MINIMUM CONFUSIBILITY TRAINING OF CONTEXT DEPENDENT DEMIPHONES

*Albino Nogueiras-Rodríguez** *José B. Mariño*

Research Center TALP, Department of Signal Theory and Communications.
Universitat Politècnica de Catalunya. Barcelona, SPAIN.
{albino,canton}@gps.tsc.upc.es

10.21437/Eurospeech.1999-686

ABSTRACT

During the last years two different approaches have been widely used in order to improve the acoustic modeling in continuous speech recognition systems: discriminative training algorithms and context dependent subword units. However, while the use of each of these techniques leads to much better results than standard maximum likelihood trained phone models, their combination, i.e. discriminative training of context dependent units, has revealed to be a much more difficult task. In this paper we deal with minimum confusibility training of demiphones using TIMIT database. By applying this approach —recently introduced by the authors—, the string error rate in the recognition of TIDIGITS using demiphones is reduced some 24% with respect to maximum likelihood training. This improvement is added to the 8% reduction already provided by demiphones with respect to minimum confusibility trained phones.

1 INTRODUCTION

1.1 Context Dependent Subword Units for Continuous Speech Recognition: the Demiphone

Continuous speech recognition (**CSR**) is based on the assumption that speech can be seen as the concatenation of short meaningless sounds, called subword units. The main feature of CSR systems is that they are able to recognise no matter which task without the need of training task specific models. In order to do so, the set of subword units should accomplish three conditions:

1. They should provide a complete and unambiguous transcription of any possible oral message.
2. The different subword units should be distinguishable one each other by means of their acoustic properties.
3. The acoustic properties of each realisation of a given unit should not depend on its contexts.

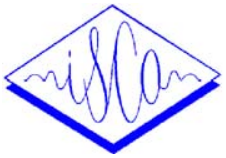
*This research was supported by the CICYT of the Spanish government under contracts TIC98-0423-C06-01 and TIC98-0685

The first two conditions relate to the ability of the system to undertake the recognition of a given task in terms of the corresponding subword units. The last is needed in order to ensure that a set of models trained using a general purpose database will be essentially identical to those which would be obtained if the database had been specific to the task.

The phone is the most evident choice as subword unit: it provides a complete and almost unambiguous transcription of any possible sentence, and different phonemes have acoustic properties that enable —to a certain point— the distinction between its realisations and those corresponding to any other one. Besides, their acoustic properties are relatively homogeneous: different realisations of a certain phone are more similar to the rest of realisations of the phone than to most of the realisations of other ones. Thus, with the only aid of a phone bigram, it is easy to achieve phone error rates below 40% —complete results in the recognition of TIMIT are provided later in this paper—, using maximum likelihood trained phone models.

Nevertheless, the context independence condition is not completely satisfied by the phone. This is so because the acoustic properties of a phone depend at a great deal on the sounds that circund it. This effect is called *coarticulation*, and may have different causes: physiological inerce and preparation in the articulation of sounds, finite length analysis windows, etc. Although the phone does not completely satisfy the context independence condition, it is usually assumed that the differences produced by the context will be of lower magnitude than those due to the kind of phone. If this is the case, we can train a model that somehow capture the properties of the whole of possible contexts of a certain phone by guaranteeing that enough samples of each of them appear in the training set. This leads to the so called context independent phones, which must be trained using phonetically balanced databases.

A different approach consists in explicitly modeling the context dependency. This means that a different unit is used for each phone depending on the context in which it is found. The main drawback is that the number of units grows dramatically as a longer context is considered for each phone. The demiphone is one of the possible solutions to this problem. It con-



6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

sists in considering that only the neighbouring phones are responsible of the effects of coarticulation and, furthermore, that this effects are mostly concentrated in the zone of the transitions between the current phone and its neighbours. Thus, the acoustic characteristics of the initial part of a phone are supposed to be much more affected by the previous phone than by the following one, and the opposite happens with the final part. This two assumptions enable us to undertake the problem of coarticulation dividing each phone in two sequential parts: an initial demiphone, which explicitly models the dependency on the previous phone; and a final demiphone, which does the same with the following phone. In different experiments carried out in Spanish [2, 3], the demiphone has revealed to be a good compromise between complexity and accuracy in the characterisation of coarticulation, clearly outperforming other typical solutions —as triphones— even in the case of the most sophisticated state-tying frameworks.

1.2 Discriminative Training for CSR: Minimum Confusibility Training over Short Chains of Subword Units

Discriminative training (**DT**) is one of the most appealing techniques in order to get high performance acoustic models. Instead of maximum likelihood, whose effectiveness relies on the correctness of the model election, i.e. in the fact that speech behave as a Markov chain in the case of hidden Markov models (**HMM**'s), DT directly aims at the minimisation of the number of errors committed in the recognition. This is usually accomplished by somehow incorporating the decision rule used in the recognition inside the training algorithm. Thus, a common feature to all the discriminative frameworks proposed so far is that they rely on the most probable confusions of the training utterances. The acoustic models are reestimate in such a way that the score of the correct sentence is increased and those of the incorrect hypotheses decreased.

