6	35.8	37.8	27.4	54.1	33.5
recognized		72	345k	23.3	31.0	43.4	34.6	39.5	36.4	59.5	34.1
recognized	0.5	62	335k	23.1	30.9	43.5	33.7	39.6	37.8	60.2	34.1
+	0.7	56	318k	23.5	30.4	44.3	33.9	39.4	36.1	60.4	34.1
conf. meas.	0.9	47	294k	24.0	30.5	45.7	32.8	40.5	37.1	58.5	34.3

domains, cf., [15]. The language model perplexity was 221.2 on a 72 h subset of the Broadcast News'97 training corpus which was used for training purposes, 218.1 on the Broadcast News'96 eval test set, and 213.7 on the Broadcast News'98 eval test set.

III. BOOTSTRAPPING WITH AN OPTIMIZED SYSTEM

In a first experiment, the effect of recognition errors in the automatically generated transcriptions was studied. The training corpus for all of the following experiments is a subset of the Broadcast News'97 training corpus which consists of 72 h of acoustic data (music and commercials were filtered). The segmentation of the training corpus is based on the official script files provided by NIST and was done at the Philips Research Laboratories in Aachen. A number of corrections was carried out afterwards for those sentence that could not be forced-aligned to very simple acoustic models. For all experiments, gender-independent within-word phone models were used. The testing corpus is the official Broadcast News'96 evaluation corpus [16]. It should be noted that all word and phoneme error rates in the following sections were computed directly on the recognition output.

The training corpus was transcribed automatically using a speech recognition system which was trained previously on the Broadcast News '96 training corpus with manual transcriptions. The word error rate achieved on the 72 h subset of the Broadcast News'97 training corpus was 32.5%. Using these erroneous transcriptions, the standard training procedure was used to train a within-word speech recognition system with different confidence tagging thresholds. Table I summarizes the experimental results. The first line gives the word error rates for the different conditions of the Broadcast News'96 testing corpus for the speech recognition system trained with the manual transcriptions of the 72 h subset of the Broadcast News'97 training corpus. As the second line of Table I shows, the word error rate increases by only 1.8% relative using the recognized transcriptions instead of the correct transcriptions. This result is in fact very surprising, bearing in mind that the word error rate of the transcriptions of the training corpus is 32.5%. On the F0 and F1 condition there is no degradation at all, the word error rate even decreases for the F0 condition. In this particultar experimental framework, the word error rate cannot be reduced with our confidence measure. Only the amount of training data and thus the number of Gaussian densities can be reduced with confidence measures without any loss in performance.

These results can be attributed to two opposed effects: if the recognizer used to transcribe the data is trained on large amounts of material as in the experiments above, most of the incorrectly recognized words in the transcriptions will be acoustically very similar to the words originally spoken. The negative impact of these errors is thus only small since the acoustic models are defined on a phonetic level. Confidence measures cannot improve the performance since they do not only exclude words from the training which might be erroneous but since they also reduce the amount of training material for the acoustic models. The trade-off between these two effects is an obvious explanation for the above results.

In order to accumulate further evidence for these assumptions, the phoneme error rate (to be more concise, a lower bound of the phoneme error rate) on the training corpus was computed. Since the manually constructed reference transcriptions come without pronunciation variants, a phoneme graph of all possible pronunciation variants of the manual transcription was aligned with the phonetic transcription of the recognized sentence which may contain pronunciation variants. For the computation of the Levenshtein–Alignment, an algorithm was used which was originally implemented to compute the word graph error rate, a lower bound for the word error rate which can be achieved with a given word graph, cf., [14] for details. While the word error rate on the 72 h subset of the Broadcast News'97 training corpus is 32.5% (5.5% deletions, 4.7% insertions, and 22.3% substitutions), the phoneme error rate is roughly half as high with only 17.3%, of which 4.7% are deletions, 3.4% insertions, and 9.2% substitutions. This outcome supports the hypothesis that the recognition result is phonetically rather similar to the originally spoken words despite the high word error rate.

