ITERATIVE OPTIMIZATION OF SUB-WORD TEMPLATES FOR SPEECH RECOGNITION

Carlo SCAGLIONA(*) , Cesare VICENZI(*)
Angelo CARBOSINO(*) & Donatella SCIARRA(*)

(*)Central Research Service
Eletronica San Giorgio, ELSAO S.p.A.
Via G. Puccini, 2 - 16154 GENOA - ITALY

ABSTRACT

An iterative optimization procedure to derive reliable sub-word templates from training speech material is described. At each step sub-word occurrences are located and new templates computed in such a way that the overall distance of speech material from the language model approaches an absolute minimum. Different options are discussed and experimental results presented.

INTRODUCTION

In recognizing speech by the use of sub-word units, an automatic derivation of a reliable and representative set of templates is perhaps the most critical point. The key steps in this process are the identification of occurrences of various sub-word units in a set of training utterances and the subsequent derivation of representative templates. The location of sub-word acoustic events in fluent speech is not an easy task and it is particularly troublesome for diphones that follow our definition: a collection of short transitional sounds, only few frames long, and stationary sounds represented by single-frame templates. Earlier training methods for diphones-based recognizers used hand-derived templates from a primary "bootstrapping" speaker to locate events by means of "forced recognition", i.e. a recognition step in which only valid interpretations are allowed. \(^{(2)}\) The working hypothesis here is that the bootstrapping speaker's templates are good enough for locating events, even if they would give poor results during recognition. However, the assumption proved to be sufficiently correct when small vocabularies were dealt with \((2,3,5)\) and also for some speakers when the whole diphone inventory for the Italian language was needed to recognize large vocabularies of isolated words \((4)\). For some other speakers, however, automatic diphone bootstrapping resulted in low recognition rates, basically due to the assumption that diphone occurrences were known when they appeared clearly as outliers and there were enough correctly located occurrences of a diphone.

The problem we faced was that of finding a way of eliminating segmentation errors and of deriving templates that were truly representative of the training speech material. On the other side, we knew that in the field of recognition by Hidden Markov Models, a powerful method exists to learn model probabilities: the Forward-Backward algorithm \((6)\). This is a recursive algorithm that adjusts model probabilities at each step, in such a way as to increase the probability of the sequence of training observations given the model \((7)\). It is experimentally shown that the F-B algorithm computes good models and produces a good segmentation of training speech material, even when applied to language models based on sub-word units, without any manual labeling of the speech material and starting from very poor initial estimates \((8)\).

While the physical phenomenon remains the same whether we model it by statistical or prototypical methods, we focused on the implementation of a template based training method with much the same properties of the above mentioned F-B algorithms. The present paper describes this method and discusses the different aspects and possible options, both at the model and experimental points of view. Experimental results from three speakers, where all the diphones of the Italian language had to be learnt to recognize a 10,000 word vocabulary, are presented and discussed.

THE ITERATIVE TRAINING PROCEDURE

Similarly to the F-B algorithm for training HMM's, the problem of our training procedure is: find the template set that optimally represents the training speech material by minimizing its overall spectral distance. Towards this goal we proceed iteratively, where at each step templates computed at the previous step are used to locate further sub-word occurrences in the training material; new templates are then computed so as to reduce the average spectral distance of each diphone from the relative occurrences. In this way, the overall distance gets closer to the absolute minimum at each step, even if we cannot prove that the absolute minimum is reached, since the form of the distance functional is not known.