The dependency on the decision rule used in the recognition is a major drawback of DT: if we apply a different rule —i.e. vocabulary, language model, etc.— in the training and recognition phases, we have no guarantee that reducing the number of errors in the training set will result beneficial in the recognition of the task. As a consequence, their application to subword based CSR has resulted much more cumbersome than to task dependent systems. It has not been until rather recently that DT frameworks aimed at minimising the error rate in the recognition of actual CSR tasks using task independent training databases have been proposed [1, 5, 4].

In the proposal of the authors, a novel loss function, the *confusibility*, is minimised using as training utterances short chains of subword units taken from a general purpose database. The confusibility is similar to the classification error but, instead of being a measure of the expected number of misrecognised utterances, it provides a measure of the total number of incorrect hypotheses that could be misrecognised if the grammar of the task allowed them [5]. On the other hand, the use of chains of a few subword units as training utterances represent a compromise between the ability to consider all possible errors and the possibility of training the models using a database of reasonable size.

The combination of the confusibility criterion with the use of short chains of subword units, enables us to consider the minimisation of the task dependent confusibility using task independent databases by incorporating in the reestimation formulae a measure of the relevance of each confusion between short chains of phones in the recognition of the actual task [4]. In a first approximation, the relevance of an error is given the value of the frequency of appearance of this error in the set of all the possibles error in the task. This frequency may be estimated with the aid of a bigram which, in the case the task is not known, is made equal to that of the language —English in our experiments—, leading to a task independent formulation. By applying consecutively both task independent and task adapted minimum confusibility training using TIMIT, the string error rate in the recognition of TIDIGITS strings was reduced in [4] from 7.5% to 5.1% —a 32% reduction—.

2 DISCRIMINATIVE TRAINING OF CONTEXT DEPENDENT SUBWORD UNITS FOR CSR

Given the good results provided by both context dependency modeling and DT, their combination, i.e. discriminative training of context dependent units would be a very interesting approach if the benefits of both techniques accumulated. Unfortunately, this kind of training is much more difficult to perform due to the following reasons:

1. The needs of training material grow dramatically when context dependent units are to be trained using DT algorithms.

2. Both techniques are somehow overlapping: discriminative training is expected to improve the modeling of the most confusable contexts, while context dependency improves the modeling of them all.

3. Context dependent units already represent a very low error rate situation, so further improvements are harder to get.

As a consequence, many times either it is impossible to improve with DT the maximum likelihood trained context dependent unit models, or the result achieved is worse than that obtained by applying DT to context independent ones.

2.1 Minimum Confusibility Training of Context Dependent Demiphones

Despite of the aforementioned difficulties, the application of DT to context dependent subword units should not differ much from its application to context independent phones. All that should be done is to provide enough training material, and expect that the situation given by the explicit context dependency does not represent a local minimum in the error rate such that DT is unable to improve it.

In our case, we applied minimum confusibility training to demiphones. The objective is the minimisation of the number of errors committed in the recognition of a given task —TIDIGITS strings— using acoustic models trained with a task independent database —TIMIT—.

Demiphones represent a specially well suited unit for the reestimation of its acoustic models using DT because, while providing a very good representation of coarticulation effects, the number of units needed to cover a significant portion of any lexicon is much lower for them than for triphones. In this way, demiphones provide results as good or better than triphones without the need of increasing excessively the number of units and complexity of the system [2]. Thus the difficulties of DT in the reestimation of context dependent unit models are of lower magnitude for demiphones than for triphones.

Although the higher immunity to lack of training material of demiphones in front of other alternatives, it was also notorious in the case of task adaptation. In this case only the segments in the training set for which all the transitions between units are present in the task participate in the training. In the case of segments of only five demiphones —some three phones—, less than the 1% of the available segments in TIMIT may occur in the strings of digits task. In order to get rid of this problem, a smoothed version of the bigram of the task was used. This smoothed bigram is constructed adding to the probability of each transition to and from a subword unit present in the task, a fraction of the probability that this transition is produced in the training set.

3 EXPERIMENTATION

In order to assess the usefulness of DT in the reestimation of demiphones we performed two different series of experiments: TIDIGITS strings recognition and speaker independent phone recognition. In both series we used the same set of 344 demiphones selected with a clustering algorithm. Semi-continuous HMM's of two states, where each of them must be visited, were trained, using TIMIT database, both for the context independent demiphones —which are equivalent to context independent phone models of four states—, and for the context dependent ones. The rest of details are identical to the experimentation framework presented in [4] except for the training set used. In the TIDIGITS strings recognition experiments we used the whole male part of TIMIT instead of just the training part in order to increase the training material available. In the case of the phone recognition experiments, the male test corpus was kept apart of the training in order to provide an estimation of the phone recognition performance in speaker independent conditions.