Another open question is how silence frames, i.e., feature vectors which are aligned to the pause model, should be treated if they occur before or after words which are tagged as incorrect. In principle, it is unclear whether these feature vectors should be used to train the silence model or not. In [1], silence frames are excluded from the training of the silence model if they are adjacent to words which are tagged as incorrect. In order to study the effect of silence frames, all silence frames between two words tagged as incorrect were excluded from the training in a first step. In a second step, all silence frames adjacent to words which were tagged as incorrect were omitted. Table II summarizes the

TABLE II
WORD ERROR RATES ON THE BROADCAST NEWS'96 EVALUATION CORPUS WHEN OMITTING SILENCE FRAMES ADJACENT TO WORDS THAT WERE TAGGED AS INCORRECT. THE TAGGING THRESHOLD WAS 0.5 IN THIS EXPERIMENT

	corp.		WER [%] on eval '96							
silence omitted	[h]	#dns	F0	F1	F2	F3	F4	F5	FX	total
none	62	335k	23.1	30.9	43.5	33.7	39.6	37.8	60.2	34.1
between	60	310k	23.3	30.2	44.5	33.9	40.2	36.1	60.8	34.1
adjacent	56	328k	23.2	30.7	44.6	32.7	40.2	34.4	60.4	34.1

TABLE III
WORD ERROR RATES ON THE BROADCAST NEWS'96 EVALUATION CORPUS USING THE BROADCAST'97 TRAINING CORPUS FOR AUGMENTING THE BROADCAST NEWS'96 TRAINING CORPUS

				WER [%] on eval '96							
corpus	transcr.	[h]	#dns	F0	F1	F2	F3	F4	F5	FX	total
train '96	manual	96	318k	23.6	31.8	40.9	32.6	37.7	28.2	54.5	33.2
+ train 97	manual	168	367k	22.7	20.7	38.9	32.3	37.2	30.9	50.5	32.0
+ train 97	recog.	168	398k	22.8	21.1	38.9	32.4	37.2	31.2	51.6	32.3

experimental results. There is only a negligible gain in recognition accuracy on some of the conditions (four errors in total on the complete training corpus) in case of omitting all feature vectors which were aligned to the silence model. Although the improvements are statistically insignificant, this approach was defined as the standard method and was used for all of the following experiments.

In summary, the experimental results show that the word error rate hardly increases in comparison with the correct transcriptions if the speech recognition system is trained on transcriptions which were recognized with an optimized system. Since the negative effect of recognition errors was smaller than expected, the Broadcast News'97 training corpus was used with the automatically recognized transcriptions to augment the manually transcribed Broadcast News'97 training corpus and to train new acoustic models. For comparison, the Broadcast News'96 training corpus was also augmented with the Broadcast News'97 training corpus and the original, manually generated transcriptions. Table III shows the experimental results. The first line contains the word error rates achieved on the Broadcast News'96 evaluation test set with a system trained on the manual transcriptions of the Broadcast News'96 training corpus. For the second line, this training corpus was augmented with the Broadcast News'97 training corpus and the corresponding manual transcriptions. For the third line, the Broadcast News'97 training corpus was used with the automatically recognized transcriptions to augment the Broadcast News'96 training corpus. In this case, the confidence measures were not used since there was almost no effect as shown in Table I. As the experiments show, the total word error rate is reduced from 33.2% in the first line to 32.0% in the second line, a relative reduction of 3.6%. Using recognized transcriptions instead of the correct ones for the Broadcast News'97 training corpus, the total word error rate is reduced to 32.3%, a relative reduction of 2.7% in comparison with the first line. The loss in performance from 32.0% to 32.3% which can be attributed to the recognition errors in the transcriptions is thus only 0.9%. As the experiment clearly shows, the automatically transcribed training corpus can be used successfully to augment an already existing training corpus and to reduce the word error rates on the testing corpus.

IV. BOOTSTRAPPING WITH A LOW-COST SYSTEM

The scenario for the following experiments is as follows: it is assumed that the 72 h of the Broadcast News'97 training corpus are not transcribed, but chopped into suitable audio segments. It is also assumed that no initial acoustic models, no initial phonetic CART, and no initial LDA matrix are available which might be used to recognize the training corpus as in the previous section. On the other hand, it is assumed that a language model is available. The same language model as in the previous experiments was used, cf., page 6.

In such a scenario, it appears to be straightforward to transcribe a small amount of the training corpus manually, to train a speech recognition system on this small corpus and to generate transcriptions for the rest of the training corpus automatically. In this experimental framework, it is left open what part of the data should be transcribed manually. For the following experiments, the subset of the training corpus which has to be transcribed manually is defined as the set of training utterances, i.e., acoustic segments, with a duration of less than a specified maximum.[1] The motivation for this particular definition was that a linear time alignment of the feature vectors and the HMM states has to be used initially if no better acoustic models are available. Experiments have shown, that the linear alignment path is a too crude approximation to the optimal alignment between the acoustic feature vectors and the HMM states if the acoustic segments are very long. Also, the effect of possible transcription errors which may occur even in manually generated transcriptions and which might jeopardize the alignment, can be aggravated.

