TASK ADAPTATION OF SUB-LEXICAL UNIT MODELS USING THE MINIMUM CONFUSIBILITY CRITERION ON TASK INDEPENDENT DATABASES

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

Universitat Politècnica de Catalunya. Barcelona, SPAIN.

{albino,canton}@gps.tsc.upc.es

ABSTRACT

Discriminative training is a powerful tool in acoustic modeling for automatic speech recognition. Its strength is based on the direct minimisation of the number of errors committed by the system at recognition time. This is usually accomplished by defining an auxiliary function that characterises the behaviour of the system, and adjusting the parameters of the system in a way that this function is minimised. The main drawback of this approach is that a task specific training database is needed. In this paper an alternative procedure is proposed: task adaptation using task independent databases. It consists in the combination of acoustic information—estimated using a general purpose training database—, and linguistic information—taken from the definition of the task—. In the experiments carried out, this technique has led to great improvement in the recognition of two different tasks: clean speech digit strings in English and dates in Spanish over the telephone wire.

1 INTRODUCTION

One of the most appealing features of sub-word based continuous speech recognition (**CSR**) is that the acoustic models can be trained using a general purpose speech database, instead of requiring a task specific one. The underlying assumption is that any word may be seen as the concatenation of small sub-lexical units independent of the context they are in, either because this context is explicitly modelled—when context dependent units are used—or because its effects are discarded—using context independent ones—. In any case, and because of this independence, any speech database where the different sub-lexical units are correctly represented suffices to train the acoustic models. Recognition of any task only requires then the knowledge of how the sentences allowed by the task are represented as a function of these units.

Discriminative training (**DT**) of sub-lexical unit models for CSR is a major challenge in acoustic modelling. This is so because most DT implementations proposed so far rely on the optimisation of the performance of the system at recognition time. While this optimisation is quite straightforward for task specific systems—those where the acoustic models are trained using a task specific speech database—, it is rather more cumbersome when the task is not known at training time or only a general purpose database is available. In these last cases, a sub-lexical units based CSR framework must be used. The main problem in applying DT to this kind of system is that its behaviour at recognition time cannot be accurately represented. One way to overcome this drawback consists in increasing the discrimination between the different sub-lexical units in a grammar free environment, and expect that this increased discriminative strength results in an improvement of the performance in no matter which task. Several such frameworks—which are referred to as task independent DT—have been proposed recently [2, 4], leading to notorious improvements in the performance of CSR systems. Nevertheless, this approach does not take any profit of the task definition when it is known but a task specific database is not available. In this paper we consider the utility of embedding the knowledge about the task characteristics in our previously proposed task independent DT framework [4]—minimum confusibility training (**MCT**)—in order to focus it to the minimisation of the errors that may actually be involved when a task is to be recognised.

2 TASK ADAPTATION USING MCT ON SHORT CHAINS OF SUB-LEXICAL UNITS

Grammar free sub-lexical units recognition and task specific recognition represent rather different situations. While in the first acoustic modelling is determinant, in task specific recognition, the grammar may prevail in many of the cases. For instance, using maximum likelihood (**ML**) trained HMM's, the phone error rate recognising TIDIGITS is some 40%. If the grammar of the task—strings of digits—is used, this rate falls below 2% with the same acoustic models. This drastic reduction in the phone error rate represents that most of the phone units will be correctly recognised or forced to be so by the grammar of the task. In this situation, it will usually be likely to find chains of one or more units correctly recognised. These correctly recognised segments enable us to consider each lexical error as the concatenation of correctly and incorrectly recognised segments. Moreover, if a segment of two or more units is correctly recognised, we can expect that not only the recognised transcription will be correct, but also the points where the transition between units will occur. For instance, let's consider the confusion between digits 'five' and 'nine'. The transcription in acoustic units of both words is `[f ay v]` and `[n ay n]`, respectively. The presence of a common central unit has an isolating effect between the initial and final segments. We can consider that this

[*] This paper was supported by the CICYT, Spanish government, under contract TIC95–0884–C04–02, and the CIRIT, Generalitat de Catalunya, through the CREL.

lexical error requires, to be committed, that at least one of two segmental errors is possible: either [f ay] for [n ay], or [ay n] for [ay v]. Moreover, we can expect that the possibility of confusing [f ay] with [n ay] will not depend on the possibility that [ay n] is confused with [ay v]. This consideration led the authors to the proposal of a segmental DT approach where short chains of sub-lexical units are taken as training material [4].