Name	Error	Sust	Inse	Dele	Goal	Corr
BasePhon	2.10	1.0	0.6	0.5	98.4	94.1
IndePhon	1.69	0.9	0.4	0.4	98.7	95.3
TaskPhon	1.76	0.9	0.4	0.5	98.6	95.1
BothPhon	1.48	0.7	0.4	0.4	98.9	95.8
BaseDmPh	1.27	0.4	0.4	0.4	99.1	96.3
IndeDmPh	1.10	0.4	0.4	0.3	99.2	96.9
TaskDmPh	0.96	0.3	0.3	0.3	99.4	97.2
BothDmPh	1.05	0.4	0.4	0.3	99.3	97.0

Table 1: TIDIGITS strings recognition results using several DT frameworks and either context independent phones or demiphones trained with TIMIT database.

3.1 TIDIGITS Strings Recognition using TIMIT Trained Subword Units

In the case of TIDIGITS strings recognition, eight different training frameworks were compared:

- `BasePhon` Maximum likelihood trained context independent phone models.
- `IndePhon` Task independent —language adapted— minimum confusibility phones.
- `TaskPhon` Task adapted minimum confusibility phones.
- `BothPhon` Consecutively task independent and task adapted minimum confusibility phones.
- `BaseDmPh` Maximum likelihood trained demiphone models.
- `IndeDmPh` Task independent —language adapted— minimum confusibility demiphones.
- `TaskDmPh` Task adapted minimum confusibility demiphones.
- `BothPhon` Task independent and task adapted minimum confusibility demiphones.

Complete results on this experiment are shown on Table 1, where:

Framework	Error	Sust	Inse	Dele	Goal
BasePhon	39.0	23.8	7.9	8.1	68.1
IndePhon	34.3	21.0	6.6	6.8	72.2
BaseDmPh	33.2	20.5	6.3	6.5	73.0
IndeDmPh	30.6	19.1	5.6	5.8	75.1

Table 2: Speaker independent TIMIT phone recognition results using several DT frameworks and either context independent phones or demiphones.

Error	Digit error rate, equal to the sum of substitutions, insertions and deletions.
Sust	Digit substitution rate.
Inse	Digit insertion rate.
Dele	Digit deletion rate.
Goal	Percentage of correctly recognised digits.
Corr	Percentage of correctly recognised strings.

The first remarkable thing about these results is the benefit shown by the use of demiphones. Thus, even the best result with DT trained context independent phone models is worse than that of ML trained demiphones. It is also remarkable that DT is still able to improve the performance of demiphones models yielding a notorious 24% of reduction in the string error rate. Finally, it is also remarkable the different behaviour of phones and demiphones with respect to DT: while both DT techniques, task independent and adapted, show cumulative benefits for phone models, this is not the case for demiphone models, leading all three frameworks to a similar result.

3.2 Speaker Independent Phone Recognition

In the case of phone recognition, a bigram of the acoustic unit was used. This means that the language models used for phones and demiphones were not the same —a bigram of demiphones is equivalent to a trigram of phones—. Only four of the above mentioned frameworks are tried: BasePhon, IndePhon, BaseDmPh and IndeDmPh, with the same meanings as there. Table 2 shows the complete recognition results. Once again ML trained demiphones outperform both ML and DT trained phones, but the more detailed language model employed in the case of demiphones is surely benefiting them. It is also remarkable the notorious improvement yielded by DT both for phones and for demiphones.

4 CONCLUSSIONS

This paper shows that the minimum confusibility criterion can successfully improve the recognition performance of context dependent units using task independent training databases. Applied to the reestimation of demiphone models using TIMIT, the method yields a global 50% reduction in TIDIGITS string error rate with respect to standard maximum likelihood context independent phones.

References

[1] C.-H. Lee, B.-H. Juang, W. Chou, and J.J. Molina. A study on task-independent subword selection and modeling for speech recognition. In *Proc. of ICLSP'96*, pages 1820–1823, 1996.

[2] J.B. Mariño, A. Nogueiras, and A. Bonafonte. The demiphone: an efficient subword unit for continuous speech recognition. In *Proc. of EUROSPEECH'97*, pages 1215—1218, 1997.

[3] J.B. Mariño, P. Pachès-Leal, and A. Nogueiras. The demiphone versus the triphone in a decision-tree state-tying framework. In *Proc. of ICSLP'98*, 1998.

[4] A. Nogueiras and J.B. Mariño. Task adaptation of sub-lexical unit models using the minimum confusibility criterion on task independent databases. In *Proc. of ICSLP'98*, September 1998.

[5] A. Nogueiras and J.B. Mariño. Task independent minimum cofusibility training for continuous speech recognition. In *Proc. of ICASSP'98*, pages 477–480, May 1998.