TASK INDEPENDENT MINIMUM CONFUSIBILITY TRAINING FOR CONTINUOUS SPEECH RECOGNITION

Albino Nogueiras-Rodríguez and José B. Mariño[1]

Universitat Politècnica de Catalunya

Barcelona, Spain

albino@gps.tsc.upc.es

ABSTRACT

In this paper, a task independent discriminative training framework for subword units based continuous speech recognition is presented. Instead of aiming at the optimisation of any task independent figure, say the phone classification or recognition rates, we focus our attention to the reduction of the number of errors committed by the system when a task is defined. This consideration leads to the use of a segmental approach based on the minimisation of the confusibility over short chains of subword units. Using this framework, a reduction of 32% in the string error rate may be achieved in the recognition of unknown length digit strings using task independent phone like units.

1. INTRODUCTION

Since their first proposal in the late 1980's, discriminative training techniques have become a major trend in automatic speech recognition. Their success is due to the fact that they do not rely on the main assumption done in conventional training: the correctness of the model employed to characterise speech. For instance, maximum likelihood trained hidden Markov models based systems would lead to minimum risk classifiers just in the case that speech actually behaved as a Markov process whose parameters could be estimated from a finite amount of training data. As this is not the case, the solution obtained in this way is just suboptimal.

In order to overcome this suboptimality, discriminative training, in its various implementations, aims directly at the maximisation of the probability that the correct word is recognised by increasing the difference between the score of this one and that of the incorrect ones, i.e. the discriminative strength of the system. In order to increase this discrimination, several approaches have been proposed [2, 5]. Many of them share a common basis: for all the utterances in the training set, first the most likely confusions are estimated, through an N-Best search, and then the difference between the score of the correct transcription and that of the incorrect ones is augmented in an iterative procedure. This scheme has been proved to lead to much higher performances than those achieved by conventional training in several different tasks: isolated and connected word recognition systems, for instance.

Continuous speech recognition is a major challenge in spoken language processing. The main assumption is that speech may be seen as the concatenation of short meaningless sounds, called subword units, taken from a finite and complete set, and such that any of these units may be distinguished from all the others by means of its acoustical properties. This assumption enables us to undertake any vocabulary or grammar specific recognition task by building a model for each sentence of the task through concatenation of the models of the respective units that form it. The main asset of this approach is that we no longer need task specific training databases in order to recognise any kind of application.

Unfortunately, the extension of DT techniques to subword based continuous speech recognition systems is not as immediate as in the case of task dependent systems. The trivial solution: optimisation of a task where the lexicon is built from subword units, either isolated or connected, does not always lead to the expected results. In this paper, we propose a framework that, while maintaining its task independent nature, performs DT on subword units, leading to notorious improvement in task specific continuous speech recognition. The differences with previously published works rely on the objective function we use to optimise the system in task independent databases, leading to a minimum confusibility formulation, and the kind of utterances employed in estimating such function, leading to the use of chains of phones.

2. MINIMUM CONFUSIBILITY TRAINING

While being its most appealing feature, not relying in task dependent training databases is at same time the main difficulty in applying DT to subword based CSR systems. If we know the task in which the models will be used and dispose of task specific utterances, DT may be

[1] This work has been supported by the grants TIC 95-1022-C05-03 and TIC 95-0884-C04-02 of the Spanish government.

implemented quite straightforward, because we can get knowledge about the actual errors that may be committed when recognising the task. We know that all the utterances used during the training might be found at recognition time, and we can restrict the search of candidate errors to those allowed by the grammar of the task. When this is not the case, and either we do not know how the task is, or a task dependent database is not available, we have no guarantee that the errors we can take into account, those committed in the recognition of task independent utterances, reflect the actual behaviour of the system. One way to overcome these difficulties is focussing the discriminative training procedure to reducing the number of errors committed in a task independent recognition.

Although other approaches have also been proposed, the most popular techniques in DT rely in the definition of a continuous differentiable function that reflects somehow the discriminative strength of the system. Once defined such a function, a gradient search along it may be used to optimise the parameters of the system. Two quite different functions have been widely used for these purposes: the mutual information and the classification error. The first is an information theory derived function that reflects the class separability of the models. As it is a figure of merit of the system, the objective of DT will be its maximisation, leading to maximum mutual information estimation (MMIE) [5]. The second works by defining a smooth approximation of the empirical error rate, leading to minimum classification error (MCE) training [2]. Although they are based in rather different foundations, they share several features and result in very similar optimisation procedures. Results published so far show that the results achieved with either technique are quite the same, being slightly better for MCE formulation [3,6].

One negative feature common to both methods is the shadowing effect. This effect consists in the fact that, among all the incorrect hypothesis considered for each training utterance, the most likely will be the one the method will more heavily avoid. Moreover, if the first incorrect hypothesis is much more likely than the correct one, then no reestimation of the system will be carried out, neither with this most likely error, nor with the rest of the hypothesis. This is an undesired feature when the models are to be used in task specific recognition, because it is probable that, when the task is defined, this most likely hypothesis is not allowed by the task. As a matter of fact, when a task independent recognition is carried out, it is almost impossible that the recognition result fit the task constraints.

