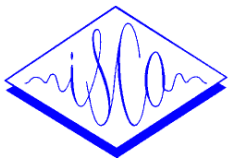

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

MAXIMUM LIKELIHOOD BASED DISCRIMINATIVE TRAINING OF ACOUSTIC MODELS[1]

Albino Nogueiras-Rodríguez and José B. Mariño
e-mail: albino@tsc.upc.es
Universitat Politècnica de Catalunya
Campus Nord D-5, C/ Gran Capità s.n.
08034 Barcelona
SPAIN

ABSTRACT

In this paper, a framework for discriminative training of acoustic models based on Generalised Probabilistic Descent (GPD) method is presented. The key feature of our proposal, Maximum Likelihood based Discriminative Training of Acoustic Models (MLDT), is the use of maximum likelihood trained HMM's instead of the original speech signal. We focus our attention in performing discriminative training applied to a discrete hidden Markov models continuos speech recogniser, achieving a 4.6% error rate reduction on a Spanish speaker-independent phoneme recognition task.

1. INTRODUCTION

Hidden Markov models, HMM's, represent the major approach to statistical modelling of speech signals for speech recognition tasks. Indeed, they provide a natural and highly reliable way of recognising speech for a wide range of applications and integrate well into systems incorporating both task syntax and semantics [1]. The underlying assumption of the HMM approach is that the observed signal can be well modelled if the parameters of the model are carefully and correctly chosen. Nevertheless, there are several problems with this approach, namely, we need to make the following assumptions: we are able to estimate the parameters of the model from a finite training set, and maximising the likelihood of the models will lead to maximum discrimination. An alternative to statistical modelling is discriminative training. In discriminative training we are no longer concerned with the correctness of the models or their likelihood whenever they lead to a maximum discriminative situation.

During the last few years, several discriminative training approaches have been proposed for acoustic modelling of speech signal [2], [3]. All of them are based upon the opposition between each of the utterances of the training set and all the models of the system. In this paper, we propose an alternative method for performing discriminative training. The key feature of our approach is to perform discriminative training using the information present in maximum likelihood trained HMM's instead of the original speech utterances. Discriminative training is carried out using the *generalised probabilistic descent* (GPD) method [3], so our system inherits many of its advantages, including its theoretically consistent formulation.

The main advantages of our method are: easier extension to continuos speech than conventional discriminative methods, little speaker adaptation, capability of *off-line* implementation and linguistical knowledge embedding for sub-lexical based semantic recognition and task orientated applications.

2. MAXIMUM LIKELIHOOD BASED DISCRIMINATIVE TRAINING OF ACOUSTIC MODELS

Consider a set of training samples $\varphi = \{x^1, x^2, ..., x^N\}$, where each x^n is known to belong to one of M classes $C_i, i = 1,2,...M$. The task of minimum error classifier design is to find the classifier parameter set, denoted by $\Lambda = \{\lambda_1, \lambda_2, ..., \lambda_M\}$, such that the probability of misclassifying any x is minimised. However, direct minimisation of the empirical error rate function is very difficult due to its not-continuos form. In order to avoid this problem, we shall use a smooth loss function approach of it, $L(x, \Lambda)$.

GPD training, as presented in [3], is carried out via a three step procedure for defining a parametrically smoothed version of the recognition error rate function: definition of a set of discriminant functions, $FD_m(x_t, \Lambda)$

[1] This research was supported by CICYT under contract TIC92-1026-C02-02

one for each of the **M** classes to be recognized; definition of a misrecognition measure based on the set of discriminant functions; and definition of a smooth 0-1 loss function based on the misrecognition measure, $L(x,\Lambda)$. The actual choice of these functions is free as long as we ensure that the loss function and its derivative approach the empirical error rate and its derivative. GPD theorem proves that if we adjust Λ according to $\Lambda_{n+1} = \Lambda_n + \partial\Lambda_n = \Lambda_n - \varepsilon_n U_n \nabla_{\Lambda_n} L(x^n, \Lambda_n)$ where U_n is a properly designed positive definite matrix, and $\{\varepsilon_n, n \geq 1\}$ is an infinite sequence of small positive numbers such that $\sum_{n=1}^{\infty} \varepsilon_n \to \infty; \sum_{n=1}^{\infty} \varepsilon_n^2 < \infty$, the algorithm converges to a Λ^* which results in a local minimum of $L(x,\Lambda)$.