[1]As before, the segmentation of the training corpus is based on the official script files provided by NIST.

TABLE IV
SIZES OF THE DIFFERENT SUBSETS OF THE BROADCAST NEWS'97 TRAINING CORPUS FOR DIFFERENT MAXIMUM SEGMENT LENGTHS AND ERROR RATES ACHIEVED WITH THE ACOUSTIC MODELS TRAINED ON THESE SUBSETS FOR THE COMPLETE 72 H OF THE BROADCAST NEWS'97 TRAINING CORPUS AND THE BROADCAST NEWS'96 EVALUATION CORPUS

dur.	corp.		errors [%] on train '97		WER [%] on eval '96							
[s]	[h]	#dns	WER	PER	F0	F1	F2	F3	F4	F5	FX	total
2	1.2	9k	67.5	44.0	67.9	66.9	82.9	75.5	71.5	74.2	81.2	71.3
4	3.1	23k	51.1	30.3	48.5	49.0	72.1	56.8	54.8	62.5	73.7	54.9
6	5.6	40k	43.6	24.7	39.4	42.5	64.8	47.9	48.6	50.5	68.1	47.5

TABLE V
EXPERIMENTAL RESULTS FOR INITIAL ACOUSTIC MODELS TRAINED WITH 1.2 H OF MANUALLY TRANSCRIBED MATERIAL. THE WORD ERROR RATE OF THE RECOGNIZED TRANSCRIPTIONS WAS 67.5% AND THE PHONEME ERROR RATE 44.0%

type of		corp.		WER [%] on eval '96							
transcr.	thresh.	[h]	#dns	F0	F1	F2	F3	F4	F5	FX	total
none		1.2	9k	67.9	66.9	82.9	75.5	71.5	74.2	81.2	71.3
recognized		72	440k	39.2	44.2	67.8	48.8	51.8	54.8	74.5	49.9
recognized	0.70	23	320k	35.1	39.2	65.0	43.5	48.9	47.2	71.7	45.2
+	0.90	14	251k	34.0	38.1	60.4	41.6	47.1	48.8	73.4	44.0
conf. meas.	0.95	11	218k	36.3	40.0	61.3	42.6	48.4	48.2	72.2	45.5
corr. recognized		28	346k	29.5	35.3	55.1	37.1	44.1	48.2	67.1	40.0
manual		72	273k	24.1	31.2	41.6	35.8	37.8	27.4	54.1	33.5

Assuming that such a subset of the training corpus was transcribed manually, the acoustic feature vectors and the HMM states are aligned linearly in order to estimate the parameters of single-density monophone HMMs. These models are then used to compute a better initial time alignment for the first training phase and to start the unsupervised training procedure. In the following, experimental results are presented for different maximum durations of the audio segments, i.e., for different sizes of the resulting initial hand-transcribed subset of the training corpus, and for different confidence tagging thresholds.

The second column of Table IV gives the size of the training corpus subsets for the different maximum segment lengths, given in the first column. These subsets of the training corpus were transcribed manually (the official transcriptions of the Broadcast News'97 training corpus were used to simulate the process of manual annotation) and acoustic models were trained with these transcriptions with the standard training procedure. Table IV also contains the word and phoneme error rates on the Broadcast News'97 training corpus which were achieved with the different initial acoustic models. As the experiments show, the word error rates on the training corpus are 34% to 108% higher (relative in comparison with a word error rate of 32.5% achieved with an optimized speech recognition system). The word error rates on the different conditions of the Broadcast News'96 evaluation corpus are also very high. Despite the high word error rates, the remaining parts of the training corpus which were not transcribed manually were used with the recognized transcriptions to augment the small portion of manually transcribed training material. The complete training corpus was then used to train new acoustic models.