In order to avoid the shadowing effect, we propose an alternative formulation of the discriminative function: the empirical confusibility, leading to minimum confusibility (MC) estimation. We define the confusibility as the expected number of errors that the system might commit if the task allowed them. The individual confusibility of one utterance is the number of incorrect hypothesis that would be misrecognised. The global confusibilty of a system is the sum of the individual ones for all the utterances. The main difference with MCE is that each possible error is considered independently of the rest, so a gradient descent procedure will lead to reestimation formulae in which contributions due to different errors are also independent.

As opposed to MCE, we are not concerned with the probability that the utterance is misrecognised, but with the total number of incorrect hypothesis that are more likely than the correct one. As all the hypothesis are given the same relevance, the reestimation procedure will focus its efforts in avoiding the errors which are more easily removed. Without any prior knowledge about the task, this procedure seems more adequate than focussing in the most likely error. It should be noted that, once again, MC criterion leads to reestimation formulae that are very close to those of MCE and MMIE. Nevertheless, the decoupling between the different hypothesis in MC has an additional interest. If the task is known, we could estimate the relevance of each error with the only knowledge of its grammar. The authors are currently developing a system that uses this knowledge in order to perform task adaptation using task independent databases.

3. SEGMENTAL DISCRIMINATIVE TRAINING USING CHAINS OF SUBWORD UNITS

Independently of the form of the function we employ as a measure of the performance of the system in the recognition of the utterances, another difficulty in the application of DT to task independent continuous speech recognition is the election of the utterances themselves. Two different solutions have been proposed: whole sentence training and segmental training. The first one is equivalent to optimising the phone recognition rate. It is based in considering that the same errors that we may find in a task specific recognition will also appear in the phone recognition of a task independent database. Using whole sentences in DT has, however, one main drawback. The training material is poorly profited. This is because, as the length of the utterance grows, the number of hypotheses that differ in just some units increases dramatically. This means that just the most probable hypothesis and others very similar to it will be used, even in the case of estimating a high number of them. For instance, segments correctly recognised in the first hypothesis will probably be correctly recognised in all the rest of them, so their contribution will be void. As a result, the algorithm will avoid the commission of the most probable errors, but will not increase the discrimination versus near misses.

In order to increase the effectiveness of the training, a segmental approach has also been proposed. It consists in partitioning each training utterance in smaller segments. DT aimed to reducing the number of misrecognised segments is then performed. This scheme is founded on the fact that most errors committed in the recognition of a task usually only involve segments of the utterance. Each misrecognised utterance can be seen as the concatenation of smaller segments, each of them being either correctly recognised or not. By avoiding the commission of error on each of these segments, the number of errors when recognising the task will be reduced.

The idea of segmental training is similar in the cases of maximum likelihood and discriminative training. In the former, the assumption that each oral message can be seen as the concatenation of independent subword units, enables the training of the system to be task independent. In DT, however, the training material should no longer be the concrete utterance but the errors in which it can be involved. In order to do so, two conditions must be verified. First, we should not apply any prior assumption about which the segmentation of the utterance is, because errors committed in CSR usually involve errors in this segmentation. Second, the segments should be formed of complete strings of complete subword units, and only such strings should be allowed to be considered as incorrect hypotheses. This is so because the speaker will never utter just a piece of unit, and the system will only allow complete units to be recognised.

These two conditions are somewhat contradictory: while the first forbids the use of any prior knowledge about the segmentation, the second imposes it. As a compromise solution, the authors of this work propose the use of chains of subword units. Each sentence in a standard task independent training database is divided in overlapping segments of a reduced number of phones. As each segment is short in terms of the number of acoustic units, an N-best recognition procedure will be able to provide a high number of incorrect hypotheses even in the case that the segment would have been correctly recognised in its first hypothesis.

Also, as the segmentation is only assumed to be known at the extremes of the segment, segmentation errors, including insertions and deletions, will be allowed in the interior of the segments. Using overlapping segments relaxes the effect of the prior knowledge of the segmentation because each piece of speech will appear just twice at each extreme of the segment, while it will be present in the middle of it K−2 times, being K the length in units of the segment. Furthermore, this effect may be further reduced by forcing all the hypotheses considered to begin and end with the correct unit. As both extremes will present the same units both in the correct and incorrect hypotheses, their influence in the discriminative training procedure will be completely cancelled.

