Iterative Optimization of the Data Driven Analysis in Continuous Speech

T. Kuhn, S. Kunzmann, E. Nöth, S.Rieck, E. Schukat-Talamazzini

Lehrstuhl für Informatik 5 (Mustererkennung),
Friedrich-Alexander-Universität Erlangen-Nürnberg,
Martensstraße 3,
8520 Erlangen, F.R. of Germany

Abstract: We present an iterative method to optimize the word recognition rate for a data driven analysis in continuous speech by using a large set of speech samples. After a short description of our system environment a bootstrapping method for an iterative parameter estimation will be discussed. The initialization of the bootstrapping procedure is done by using a limited amount of hand labeled training data to estimate the statistical parameters roughly. In the second step the statistical parameters are estimated more exactly on the basis of unlabeled training data. Some experimental results for the bootstrapping method performed on unlabeled training data in comparison with results achieved by parameter estimation on labeled training data will be given.

1 Introduction

The acoustic phonetic decoding in speech recognition systems is often done by statistical methods like a *Bayes Classifier* or *Hidden Markov Models* (HMM) [6] which requires a large amount of labeled training data at the various linguistic levels. This means that the training data have to be hand labeled by an expert. This is a very time consuming process. Therefore the enlargement of a speech database is expensive. On the other side a large amount of training data is necessary for a robust parameter estimation. But there are further reasons why hand labeled speech samples might be necessary:

- A hand labeling at the various linguistic levels is needed for the evaluation of the recognition algorithms, e.g in our system environment the word recognition rate is computed on the basis of a word hand labeling. Such an evaluation criterion is useful if one wants to locally optimize the word recognition module and in the case of spontaneous speech with hesitations and interruptions. In the latter case the mostly used criterion 'word accuracy' is not appropriate.

- In an investigation that is currently being done at our institute we want to find out how well the position of the focal accent can be predicted with intensity. However it is not clear over which part of the speech signal (syllable nucleus, syllable, word, or phrase in focus) the intensity should be computed. Thus, a hand labeling of the utterances is necessary for this kind of basic research [5].

2 System Environment

Our data driven recognition of a continuously spoken utterance is part of the speech understanding and dialogue system EVAR [4] and can be divided into the following steps:

- Feature extraction and phone component labeling.
- Phone segmentation and classification.
- Generation of word hypotheses.
- Search for constituents with a context free grammar.

In the feature extraction module the speech signal is partitioned into consecutive frames of 12.8/10 msec., sampled with 12/16 Khz, and quantized with 12/16 bit. For each frame a feature vector with 9 cepstral components and 2 temporal derivatives is computed. Each frame is classified into phone components with a *Bayes Classifier* and labeled with the best n choices ($n \leq 5$).

The phone segmentation and classification is divided into four steps which are described in the following [1, 7]:

1. An initial partition is generated by searching for homogeneous ranges in the speech signal using the first alternative of the phone component stream. All frames labeled with the same phone component are grouped together to segments which are normally shorter than a phone because of the coarticulatory phenomena. The segment stream is carefully smoothed in order to reduce the number of potential segments. For instance, frames whose left and right neighbors have the same first alternative are eliminated.

2. In order to recognize plosives and diphthongs which consist of different phone components and in order to take into account classification errors a segment graph is generated by considering alternative endpoints for each starting point. The set of alternative endpoints is given by the initial partition. The size of the segment graph is restricted by limiting the maximum length of a phone. Initial segments with a length greater than the maximum phone length are not partitioned (e.g. periods of silence).

3. Each edge of the segment graph is compared with all phones of an inventory using discrete density HMM's. The result is a scored segment graph. Currently we distinguish between 44 phones which are modeled by 59 phone component HMM's.

4. Using a heuristic search procedure (A^*-Algorithm) the global optimal path is found in the scored segment graph. The output is a linear sequence of segments where each segment is labeled with the five best scoring phone classes.

The stream of phone hypotheses is taken as input data for the generation of word hypotheses with discrete density HMM's [1, 2]. A subword unit approach is used to reduce the amount of training data. All phones are modeled by the same edge oriented elementary HMM. In figure 1 alternative elementary HMM's are shown. The model *SID* was chosen for

the experiments. The word model is generated automatically by replacing each phone in the word by the corresponding HMM. In the recognition process a set of word hypotheses is generated via a word spotting technique (*vertical summation* [8]). The prefix equivalence of two consecutive word models is used to reduce the computation time.