The Single Template Case

To explain the algorithm more easily, let's assume that only one template for each basic unit is looked for. Let \(U(i,j)\) be the units (diphones in our case) and \(T(k)={t(j,k)}\) the relevant template at the \(k\)-th iteration. Let \(t(j,0)\) be an initial guess of the templates, derived in some way from some speaker. At \(k\)-th iteration, an optimal alignment of the speech training material with the language model is found by means of a Dynamic Programming algorithm that minimizes \(D(j,k)\) which is the sum of spectral distances between each input frame and the language model \(D(k)\). Owing to the certain number \(i\) of input frames is associated to a given spectral state of a given basic unit, as depicted in Fig. 1. Here we have represented a unit made of 5 spectral states (such as a transitional diphone \((4)\)) that are traversed without any possibility of skip and duplication, and a unit represented by a single spectral state that can be repeated an arbitrary number of times, within fixed limits (such as a stationary diphone \((4)\)). If we indicate by \(s(i,j)\) the \(i\)-th spectral state of unit \(U(i,j)\), we have that the \(i\)-th number \(n(i,j,k)\) of input frames are associated to it by the optimal alignment algorithm. Let \(d(i,j,k)\) be the average distance at \(k\)-th iteration of those \(n(i,j,k)\) frames from the speech template \(f(i,j,k)\) associated to \(s(i,j)\). Let's now compute the new templates \(f(i,j,k+1)\) so as to reduce, and possibly minimize, their average distances from the \(n(i,j,k)\) relevant frames associated to \(s(i,j)\). If the parameter space is a metric space and the spectral distance used is the squared euclidean distance, the average distance of each input frame associated input frame is lowered, the overall distance is also lowered to a value \(D'(k)\leq D(k)\).
At next optimal alignment step with the new spectral templates, the overall distance will be $D(k+1) < D(k)$. In fact, either the new optimal alignment is the same as that at previous step (and in this case the training procedure stops) or it is different, which implies $D(k+1) < D(k)$, by the very definition of optimal alignment algorithm. At each iteration, the overall distance gets thus closer to the absolute minimum, even if it cannot be excluded that the procedure may stop at a relative minimum.

At next optimal alignment step

$D(k+1) < D(k)$

algorithm.

gets thus

usually each template is a simple sequence of parameter vectors; now the entire "multiple" template is a single sequence of multiple parameter vectors, as depicted in Fig. 3. This is somewhat like considering each spectral state independently; the complete model actually closely resembles that of Partitioned Gaussian Autoregressive Hidden Markov Models (F13). This new definition clearly requires a modification in the spectral distance computation between an input frame and a spectral state of a unit; now the minimum distance of that frame from the different spectral templates representing the spectral state has to be selected.

Fig. 2 - Spectral templates at iteration $k$, indicated with $\circ$, are moved to positions indicated with $\triangle$ to reduce average distance from associated frames.

The Multiple Template Case

The above described procedure remains essentially the same when each spectral state $s(i,j)$ of basic unit $u(j)$ is represented by $N$ parameter vectors, i.e. $N$ spectral templates $f(i,j,n,k)$. Again, after $k$-th optimal alignment step, we have $n(i,j,k)$ input frames associated to spectral state $s(i,j)$; their average spectral distance from the relevant spectral templates is $d(i,j,k)$. Clearly now, the average distance from a set of templates is intended in the same sense as the average distortion in a vector quantizer (F11); for each input frame we consider the distance from the closest spectral template. Fig. 2 shows an example in a 2-dimensional parameter space, with $M=3$ spectral templates. The spectral templates used for $k$-th optimal alignment are modified so that the average distance $d(i,j,k)$ of the $n(i,j,k)$ input frames is reduced to a lower value $d'(i,j,k)$. The same considerations done before hold; in particular, at $(k+1)$-th optimal alignment step, the overall distance will be $D(k+1) < D(k) < D(k)$ (with $\neq$ sign holding when we have identical optimal alignments for two consecutive iterations, and no template is modified).

Please note that the definition of multiple templates for a transitional unit is different from what we are used to. Usually each template is a simple sequence of parameter vectors; now the entire "multiple" template is a single sequence of multiple parameter vectors, as depicted in Fig. 3. This is somewhat like considering each spectral state independently; the complete model actually closely resembles that of Partitioned Gaussian

Autoregressive Hidden Markov Models (F13). This new definition clearly requires a modification in the spectral distance computation between an input frame and a spectral state of a unit; now the minimum distance of that frame from the different spectral templates representing the spectral state has to be selected.

EXPERIMENTAL EVALUATION

Iterative Procedure Evaluation

The Iterative Training Procedure tests were performed on training sets collected from three male speakers, named OR, AC, and PF in the following. The speech material (training set) from each of them consisted in the utterance of 36 medium/long duration sentences (4 to 4.5 minutes of speech per speaker). Such sentences were extracted from a newspaper weather report and then adapted in order to contain at least a complete times all the diphones for the Italian language. The diphone inventory basically consists of 237 units, 23 of which are "stationary" ones; taking into account also the context of such stationary diphones, the number of units raises to 267. Each spectral state of diphones is represented by one or more parameter vectors. The parameters are 16 lifted Mel-Based Cepstra plus an overall energy value (in dB) for that frame, computed every 12.8 ms on a 25.6 ms Hamming window of speech. The training sentences were collected using a close-talking microphone in very noisy conditions (80 dB as an average) showing strong and frequent peaks up to 90 dB.