Table V contains the experimental results for the unsupervised training procedure initialized with the system trained on only 1.2 h of speech. As the experiments show, the word error rate can be reduced by 30% relative from 71.3% to 49.9% using the recognized transcriptions for the previously untranscribed parts of the training corpus in combination with the manually generated transcriptions for the 1.2 h of training data. The word error rate of 49.9% is still very high in comparison with 33.5% achieved with the manual transcriptions of the 72 h of training data. On the other hand, only 1.2 h of the training corpus had to be transcribed manually. The manual effort of transcribing could thus have been reduced drastically in a real-world application where no manual transcriptions would have been available. Table V also shows that the additional use of confidence measures reduces the word error rate to 44.0%, a further relative reduction of 11.8%. The total relative reduction in comparison with the system trained on 1.2 h of data is thus 38.3%. Although the word error rates on the testing corpus are very high, the experiments show the potential of the unsupervised training method. The maximum possible reduction of the word error rate which could be achieved ideally with confidence measures is presented in the second-last line of Table V. 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.

Table VI contains the experimental results for the unsupervised training procedure initialized with 3.1 h of manually annotated speech. The conclusions which can drawn from

TABLE VI
EXPERIMENTAL RESULTS FOR INITIAL ACOUSTIC MODELS TRAINED WITH 3.1 H OF MANUALLY TRANSCRIBED MATERIAL. THE WORD ERROR RATE OF THE RECOGNIZED TRANSCRIPTIONS WAS 51.1% AND THE PHONEME ERROR RATE 30.3%

type of transcr.	thresh.	corp. [h]	#dns	WER [%] on eval '96 F0	F1	F2	F3	F4	F5	FX	total
none		3.1	23k	48.5	49.0	72.1	56.8	54.8	62.5	73.7	54.9
recognized		72	454k	30.2	37.9	57.9	39.2	44.6	44.5	68.8	41.7
recognized	0.70	35	393k	28.3	34.9	55.4	37.0	43.6	44.1	66.5	39.4
+	0.90	25	351k	28.8	34.6	54.8	36.0	43.2	40.1	66.3	39.2
conf. meas.	0.95	21	327k	28.8	35.4	55.6	36.8	44.4	42.5	66.9	39.8
corr. recognized		40	402k	25.9	32.6	47.9	34.5	39.9	44.5	60.9	36.1
manual		72	273k	24.1	31.2	41.6	35.8	37.8	27.4	54.1	33.5

TABLE VII
EXPERIMENTAL RESULTS FOR INITIAL ACOUSTIC MODELS TRAINED WITH 5.6 H OF MANUALLY TRANSCRIBED AUDIO MATERIAL. THE WORD ERROR RATE OF THE RECOGNIZED TRANSCRIPTIONS WAS 43.6% AND THE PHONEME ERROR RATE 24.7%

type of transcr.	thresh.	corp. [h]	#dns	WER [%] on eval '96 F0	F1	F2	F3	F4	F5	FX	total
none		5.6	40k	39.4	42.5	64.8	47.9	48.6	505.5	68.1	47.5
recognized		72	460k	27.1	34.4	56.2	36.5	42.7	40.1	66.9	38.8
recognized	0.50	51	430k	26.4	33.3	53.4	35.4	41.3	41.1	65.6	37.6
+	0.70	42	416k	26.0	32.0	52.9	34.2	41.1	38.1	64.8	36.8
conf. meas.	0.90	32	388k	26.5	33.4	55.5	34.6	41.7	39.1	61.7	37.4
corr. recognized		46	421k	25.3	31.4	47.2	33.4	40.4	39.1	60.4	35.3
manual		72	273k	24.1	31.2	41.6	35.8	37.8	27.4	54.1	33.5

these results are essentially the same as for the previous set of experiments. The main difference is the reduced word error rate on the testing corpus. As the results show, the word error rate on the testing corpus was reduced from 54.9% to 41.7% with the acoustic models trained on the automatically generated transcriptions of the complete training corpus. This amounts to 24.0% relative. Using the confidence measure to detect possible recognition errors, the word error rate decreased to 39.2% for the optimal tagging threshold, a further improvement of 6.0% relative which is significantly less than for the experiments presented in Table V. The maximum possible reduction of the word error rate as presented in the second to last line is also smaller. Obviously, the gain which can be achieved with confidence measures is higher for higher initial word error rates, i.e., for systems trained with less hand-transcribed training material and with more recognition errors.

The last set of experiments with 5.6 h of manually transcribed training data is summarized in Table VII. In this experimental setup, the word error rate was reduced from 47.5% to 38.8% using all of the 72 h of training data. The error rate was thus reduced by 18.3% relative. The use of the confidence measure further reduced the word error rate to 36.8%. As the three previous tables clearly show, the relative reduction of the word error rate becomes the smaller, the larger the amount of initial hand-transcribed training material is. The additional relative reduction which can be achieved with the confidence measure also decreases for increasing sizes of the training corpus.