In general, we can suppose that the loss function will only depend on the degree of confidence of the assignations $x_k \to C_i, i = 1,2,...,M$, where $X_k = \{x / x \in C_k\}$. Also, if we denote these functions by $FAc_i(x_k, C_i, \Lambda)$, the only dependence with the system of each of them will usually be with the model corresponding to the class we are considering, so $FAc_i(x_k, C_i, \Lambda) = FAc_i(x_k, \lambda_i)$. Finally, and also in the general case, each λ_i can be seen as a set of states where the signal can be at at any time. If we denote as $Q_i(x_k) = \{q_i(x_k^1), q_i(x_k^2), ..., q_i(x_k^T)\}$ the sequence of states along λ_i that best fits signal x_k, our assignation function can be expressed as follows:

$FAc_i(x_k, \lambda_i) = \sum_{t=1}^{T} FT_i^{q_t}(x_k^t, \lambda_i^{q_t})$

And by proper manipulation of the actualisation formula of GPD, we get

$\lambda_i^{s'} = \lambda_i^s - \varepsilon U \sum_{t=0}^{T} \int_{x_k^t} \nabla_{\lambda_i^s} l_k^{u,s}(x_k, \Lambda) \cdot f(q_i^t = s / x_k, \lambda_i) \cdot f(x_k^t) \cdot dx_k^t \cdot f(C_k)$

In this formula, the main contribution is due to the terms $FT_i^s(x_k^t, \lambda_i^s)$ and $\nabla_{\lambda_i^s} FT_i^s(x_k^t, \lambda_i^s)$, so we can make the following approximations

$FAc_i(x_k, \lambda_i) \approx \overline{FAc_i^t}(x_k, \lambda_i) = E\{\sum_{t'=0}^{t-1} FT_i^{q_{t'}}\} + FT_i^{q_t} + E\{\sum_{t'=t+1}^{T} FT_i^{q_{t'}}\}$

$l_k(x_k, \Lambda) \approx \overline{l_k^t}(x_k, \Lambda) = l_k(\overline{FAc_i^t}(x_k, \lambda_i)), i = 1,2,...,M$

leading to

$\lambda_i^{s'} = \lambda_i^s - \varepsilon U \sum_{t=0}^{T} \int_{x_k^t} \nabla_{\lambda_i^s} \overline{l_k^{u,s}}(x_k, \Lambda) \cdot f(q_i^t = s / x_k, \lambda_i) \cdot f(x_k^t) \cdot dx_k^t \cdot f(C_k)$

The advantage of this last expression is that it can be calculated with the only knowledge of the state sequence that best suited the signal along its class maximum likelihood trained HMM. What is more, if we have HMM's accurate enough, we need the actual signal no more because we can generate sequences of states via Monte-Carlo method and employ them instead of the actual speech samples. It must be remarked that we are no longer concerned with the ability of these models, in the following *generator HMM's*, to discriminate between the different classes but with their ability to *remember* the training utterances. And this ability is exactly what we maximise when we train them by applying maximum likelihood criterion. What is more, we can get generator HMM's as accurate as needed by increasing their freedom degrees, (number of states, transitions between them, etc.), and, at a certain point, they will behave as finite states machines able to reproduce exactly each training utterance. In this case both conventional GPD and MLDT will behave the same.