Segmental DT using short chains of sub-word units represents a tradeoff between training availability and CSR errors representation. Each utterance in the training database is divided into segments of a few units. Each of these segments is then used as if it were an utterance completely independent of the rest. An N-Best search is performed on each segment requiring that the initial and final units of the segment are correctly recognised. By fixing the recognition of the extreme units we ensure that the method is little influenced by the segmentation in units of each sentence, so a maximum likelihood forced recognition is enough to determine this alignment. There is a compromise in the length of the segments between the availability of training material, inversely proportional to their length, and their ability to characterise actual CSR errors, which grows with it. In our experiments, a value of five phones was used.

Besides, and considering the same example as above, it must be remarked that it is not necessary that both errors are possible. If just one of the segments would be misrecognised by itself, the commission or not of the lexical error will depend on the ability of the system to reject the wrong hypothesis for the other segment. Moreover, only the protection against the commission of these two sub-lexical errors intervenes in the possibility that 'five' and 'nine' are confused. It does not matter much if there are other sub-lexical errors more prone to be committed. For instance, [m ay] is much more prone to be confused with [n ay] than [f ay], but it does not participate in the confusion between 'five' and 'nine' or any other digit. In this situation we found that minimum classification error or maximum mutual information training was not suited for CSR because they focus mainly on the reduction of the possibility of committing the most likely confusion for each utterance. We proposed instead the use of a minimum confusibility training (**MCT**) approach [4].

We define the *confusibility* of a system as the expected number of errors that it is possible to commit in the recognition of a task. This means that each utterance contributes to the global confusibility of the system as many times as the number of different sentences in the grammar of the task whose likelihood is superior than that of the uttered one. Our purpose is to formulate a task dependent confusibility measure of the system using a general purpose phonetically balanced training database and the knowledge of the grammar of the task. The main idea is to learn the acoustic characteristics from the training database and, taking into account which are the errors allowed by the task grammar, adapt the parameters of the system in such a way that this measure is reduced. In order to do so, we first consider the case in which the training database has enough utterances of each possible sentence of all possible tasks in a given language. As all tasks are completely represented, it is possible to consider only those utterances belonging to the task being optimised and then performing task dependent MCT on them. As the frequency of appearance of the sentences of the task may be different in the training database and the task, the formulation of the confusibility must take this fact into account.

2.1 Task Dependent Confusibility Formulation

Let's be Λ the parameters of the system being optimised; $\mathbf{W}$, the grammar of the task; and $x_n \in X$, the nth utterance in the training database. As the lexical contents of the utterance must be known in advance to performing DT, we shall refer to x_n^i denoting a utterance corresponding to word w_i in the vocabulary of $\mathbf{W}$[1]. Being $g_j(x_n^i) = \log \Pr(x_n^i, \Theta^j/\lambda_j)$ the log-likelihood of utterance n along its best path, Θ^j, through the acoustic model of word j, λ_j, we define the possibility of committing error $[\, w_i \rightarrow w_{j \neq i} \,]$ for x_n^i as:

$$\mathcal{P}_{ij}^{x_n^i}(\Lambda) = \begin{cases} 1 & g_j(x_n^i) \geq g_i(x_n^i) \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

$$\approx \frac{1 + \tanh(\frac{g_j(x_n^i) - g_i(x_n^i)}{G_0})}{2} \tag{2}$$

Where Equation 2 is used instead of Equation 1 in order to guarantee continuity—necessary to allow a gradient-descent optimisation procedure—. The shape of this function is essentially the same as the sigmoid used in [2], but leading to a more compact notation, and it is only controlled by G_0. If this value is close to zero, Equation 2 will behave as a hard limiter. If a high value is taken, then the function will almost be linear. In our implementation, we use the standard deviation of $g_j - g_i$ for the two first hypotheses of each utterance in the whole population. In this way, we ensure that Equation 2 will be nearly linear for the errors with likelihood closest to that of the correct transcription—those which are the most easily removed by the DT algorithm—, while limiting the effects of the extreme ones.