In summary, it can be noted that the speech recognition system trained in the setup described in Table VII performs quite well in comparison with a word error rate of 33.5% achieved with the complete manually transcribed training corpus. Still, it would be desirable to further reduce the manual effort. Since the quality of the best system in Table V which was trained on recognized transcriptions is comparable with the system initially trained with only 5.6 h of manually transcribed training material, it is straightforward to repeat the process of unsupervised training with the best system in Table V, i.e., to use these acoustic models to recognize the training corpus again and to train new acoustic models on these new automatically generated transcriptions. Ideally, the performance of such a system should be comparable to the system initially trained on 5.6 h. This approach is discussed in the following section.

V. ITERATIVE APPLICATION OF THE UNSUPERVISED TRAINING

As the experimental results show, a speech recognition system can be trained with very little manually transcribed training material if the system is used to recognize a larger, previously untranscribed training corpus which can then be used to train improved acoustic models. In order to keep the manual effort as low as possible, the process of unsupervised training should be repeated iteratively in order to increase the quality of the recognized transcriptions of the training corpus

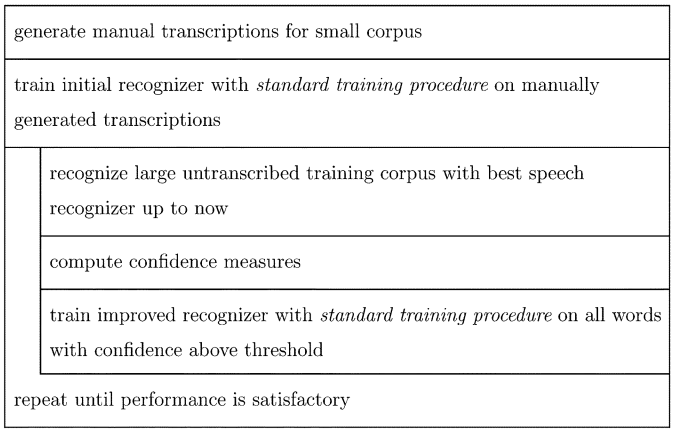

Fig. 4. RWTH iterative training procedure.

and to decrease the word error rate on the testing corpus in a (hopefully) converging process.

In order to achieve these goals, the manually transcribed 1.2 h of the training corpus were used to recognize the remaining 70.8 h. Using these transcriptions, new acoustic models were trained with the unsupervised training procedure in combination with the confidence measure. These acoustic models were then used to generate new transcriptions. This process was repeated several times. At each step, the word error rate of the transcriptions was reduced and the quality of the resulting acoustic models was increased, which were then used to recognize the untranscribed part of the training corpus again. The unsupervised training procedure is illustrated in Fig. 4 and was was repeated six times on the 72 h subset of the Broadcast News'97 training corpus.

The recognized transcriptions were used each time to train a new system which was then used to recognize the training corpus again. In order to reduce the computational effort, the iterative recognition of the training corpus was not repeated with different confidence tagging thresholds. Based on the experimental results presented in the previous section, the confidence tagging threshold was set to 0.7 during all iterations and was not changed at any time. As the left part of Table VIII shows, the word error and the phoneme error rates were reduced in every iteration. The amount of training material, i.e., the number of words with a confidence above 0.7, and resultingly the number of Gaussian densities also increased. Tables VIII and IX summarize the recognition results for the Broadcast News eval'96 test set and the Broadcast News eval'98 test set [17], respectively. In each table, the first line gives the word error rates for the initial references trained on the manually transcribed 1.2 h subset of the training corpus. The baseline word error rate with the fully tuned speech recognition system is 33.5% for the eval'96 test set and 24.7% for the eval'98 test set. As both tables show, the word error rates can be reduced significantly in each step. The performance of the final system obtained after iteration six is comparable to the performance of the system trained initially with 5.6 h of manual transcribed speech data.

The computational effort which was spent for the recognition of the training corpus and the training of the speech recognizer was extraordinarily high. The parameters of the speech recognizer used for the recognition of the training corpus were adjusted so as to run the recognition in about ten times real-time. Using 25 computers in parallel for the recognition, the complete recognition process took about 1.5 days including manual work, i.e., starting and controlling the jobs. The standard training procedure took about 15 days per iteration so that the total time spent for the iterative application of the unsupervised training algorithm was about 20 weeks. It should be mentioned that the standard training procedure was sped-up algorithmically while the experiments were in progress. If all of these algorithmic improvements had been integrated into the software used for these experiments, the total time needed for the iterative application of the unsupervised training would have been about 12 weeks. Using only one Pentium III 600 MHz CPU the iterative training would have taken about five months.