2.1 Adjust of the Loss Function to the Confusion Matrix.

As was stated before, GPD relies in a three step procedure in order to define a smooth 0-1 loss function. The actual choice of this function is free, but it has some restrictions, namely: it must be smooth, i.e. it must have finite derivative; and its derivative must be similar to that of the empirical error rate. In conventional GPD training, this is accomplished by using the same discriminant functions employed during the actual recognition phase and substituting the hard decision rule by a sigmoid. This technique leads to a cost function very similar to the empirical error rate and that tends to it when the parameters of the sigmoid tend toward their degenerated values. In the method proposed in this paper, this substitution is not enough due to the fact that the discriminant functions are also approximations of that employed during the recognition phase. As a result, if we make no corrections, the cost function we obtain is very different from the actual empirical error rate. Yet, its derivative is sufficiently accurate so as that direct application of the method provides a significant improvement in the overall performance of the system, some 1% error rate reduction in a Spanish phoneme recognition task. Nevertheless, we can still improve the performance by substituting the conventional sigmoid by a biased sigmoid and choosing the bias factor in order that the loss function adjusts the actual empirical error rate. By doing so, we obtained a 4.6% reduction in this task.

2.1.1 Confusion Matrix Evaluation.

Conventional confusion matrix evaluation for continuos speech relies on dynamic programming algorithms. Although this technique provides a quite good approximation of the overall empirical error rate, it falls whenever we are interested in a detailed confusion matrix. This is due to the fact that dynamic programming algorithms search for the best path, either minimising error count or maximising goal count, with no temporal restriction. This leads to a situation where two units are considered to be confused or correctly recognised even if they occur at very different times. While, in general, this way of evaluating the confusion matrix led to good results when applied to discriminative training, we discovered that it also led to some results that were absolutely absurd: some

confusions that are very unlikely phonetically appeared frequently, while others that were more likely to do did not appear at all. In order to adjust the loss function to a more reliable confusion matrix, we have applied a modified dynamic warping algorithm with temporal restrictions that do not allow a recognised unit to be considered confused or correct if it does not overlap in time with the actual unit. This strategy leads to worse overall results, really it should be said that conventional algorithms tend to inflate results, but it also leads to more significant improvement when the confusion matrix is employed in models reestimation.

It must be stated that this way of evaluating the confusion matrix was also employed in a genetic algorithm used for choosing topological parameters in a HMM based recognition system, leading to better results than conventional methods as well.

3. FEATURES OF MAXIMUM LIKELIHOOD BASED DISCRIMINATIVE TRAINING

3.1 Extension to Continuos Speech

The basis of conventional discriminative training is the confrontation between a known speech utterance and the acoustic models that represent its phonetic unit and all the rest of units. This confrontation is possible in isolated word recognition systems because we know exactly where the utterance starts and ends and that this utterance must belong to any of the M acoustic classes we consider, that is: we know exactly what the segmentation is. In connected word recognition systems, we can still make the assumption that the segmentation induced by the models is correct even in the case that we have an error. This is equivalent to suppose that the only errors we make are substitution errors. In continuos speech, this assumption can no longer be done: we can not confront one single speech utterance with all the acoustic models because we do not know where the utterance starts and ends. What is more, we can not even use the segmentation induced by the models themselves because it is different for every model, and so should be the utterance.

With the method proposed in this paper happens something similar. Nevertheless, as the confrontation is done between two models; we make no assumption about the segmentation induced by them. Somehow, it is equivalent to confront the two phonetic units themselves. So, if applied directly to continuos speech, it should, at least, be able to reduce substitution errors which are, usually, the main contribution to overall error rate. Actually, experimental results prove that direct application of the method improves substitution rate with little or none degradation of the insertion and deletion rates. Currently we are working on several techniques in order to improve these rates as well.

3.2 Off-Line Implementation.

One of the theoretical advantages of discriminative training is that they can still be reestimated while already in work. Nevertheless, conventional methods are not well suited for this kind of work because they need to keep the signal during all the reestimation procedure and it can not start until verification of the recognised utterance is done. This unables conventional methods to deal with systems where verification is carried out after an undetermined quantity of data has been collected, say automatic dictation systems. But even in the case of systems with immediate verification, discriminative training should be carried out at recognition time, increasing significantly the hardware requirements of the system. Unlike conventional discriminative training systems, the method proposed in this paper is specially well suited for real life applications: all we must do at recognition time is to collect the confusion matrix. Discriminative training can then be done off-line, out of office hours for example, with no degradation of the performance of the system.