We define the *utterance confusibility* of x_n^i as the number of erroneous hypotheses allowed by the grammar of the task with higher likelihood than the correct one,

$$\mathcal{UC}^{x_n^i}(\Lambda, \mathbf{W}) = \sum_{j \neq i} \mathcal{P}_{ij}^{x_n^i}(\Lambda) \cdot 1([\, w_i \rightarrow w_j \,] \in \mathbf{W}) \tag{3}$$

Where the $1([\, w_i \rightarrow w_j \,] \in \mathbf{W})$ function returns one if the grammar of the task allows the confusion between w_i and w_j, and zero otherwise. Similarly, the *class confusibility* is defined as the expected mean number of possible errors for the utterances belonging to a given word. So, being N_i the number of times word w_i appears in the training database:

$$\mathcal{CC}_i(\Lambda, \mathbf{W}) = \frac{1}{N_i} \sum_{x_n^i} \mathcal{UC}^{x_n^i}(\Lambda, \mathbf{W}) \tag{4}$$

[1] Although the following development refers to words in a vocabulary, it is simply a matter of convention. All results may be applied equally to any kind of task substituting words for sentences and vocabulary for grammar.

Finally, we shall refer to *global confusibility* or, simply, *confusibility*, as the expected number of errors that the system may commit:

$$\mathcal{GC}(\Lambda, \mathbf{W}) = \sum_i f_W(w_i)\mathcal{CC}_i(\Lambda, \mathbf{W}) \tag{5}$$

Where $f_W(w_i)$ is the frequency of appearance of word w_i in the task. Summing up Equations 1–5 we arrive to:

$$\begin{aligned}\mathcal{GC}(\Lambda, \mathbf{W}) = \sum_i f_W(w_i)\cdot \\ \frac{1}{N_i}\sum_{x_n^i}\sum_{j\neq i} \mathcal{P}_{ij}^{x_n^i}(\Lambda)\cdot 1([\,w_i \rightarrow w_j\,] \in \mathbf{W})\end{aligned} \tag{6}$$

Notice that this last expression mixes three different kinds of information: acoustic information is used in the term $\mathcal{P}_{ij}^{x_n}(\Lambda)$, and only there; N_i is the number of times each word in the task appears in the training database; finally, $f_W(w_i)$ and the explicit condition $1([\,w_i \rightarrow w_j\,] \in \mathbf{W})$, refer to the vocabulary structure of the task. Re-arranging the order of the summations in Equation 5 we arrive to

$$\mathcal{GC}(\Lambda, \mathbf{W}) = \sum_{x_n^i}\sum_i\sum_{j\neq i} \frac{\rho(w_i, w_j, \mathbf{W})}{N_i}\mathcal{P}_{ij}^{x_n^i}(\Lambda) \tag{7}$$

Where $\rho(w_i, w_j, \mathbf{W}) = f_W(w_i)1([\,w_i \rightarrow w_j\,] \in \mathbf{W})$ is a measure of the expected number of errors allowed by the grammar if one w_i is confused with w_j. We shall refer to to this measure as the *relevance* in $\mathbf{W}$ of error $[\,w_i \rightarrow w_j\,]$.