In comparison with the speech recognition system trained on the manual transcriptions of the complete training corpus, the word error rates increased by 14.3% relative on the eval'96 test set and by 18.6% relative on the eval'98 test set. Despite the computational effort of about 12 weeks, the unsupervised iterative training of the speech recognizer is an interesting alternative to having to transcribe 72 h of speech manually. It should also be noted that six iterations are not really needed to obtain reasonable results. If 2 h of speech are transcribed manually in order to bootstrap the unsupervised iterative training procedure, three iterations should be sufficient so that the total computing time with the algorithmic improvements would be in the range of five weeks.

VI. Unsupervised Training of an Across-Word System

In order to investigate the usefulness of the iterative unsupervised training procedure for the training of an across-word speech recognition system, the following approach was chosen. With the computing time of at least 12 weeks in the case of a within-word system in mind, the decision was made against a new recognition of the training corpus with an across-word system in each iteration. In fact, the recognition of the untranscribed parts of the training corpus would have taken even longer than 12 weeks since a real-time factor of ten as in the case of within-word systems could hardly have been achieved without loosing too much of the desired recognition performance, as experiments performed by colleagues at the RWTH Aachen have shown. Instead, the transcriptions generated with the within-word systems during the six iterations are used. The confidence measures which are used to detect possible recognition errors are also the same as during the experiments with the within-word systems.

In a first step, the final alignment between the feature vectors and the HMM states computed during the within-word training is used to initialize the across-word training procedure. To this end, the within-word phone model mixture labels, i.e., the labels of the within-word phone HMM states, are substituted with the across-word phone model mixture labels, but the alignment itself is not changed. For the training of the CART, the LDA matrix, and the across-word phone models, the sequence of standard training steps is then used in combination with the confidence measures. For a discussion of different aspects of the across-word model training in the RWTH speech recognition system the reader is referred to [18]. It should be kept in mind

TABLE VIII
EXPERIMENTAL RESULTS FOR THE TRAIN'97 AND THE EVAL'96 TEST SET USING ITERATIVE UNSUPERVISED TRAINING

	corp.		errors [%] on train '97		WER [%] on eval '96							
iteration	[h]	#dns	WER	PER	F0	F1	F2	F3	F4	F5	FX	total
0	1.2	9k	67.5	44.0	67.9	66.9	82.9	75.5	71.5	74.2	81.2	71.3
1	23	320k	49.2	27.2	35.1	39.2	65.0	43.5	48.9	47.2	71.7	45.2
2	49	428k	44.6	24.1	29.3	36.0	58.9	38.0	44.7	45.8	70.0	40.9
3	55	441k	43.0	23.0	27.9	35.1	57.5	37.2	42.4	45.2	66.8	39.5
4	57	446k	41.9	22.3	26.5	34.1	57.3	37.5	43.2	44.5	69.4	39.1
5	58	447k	41.2	21.8	25.9	34.3	56.7	35.9	41.5	43.1	67.4	38.4
6	59	448k	-	-	26.1	34.1	56.0	36.1	41.8	42.1	66.9	38.3
manual	72	273k	0.0	0.0	24.1	31.2	41.6	35.8	37.8	27.4	54.1	33.5

TABLE IX
EXPERIMENTAL RESULTS FOR THE EVAL'98 TEST SET USING ITERATIVE UNSUPERVISED TRAINING

	WER [%] on eval '98							
iteration	F0	F1	F2	F3	F4	F5	FX	total
0	61.0	63.6	72.6	67.1	67.1	74.0	72.5	65.5
1	27.1	35.0	57.0	39.7	38.6	55.3	51.4	36.9
2	22.5	31.8	50.7	34.7	32.6	50.6	45.9	32.0
3	21.0	31.3	49.0	34.2	31.0	44.7	45.9	30.9
4	20.2	31.0	47.9	31.3	30.3	43.8	45.0	30.1
5	19.4	30.5	48.3	32.6	29.8	46.0	44.7	29.6
6	19.6	29.8	48.3	32.8	29.0	40.4	44.6	29.3
manual	16.2	26.2	35.3	28.7	25.2	34.9	36.5	24.7