3.3 Little Speaker Adaptation

Another advantage of maximum likelihood based discriminative training in speaker independent real life tasks is that the system has little or none speaker adaptation. In conventional discriminative training, if a little set of speakers uses during a time the system very often, the models will learn so much from this little speakers set utterances that they can even *forget* its original speaker independence. This is due to the fact that, once we set into work the system, the only information it uses in reestimating the models is that given by the new utterances.

Although some speaker adaptation can still be expected in the proposed method, the way the reestimation is carried out is somewhat more robust: we avoid the typical errors committed by the recognition set, but the information we use in order to avoid them comes from a speaker independent database. Even in the case where a little set of speakers employed very much the system, it will not correct any error that would not be committed by the speaker independent training set.

3.4 Extensions to Sub-Lexical Based Semantic Recognition and Task-Oriented Applications

Sub-lexical based semantic recognition and task-oriented applications rely on the fact that it is much easier to get a phonetic balanced training database than a word balanced one. Indeed, it is possible to

implement a task orientated system where the training database is only balanced at the phonetic level with no occurrence of the lexicon to be recognised.

Conventional discriminative training applied to this kind of tasks is based upon the confrontation of the lexicon to be recognised, so, in order to perform discriminative training, we need a word balanced database [3]. In the proposed method there are two possible strategies in order to perform discriminative training of sub-lexical units applied to semantic recognition and task-oriented applications: construction of word models from the sub-lexical models for both the generator HMM's and the discriminative ones in a similar way to that of [3]; and linguistical knowledge embedding. This last technique, still under study, consists on estimating the effect a certain sub-lexical confusion has in the overall semantic recognition performance and introduce it on the reestimation formulae. The advantage of this technique is that this information can be obtained directly from the specific application characteristics.

3. EXPERIMENTAL RESULTS

Maximum likelihood based GPD training has been applied to a speaker independent Spanish phoneme recognition task based on Fuzzy Markov Models (MFM). A MFM is an acoustic model structurally identical to a discrete HMM's but with no stochastic constraints. At this moment, we have only an uncompleted version of the framework. Its main lacks are that it can be only applied to one information systems (mel frequency cepstral coefficients), and that no extensions are done to continuos speech recognition, so that we can only minimise substitution errors in this kind of tasks. Both generator HMM's and MFM's were 3 state left to right models. Signal was represented by 12 Mel-frequency cepstral coefficients quantified with a 120 vectors codebook. Generator HMM's were employed to initialise MFM's. We run 5 epochs of discriminative training. At each epoch the confusion matrix calculated over the training set (120 phrases each of them pronounced by 6 speakers) was used to reestimate the value of the bias parameter of the sigmoid. The results refer to the recognition set (50 sentences pronounced by 4 speakers, both the sentences and the speakers different from those of the training set). In the following table, **Subst** stands for the total number of substitution errors, **Inse**, for the number of insertion errors, **Dele**, for the number of deletion errors, **Goals** is the number of goals, and **Error** stands for the percentage of total number of errors over the total number of units present in the recognition base (7420 units in our case).

	Subst	Inse	Dele	Goals	Error
Initial models	2,334	1,562	794	4,292	63.22%
Epoch 1	2,256	1,527	793	4,371	61.68%
Epoch 2	2,233	1,580	779	4,408	61.89%
Epoch 3	2,182	1,588	782	4,456	61.36%
Epoch 4	2,145	1,592	795	4,480	61.09%
Epoch 5	2,073	1,585	823	4,524	60.40%

As can be seen MLDT leads to a considerable reduction of the number of substitution errors, about 12%, by augmenting the number of goals in a similar amount and with little degradation of the number of insertion and deletion errors. Further research is in progress in order to improve these quantities as well.

BIBLIOGRAPHY

[1] L.R. Bahl, P.F. Brown, P.V. De Souza, R.L. Mercer. "A New Algorithm for the Estimation of Hidden Markov Model Parameters". IEEE ICASP, April 1988.

[2] Kai-Fu Lee, S. Mahajan. "Corrective and Reinforcement Learning for Speaker-Independent Continuous Speech Recognition". Computer and Language, 1990.

[3] W. Chou, B.H. Juang, C.H. Lee. "Segmental GPD Training of HMM Based Speech Recognizer". IEEE, 1992.