2.2 Task Adaptation Formulation Using Short Segments of Phones

Equation 7 already leads to a formulation of the task dependent confusibility using a task independent training database. Nevertheless the nature of this database—it should be almost infinite to hold all possible tasks—makes its use un-affordable. Instead of that, we can expect that the possibility that a lexical error occurs will be a function of the possibility that each of the segments intervening in it is confused. In the above example, this means that the possibility of confusing `[f ay v]` with `[n ay n]` will depend on the possibility of confusing `[f ay]` with `[n ay]` and `[ay n]` with `[ay v]`. In general, we will have an application between the $M \times M$ space of all the possible errors between the M different sub-lexical segments in a language, E_S, into the $N \times N$ space of all possible errors between the N different words in the task, E_W. The shape of this function will be, in general, intricate, but there are two specially interesting situations where this function has a straight formulation: if the lexical error only requires one segment to be misrecognised, the possibility of committing the lexical error will be exactly the possibility of confusing this segment; besides, and due to the almost linear behaviour of Equation 2 for arguments close to zero, if the difference between the likelihood of the correct and incorrect transcriptions for all the segments in which a given lexical error is divided is close to zero, then the possibility of committing the lexical error will approximately be the sum of the segmental possibilities. Thus, being T the number of segments needed to commit error $[\,w_i \rightarrow w_j\,]$, $y_{n,t}^k$ each of the segments in which x_n^i is divided and s_k the acoustic contents of each segment, we can express the possibility of committing error $[\,w_i \rightarrow w_j\,]$ as:

$$\mathcal{P}_{ij}^{x_n^i}(\Lambda) \approx \sum_t^T \mathcal{P}_{s_k s_j}^{y_{n,t}^k}(\Lambda) \tag{8}$$

This expression is equivalent to considering that each lexical error contributes to the global confusibility as many times as the number of segments that need to be confused to commit it. In this way, the minimisation of the task dependent confusibility can be undertaken in terms of the segments—what we actually have available in the training database—instead of the words in the task.

$$\begin{aligned}\mathcal{GC}(\Lambda, \mathbf{W}) &\approx \sum_k\sum_{y_n^k}\sum_{j\neq k} \frac{\rho(s_k, s_j, \mathbf{W})}{N_{s_k}}\mathcal{P}_{s_k s_j}^{y_n^k}(\Lambda) \\ \rho(s_k, s_j, \mathbf{W}) &\approx f_{E_W}([\,s_k \rightarrow s_j\,])\end{aligned} \tag{9}$$

Where s_k are all the different segments found in the training database, N_{s_k} the number of times these segments are present, and the relevance given to each segmental error $[\,s_k \rightarrow s_j\,]$ is the frequency of appearance of this segmental error in the $N \times N$ space E_W of all possible errors in the task.

2.3 Simplified Calculation of the Relevance

One useful simplification of Equation 9 is considering that the uttered string and the recognised one are independent. This means that any segment of speech may be confused with all the possible segments allowed by the task. In this situation, the frequency of a given segmental error will be the product of the frequency of the uttered segment by the frequency of the recognised one. These frequencies may be estimated in a variety of ways, being the simplest the use of a stochastic language model, such as N-gram's. In our case, we use the bigram.

Even in the case that the task grammar is not known, this approximation provides a reasonable formulation for task independent MCT. It consists in substituting the language model of the task in the calculus of the relevance with a language model of the language (English, Spanish, etc.). This last language model can be seen as the model that would have a task involving all possible tasks in the language and it seems a natural choice for automatic dictation systems.

3 EXPERIMENTATION

In order to assess the effectiveness of task adaptation using MCT on short phonetic segments, we have applied several DT strategies based on MCT on two different CSR tasks: English digit strings recognition and Spanish dates recognition. In both cases, a maximum likelihood framework using context independent phone models is taken as baseline for the DT procedures. We optimise the confusibility of this framework using segments of five phones where the extreme ones are required to be correctly recognised. An N-Best search is performed on each segment of each utterance in the training set to find the twelve maximum likelihood hypotheses which are then used to compute the gradient of the confusibility. Finally, and using the algorithm depicted in [5], a gradient descent search is performed in order to minimise the confusibility of the system. Three different kinds of informations are trained: the transition and emission probabilities of the HMM's, and the weight each kind of parameter is given at each state in a similar way to that proposed using ML in [1]. Four different discriminative training strategies were tried and compared with the ML baseline (**base**):

inde Independent MCT: no knowledge from either the language or the task is used.

lang Language dependent MCT: a bigram of phones in the language is estimated from the training set and used to estimate the relevance of each error.

task Task dependent MCT: a bigram of phones is inferred from the definition of the task and used in the estimation of the relevance of the errors.

mix3 The same framework as **task** but using the result from **lang** as starting point.