4. EXPERIMENTATION

In order to assess the advantages of MC estimation on chains of subword units to perform task independent DT, we performed a series of experiments involving the recognition of TI connected digits database using task independent subword units. Semicontinuous HMM's of context independent phones were trained with the train corpus of TIMIT task independent database, using our own framework RAMSES. Due to the different gender coverage of both databases, only male speakers were used. Signal was parameterised using 30ms frames taken each 10ms. 12 MFCC parameters were extracted out of 24 mel frequency spaced bands filter bank. First and second deltas of the MFCC's and a vector consisting of the first and second deltas of energy were added. Each information was quantified using Gaussian VQ's of 256 codewords for spectral parameters and 128 for the energy one. 4 states left-to-right SCHMM's were trained for each of the 45 phones considered. We used a slightly different phone set in order to alleviate the ambiguous transcription of plosive sounds: being P the plosive sound and PCL its corresponding closure, we folded P, PCL-P and PCL into one single class. The main implication of this procedure is that confusing two different closures, or a closure and silence, will lead to an error. By doing so we achieve a more compact representation of the lexicon of speech recognition tasks.

The baseline system was reestimated using MCE criterion over single phones, chains of 5 phones and whole sentences, and using MC criterion over chains of 5 phones. We performed two kinds of recognition experiments with all four DT frameworks and the baseline. First, we performed phone recognition of the core corpus of TIMIT, both isolated and connected. The phone classification error rate is the objective function optimised when MCE is applied to phone segments, and the phone recognition error rate, which involves not only substitutions but insertions and deletions as well, is the one optimised when whole sentences are used. We also performed digit string recognition using the test set of TI database.

Table I shows the results obtained with the four methods and the standard maximum likelihood system. Columns SEGM and CRITER define the experiment as a function of the segment considered (phone, chain of phones or whole sentence) and the criterion adopted in DT (ML: maximum likelihood baseline, MCE: minimum classification error;

MC: minimum confusibility). In column labelled ERROR, the phone error rate, with all three kinds of error given the same cost, is depicted. Column CLASS shows the percentage of misrecognised phones when the segmentation is a priori known. DIGIT and STRN stand for the digit and string error rate in the recognition of unknown length digit strings.

SEGM	CRITER	ERROR	CLASS	DIGIT	STRN
BASELINE ML		39.9	38.9	3.7	10.9
PHONE	MCE	38.7	29.5	5.4	15.0
CHAIN	MCE	30.8	31.3	2.9	8.2
SENT	MCE	31.2	32.8	3.2	9.0
CHAIN	MC	31.4	31.1	2.5	7.4

The behaviour of segmental DT using phones as basic unit is quite remarkable. It achieves the best result in phone classification, with a 25% reduction in the error rate. Nevertheless, the results obtained in the task recognition are worse than the baseline. This is due to the fact that phone classification reflects poorly the mechanisms actually involved in continuous speech recognition. The other three DT frameworks performed noticeable better. They lead to quite the same results in task independent recognition (with a 22% reduction in the phone recognition rate). Nevertheless, in the recognition of the digit strings, DT using chains of phones is clearly a better choice than whole sentences, independently of the optimisation criterion used, MCE or MC. Finally, the best result is obtained when the framework proposed in this paper is used, achieving a total reduction of 32% in the string error rate.

5. CONCLUSIONS AND FUTURE WORK

In our opinion, the results presented in this work confirm the usefulness of applying the minimum confusibility criterion over chains of phones in order to perform task independent discriminative training of subword unit models for continuous speech recognition. They not only represent some 32% string error reduction in the recognition of digit strings, but also outperform different frameworks proposed so far. Yet, the framework herein presented is just a simplification of a more ambitious one that is current subject of research by the authors and that, by including linguistical knowledge, is able to reduce the string error rate to some 5%, using task and context independent subword units.

The authors wish to thank Asunción Moreno for her valuable collaboration in this work.

6. REFERENCES

[1] A. Bonafonte, J.B. Mariño and A. Nogueiras, *SETHOS: The UPC Speech Understanding System*, Proc. ICSLP'96, pp. 2151-2154, 1996.

[2] W. Chou, B.-H. Juang and C.-H. Lee, *Segmental GPD Training of HMM based speech recognizer*, Proc. ICASSP'92, pp. 473-476, 1992.

[3] M. Kurimo, *Comparison Results for Segmental Training Algorithms for Mixture Density HMM's*, Proc. EUROSPEECH'97, pp. 87-90, 1997.

[4] C.-H. Lee, B.-H. Juang, W. Chou and J.J. Molina-Pérez, *A Study on Task-Independent Subword Selection and Modelling for Speech Recognition*, Proc. ICSLP'96, pp. 1820-1822, 1996.

[5] Y. Normadin, R. Lacouture, R. Cardin, *MMIE Training for Large Vocabulary Continuous Speech Recognition*, Proc. ICASSP'94, pp. 1367-1370, 1994.

[6] R. Schlüter, W. Macherey, S. Kanthak, H. Ney and L. Welling, *Comparison of Optimization Methods for Discriminative Training Criteria*, EUROSPEECH'97, pp. 15-18, 1997.