TABLE X
ACROSS-WORD RECOGNITION RESULTS FOR THE EVAL'96 TEST SET USING ITERATIVE UNSUPERVISED TRAINING

	WER [%] on eval '96							
iteration	F0	F1	F2	F3	F4	F5	FX	total
1	32.7	37.3	63.8	42.1	47.3	52.5	68.5	43.2
2	25.9	33.8	55.7	36.8	40.8	38.5	65.0	37.8
3	25.1	32.7	54.5	36.6	40.1	39.5	62.5	36.8
4	23.7	31.9	52.0	34.5	40.9	36.1	63.7	35.9
5	23.5	31.8	51.9	34.3	39.5	36.8	62.4	35.5
6	24.0	32.3	52.3	34.3	39.3	35.5	63.6	35.9
manual	20.9	26.1	38.3	32.4	35.9	30.1	49.5	29.7

that the standard training procedure is explicitly aimed at reducing the possibly negative effect of the initial alignment so that the substitution of the HMM states can in fact be used to bootstrap the across-word model training.

Tables X and XI show the experimental results for the Broadcast News'96 and the Broadcast News'98 eval test set, respectively. The baseline word error rates with a fully tuned across-word system trained on the manual transcriptions of the 72 h subset of the Broadcast News'97 training corpus are 29.7% for the eval'96 test set and 21.3% for the eval'98 test set. For the small manually transcribed 1.2 h subset of the training corpus, an across-word system was not trained because of the very limited amount of acoustic training material which would not have been sufficient to reliably estimate the CART, the LDA matrix, and the parameters of the acoustic across-word models.

TABLE XI
ACROSS-WORD RECOGNITION RESULTS FOR THE EVAL'98 TEST SET USING ITERATIVE UNSUPERVISED TRAINING

	WER [%] on eval '98							
iteration	F0	F1	F2	F3	F4	F5	FX	total
1	25.0	33.6	53.6	36.2	36.3	48.1	49.8	34.8
2	18.4	28.3	45.5	28.5	27.9	40.9	44.6	28.0
3	17.9	27.7	46.7	29.6	26.9	40.04	43.6	27.4
4	16.7	27.2	44.0	26.8	26.4	38.7	41.5	26.3
5	17.7	27.0	43.2	30.0	25.6	37.0	42.5	26.6
6	16.2	26.3	43.5	28.2	26.4	39.1	42.2	26.2
manual	13.7	21.8	32.7	21.2	22.2	26.4	33.1	21.3

As both tables show, the word error rates can be reduced significantly in each iteration as in the case of within-word systems. Comparing the above results with the word error rates achieved with the within-word systems, cf., Tables VIII and IX, it becomes obvious that the relative reduction in word error rate from within-word to across-word systems gets the smaller, the worse the quality of the recognized transcriptions is. An obvious explanation for this effect is that, although some of the recognition errors can be detected with the confidence measures, these errors still have an effect on the neighboring words. The left context of the starting phone or the right context of the ending phone of a word might thus be wrong although the word itself was recognized correctly.

In conclusion it can be noted that the word error rates increased by 20.9% relative on the eval'96 test set and by 23.0% relative on the eval'98 test set in comparison with the across-word system trained on the manual transcriptions of the complete 72 h subset of the training corpus. The deterioration of the

word error rates is slightly higher than in the within-word case, cf., Tables VIII and IX. Again, the explanation is that recognition errors may have a negative effect on the across-word contexts, even if the error is detected as such by the confidence measure. As the experiments clearly show, the iterative unsupervised training procedure can not only be used to automatically train within-word speech recognition systems but also across-word systems.

VII. Conclusions and Outlook

In this paper, 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. 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. Using the suggested method, a within-word system was initialized with only one h 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 an across-word recognizer the word error rates increased by 20.9% relative and 23.0% relative, respectively.

With the unsupervised training procedure, the manual expenditure of transcribing speech data can be reduced drastically for new application scenarios.

References

[1] G. Zavaliagkos and T. Colthurst, "Utilizing untranscribed training data to improve performance," in *Proc. Broadcast News Transcription and Understanding Workshop*. San Mateo, CA, Feb. 1998, pp. 301–305.

[2] M. Siu, H. Gish, and F. Richardson, "Improved estimation, evaluation, and applications of confidence measures for speech recognition," in *Proc. ISCA Eur. Conf. Speech Communication Technology*, Rhodes, Greece, Sep. 1997, pp. 831–834.