3.1 English Digit Strings Recognition

The English digit strings recognition is mainly the same experiment framework as that used in [4]: phone 4 states HMM's are trained using the male part of TIMIT and employed to recognise the unknown length digit strings of the male test corpus of TIDIGITS. Table 3.1 shows the results achieved at recognition, where **err** stands for the digit error rate—the sum of insertions (**inse**), deletions (**dele**) and substitutions (**subs**) of digits—, **goal** for the percentage of correctly recognised digits, and **corr** for the percentage of strings correctly recognised.

name	err	inse	dele	subs	goal	corr
base	2.7	0.8	0.8	1.1	98.1	92.5
inde	2.5	0.5	0.5	1.5	98.0	93.0
lang	2.2	0.6	0.5	1.1	98.4	94.0
task	2.3	0.7	0.6	1.0	98.3	93.6
mix3	1.8	0.5	0.4	0.9	98.7	94.9

Table 1: English digit strings recognition results.

The first remarkable thing about the results is that all four DT frameworks improve the baseline one. The best result is obtained when task adaptation is applied to the best task independent framework (**mix3**). This was the expected result and confirms that MCT based task adaptation is a valuable tool in CSR. Language adaptation has also proved to perform better than independent MCT. Although it would be expected that task adaptation also performed better than language adaptation, the result is just the inverse. We believe that this behaviour is due to the fact that task adaptation wastes most of the training material. It should be noted that any phone not appearing in any of the digits suffices to void the relevance of the whole segment. For instance, only some 5% of TIMIT segments is profited in MCT—and the available material in TIMIT is not too much to waste it!—. In the experiments we have carried out the language models used are smoothed, nevertheless the profited material still falls below 25%.

3.2 Spanish Dates Recognition

We have also applied three of the discriminative methods **lang**, **task** and **mix3** to a rather different task: recognition of Spanish spoken dates recorded through the telephone fixed network. The speech database used was the Spanish SpeechDat corpus [3]. 680 Speakers—both male and female—were extracted from four dialectal zones of Spain. A rather simplified grammar of the dates was used. The transitions between words were also adjusted in the **base** experiment, but this same value was used for the other three ones[2]. Function words are not taken into account in the recognition results, which are shown on Table 2.

name	err	inse	dele	subs	goal	corr
base	24.0	4.4	4.8	14.8	80.0	33.7
lang	21.2	2.1	5.2	13.8	80.9	38.7
task	22.0	2.1	5.4	14.4	80.2	36.7
mix3	19.6	1.9	4.9	12.8	82.3	43.7

Table 2: Spanish dates recognition results.

ACKNOWLEDGMENT

The authors wish to thank Dr. Asunción Moreno Bilbao and Mr. Marcial Chinchilla for their valuable suggestions and for their contribution in the SpeechDat experiments.

REFERENCES

[1] J. Hernando. Maximum likelihood weighting of dynamic speech features for cdhmm speech recognition. In *Proc. of ICASSP'97*, pages 1267–1270, 1997.

[2] 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.

[3] A. Moreno and R. Winsky. Spanish fixed network speech corpus. SpeechDat Project LRE-63314, 1995.

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

[5] A. Nogueiras, J.B. Mariño, and E. Monte. An adaptative gradient search based algorithm for discriminative training of hmm's. In *Proc. of ICLSP'98*, 1998.

[2] At the time of writing this paper we could not have complete results with the grammar weights adjusted for the discriminative frameworks. Note that this situation favours the recognition results of the **base** experiment in front of them.