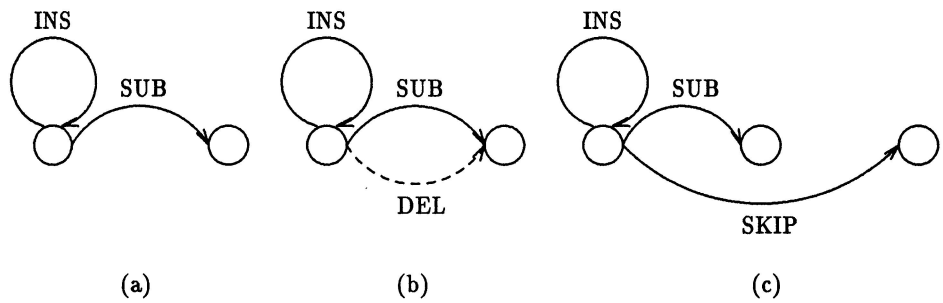

Figure 1. alternative HMM's for subword modeling: *SI* (a), *SID* (b), *SIS* (c)

Finally, a heuristic search procedure (A^*-Algorithm) is used to find word chains (constituents or sentences) in the lattice of word hypotheses. The search space is restricted by a context free grammar. Two word hypotheses are chained if they are adjacent on the time axis and syntactically correct. For the computation of a new score the word chains are modeled by HMM's on the phone level and compared with the overlapped part of the speech signal.

3 Parameter Estimation without Handlabeling

One of the main efforts in speech recognition systems is the collection of large speech databases for statistical parameter estimation because the system performance increases with more available training data [9]. Even more training data is needed for context-dependent phone models (e.g diphones, triphones) or syllable models in the recognition process [3]. Therefore we have made a large effort in expanding our speech database. For each utterance we need a hand labeling

- into phone components for the training of the *Bayes Classifier*,
- into phones for the training of the phone component HMM's and
- into words which we use for the training of the phone HMM's as well as for the computation of the word recognition rate.

Because the process of hand labeling speech is a very time consuming process a procedure for deriving a automatically a symbolic description (hand labelling) between the speech samples and the recognition units as well as a parameter training without any kind of hand labeling was developed.

3.1 Step by Step Coarsening of the Handlabeling

First we investigated the influence of a hand labeling into phones, words, or utterances on our HMM parameter training. In all of the following experiments the frame classification was done with the same *Bayes Classifier* trained with a hand labeling into phone components. The experiments were carried out with the EVAR speech sample for which a hand labeling with respect to phone components, phones and words is available. The EVAR speech sample consists of 243 utterances (about 11 min) of 6 speakers (3 female and 3 male) which was divided into a training set of 128 utterances (about 5.5 min.) and a test set of 115 utterances (about 5.5 min.). We have carried out experiments of parameter training using

- the phone labeled training data (*phone*),
- the word labeled training data (*word*) and
- the word transcription without reference to the speech signal (*sentence*).

The evaluation of the experiments was made on the word level with a lexicon of 549 words. As evaluation criterion we used the rank of the correct word hypotheses normalized by the number of segments. The results show that for the HMM parameter estimation the transcription *sentence* is sufficient because the decrease of the recognition rate can be neglected (see table 1).

	word recogniton rate in percent generating								
	1.0	*2.0*	*3.0*	*4.0*	*5.0*	*7.0*	*10.0*	*20.0*	*50.0*
	hypotheses per segment								
phone	40.7	54.4	62.5	67.3	71.6	76.9	82.3	89.5	92.7
word	41.3	52.3	59.8	65.7	70.3	74.6	81.5	89.5	92.7
sentence	42.0	52.2	58.1	63.2	68.4	74.5	79.3	88.7	92.5

Table 1. word recognition rate for step by step coarsening of the handlabeling

3.2 The Bootstrapping Procedure

Our further work was based on the training of the *Bayes Classifier* without hand labeling into phone components. Therefore we have developed a bootstrapping method which can be described as follows (see figure 2): In the initialization stage, the *Bayes Classifier* is trained with a limited amount of hand labeled speech data. In the second stage, a hand labeling into phone components is generated automatically using a HMM-based alignment procedure. This hand labeling can be used to reestimate the statistical parameters. If there is an improvement of the word recognition rate, the bootstrapping procedure stops, otherwise a new iteration begins.