[3] T. Kemp and A. Waibel, "Unsupervised training of a speech recognizer: Recent experiments," in *Proc. ISCA Eur. Conf. Speech Communication Technology*, Budapest, Hungary, Sep. 1999, pp. 2725–2728.

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

[5] ——, "Investigating lightly supervised acoustic model training," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Salt Lake City, UT, May 2001, pp. 477–480.

[6] ——, "Lightly supervised and unsupervised acoustic model training," *Comput., Speech, Lang.*, vol. 16, pp. 115–129, 2002.

[7] F. Wessel and H. Ney, "Unsupervised training of acoustic models for large vocabulary continuous speech recognition," in *Proc. Automatic Speech Recognition Workshop*, Madonna di Campiglio, Trento, Italy, Dec. 2001.

[8] H. Ney, "Acoustic modeling of phoneme units for continuous speech recognition," in *Proc. Signal Processing V: Theories and Applications, 5th Eur. Signal Processing Conference*, L. Torres, E. Masgrau, and M. A. Lagunas, Eds, Barcelona, Spain, 1990, pp. 65–72.

[9] R. Schwartz, Y.-L. Chow, O. Kimball, S. Roukos, U. Krasner, and J. Makhoul, "Context-dependent modeling for acoustic-phonetic recognition of continuous speech," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Tampa, FL, Mar. 1985, pp. 1205–1208.

[10] H. Ney and A. Noll, "Phoneme modeling using continuous mixture densities," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, New York, Apr. 1988, pp. 437–440.

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

[12] F. Wessel. (2002, Aug.) Word posterior probabilities for large vocabulary continuous speech recognition. Univ. Techno. (RWTH) Aachen, Germany. [Online]. Available: http://sylvester.bth.rwth-aachen.de/dissertationen/2002/121/02_121.pdf

[13] P. C. Woodland, C. J. Legetter, J. J. Odell, V. Valtchev, and S. J. Young, "The 1994 HTK large vocabulary speech recognition system," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Detroit, MI, May 1995, pp. 573–576.

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

[15] T. Kemp, "ILKD, Ein automatisches Indexierungssystem für Fernsehnachrichtensendungen," Ph.D. thesis, Tech. Univ. Karlsruhe, Karlsruhe, Germany, 1999.

[16] J. S. Garofolo, J. G. Fiscus, and W. M. Fisher, "Design and preparation of the 1996 HUB-4 broadcast news benchmark test corpora," in *Proc. Speech Recognition Workshop*, San Francisco, CA, Feb. 1997, pp. 15–21.

[17] D. S. Pallett, J. G. Fiscus, J. S. Garofolo, A. Martin, and M. A. Przybocki, "1998 broadcast news benchmark test results: English and nonenglish word error rate performance measures," in *Proc. DARPA Broadcast News Workshop*, Herndon, VA, Feb. 1999, pp. 5–12.

[18] A. Sixtus and H. Ney, "From within-word model search to across-word model search in large vocabulary continuous speech recognition," *Computer, Speech, Lang.*, 2001, submitted for publication.

Frank Wessel received the Diploma degree with honors and the Dr. rer. nat. degree in computer science from the University of Technology (RWTH), Aachen, Germany in 1997 and 2002, respectively.

In 1997, he joined the Department of Computer Science of the RWTH Aachen where he was with the Lehrstuhl für Informatik VI and a Member of the speech recognition group until 2002. In February 2002, he joined Vodafone D2 GmbH, Düsseldorf, Germany. His scientific interests cover all aspects of automatic speech recognition, language modeling, confidence measures, and minimum Bayes risk decoding.

Hermann Ney (M'86) received the Diplom degree in physics from the University of Goettingen, Goettingen, Germany, in 1977 and the Dr.-Ing. degree in electrical engineering from the University of Technology, Braunschweig, Germany, in 1982.

In 1977, he joined Philips Research Laboratories where he worked on various aspects of speaker verification, isolated and connected word recognition, and large-vocabulary continuous-speech recognition. In 1985, he was appointed Head of the Speech and Pattern Recognition group. From 1988 to 1989, he was a Visiting Scientist at AT&T Bell Laboratories, Murray Hill, NJ. In July 1993, he joined the University of Technology, Aachen, Germany, as a Professor of computer science. His current interests cover all aspects of pattern and speech recognition, such as signal processing, search strategies, language modeling, automatic learning, and language translation.