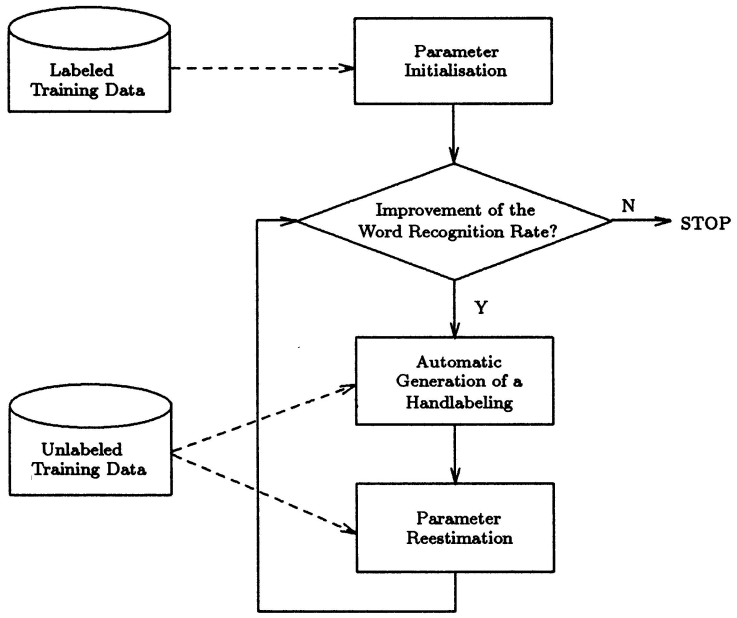

Figure 2. bootstrapping procedure

The bootstrapping procedure was tested with the SIEMENS speech sample for which only a hand labeling into phone components is available. The SIEMENS sample consists of 700 utterances spoken by 4 speakers. Each speaker spoke the SOTSCHEK corpus (100 phonetically balanced sentences) twice. The first recording session was used as training set (400 sentences, about 22 min.) The test set consisted of the remaining 300 sentences (about 16 min.) of the second recording session. For the evaluation a word hand labeling was derived automatically using the same HMM based alignment procedure as in the bootstrapping procedure.

The experiments based on a lexicon with 1300 words were carried out as follows: The parameters of the *Bayes Classifier* were initialized with 100 utterances of the EVAR speech sample (*boot_0*). The bootstrapping procedure was finished after two iterations (*boot_1*, *boot_2*), because no further improvement of the word recognition rate was observed. For comparison, we have computed the word recognition rate that can be achieved if the *Bayes Classifier* is trained with a phone component hand labeling (*hand labeled*). The results of the bootstrapping experiment (see table 2) show that the decrease of the word recognition rate is negligible especially if the high costs for hand labeling speech are taken into account. Here the same evaluation criterion was used as in section 3.1.

4 Conclusions

We have presented a method of estimating the statistical parameters without a hand labeling. At first it was shown that for the HMM parameter training the word transcription

	word recogniton rate in percent generating								
	1.0	*2.0*	*3.0*	*4.0*	*5.0*	*7.0*	*10.0*	*20.0*	*50.0*
	hypotheses per segment								
boot_0	32.7	43.6	49.9	55.1	59.7	64.8	70.0	79.1	85.8
boot_1	46.1	58.5	66.7	70.8	73.7	77.4	81.1	85.4	89.2
boot_2	48.6	60.9	67.6	72.1	75.3	78.2	80.9	86.0	89.4
handlabeled	52.3	65.0	70.2	73.0	75.8	79.1	81.3	84.6	87.5

Table 2. word recognition rate for the bootstrapping procedure

without reference to the speech signal is sufficient. Secondly a bootstrapping procedure for an automatic adaptation of the statistical parameters was presented using unlabeled training data. Only a small amount of a phone component labeling is required for the training of the Bayes Classifier in the initialization stage. The increase of the word recognition that can be achieved by using hand labeled training data is negligible.

References

[1] S. Kunzmann. *Die Worterkennung in einem Dialogsystem für kontinuierlich gesprochene Sprache.* PhD thesis, Technische Fakultät der Universität Erlangen-Nürnberg, 1990.

[2] S. Kunzmann, T. Kuhn, and H. Niemann. An Experimental Environment for Generating Word Hypotheses in Continuous Speech. In H. Niemann, M. Lang, and G. Sagerer, editors, *Recent Advances in Speech Understanding and Dialog Systems*, pages 311–316, Springer Verlag, Berlin, Heidelberg, New York, 1988.

[3] K. Lee, H. Hon, M. Hwang, and S. Majahan. Recent Progress and Future Outlook of the SPHINX Speech Recognition System. *Computer Speech & Language*, 4(1):57–69, 1990.

[4] H. Niemann, A. Brietzmann, U. Ehrlich, S. Posch, P. Regel, G. Sagerer, R. Salzbrunn, and E.G. Schukat-Talamazzini. A Knowledge Based Speech Understanding System. *Int. J. Pattern Recognition and Artificial Intelligence*, 2(2):321–350, 1988.

[5] E. Nöth, A. Batliner, and T. Kuhn. Intensity as a Predictor of Focal Accent. In *XIIème Congrès International des Science Phonétiques*, will be published in 1991.

[6] L. R. Rabiner. Mathematical Foundations of Hidden Markov Models. In H. Niemann, M. Lang, and G. Sagerer, editors, *Recent Advances in Speech Understanding and Dialog Systems*, pages 183–205, Springer Verlag, Berlin, Heidelberg, New York, 1988.

[7] A. Reißer. *Ein zeitsynchrones Segmentierungsverfahren für die Lautklassifikation mit Markov Modellen.* Technical Report, IMMD5 (Mustererkennung), Universität Erlangen-Nürnberg, 1990.

[8] E. G. Schukat-Talamazzini. *Generierung von Worthypothesen in kontinuierlicher Sprache.* Volume 141 of *Informatik Fachberichte*, Springer Verlag, Berlin, Heidelberg, New York, Tokyo, 1987.

[9] R. Schwartz and F. Kubala. Hidden Markov Models and Speaker Adaptation. In this volume.