The Iterative Training Procedure can be summarized as follows:

1. an initial template set from another speaker (bootstrapping speaker) is selected;
2. a "forced recognition" algorithm for connected speech is applied to the training set of the new speaker (training speaker); this algorithm uses an HSA description of the expected sentences allowing alternate pronunciations; matching of the training set to the current template set is described by local and global distance scores;
3. a convergence criterion is used to verify if the iterative process has to be stopped (unless we are in the first iteration step); such criterion actually checks how much
percentual decrease in the global distance score has been achieved during current iteration step; if such value goes under a predefined fixed threshold, the procedure can be stopped and the current template set is used as the final one.

(iv) using the new diphone occurrences located in (ii), a "k-means" clustering algorithm is applied to all the 267 units to be trained; for each unit, and for each frame of the unit, we obtain a predefined (maximum) number of spectral templates; a different template count for stationary and transitional diphones is available;

(v) a new temporary template set is built from templates found in (iv);

(vi) procedure goes back to (ii).

By means of the above method, four types of template sets were obtained from each speaker:

a) a "T11" template set with a single template both for stationary and transitional diphones;
b) a "T15" template set with 5 templates for stationary diphones and 1 for transitional ones;
c) a "T25" template set with 5 templates for stationary diphones and 2 for transitional ones;
d) a "REPT15" set coming out from the first step of T15 iterative procedure.

The REPT15 template set is used as a reference to evaluate, by means of recognition tests, the effectiveness of the iterative technique.

The training procedure was first applied on speaker GR using for bootstrapping a previously available template set of a fourth male speaker CV; this set was derived using another non iterative training procedure described in [2,4]. Template sets for AC were bootstrapped using the T15 set of speaker GR; sets for PP had the T15 set of speaker AC as initial templates.

Table I shows the convergence results of the iterative method for the different types of template sets required. The numbers reported refer to a so called "global distortion" score corresponding to the overall distance value obtained after forced recognition of the training set. Global distortion is normalized with respect to overall distance score obtained in the first step; it is, in fact, an almost "good" one; therefore, the refinement of the locations of diphone occurrences in the training set.

As an example of the last mentioned effect, in Fig. 4 it is shown the segmentation history of a small portion of speech during the iterative procedure.

In the upper part of the figure the energy envelope is shown together with the expected hand-made segmentation in terms of diphones; the symbols "/", "-" and "'" refer to voice-bar, silence and "shwa" diphones respectively. In the lower part, for each step of the iteration, solid lines indicate the boundaries of the occurrences; the dashed lines simply show the "location moving" effect from one step to another.

Besides the fact that after step 4 the segmentation almost agrees with the expected one, the iterative method shows its effectiveness in the recovery of misplaced initial occurrences: after the first step, diphones in the central portion of the figure are badly detected, both in terms of location and duration. For instance, diphone /da/ is located over the sequence "da ga" owing to the combined effect of its poor similarity and misplacement of preceding...
diphone "o". This also causes the next diphones "t" and "s" to be pushed to wrong speech portions as shown. In the next step, all the misplaced diphones move to an almost correct location, still providing also a more accurate determination of the adjacent sounds. Also, note in step 2 the temporary detection of a spurious silence segment immediately removed in the next step.

Recognition Tests Results

Some recognition experiments have been performed in order to verify the goodness of the described approach. The template sets collected in the way described in the previous chapter have been tested with our Large Vocabulary Isolated Words Recognizer, whose description can be found in §3. The main features of such recognition system are the "diphone spotting" approach and the diphone-based word template description. The vocabulary submitted to the recognizer was a 10,000 words one; words were taken from the street directory of the city of Rome. Of course, given that kind of vocabulary, no contextual information could be used to enhance the recognition performances. Each one of the three speakers pronounced once 1,000 (or more) spoken units under the same noise conditions previously mentioned; such names were selected one every 10 names in the complete directory, and were thus the same ones for each speaker. These test sets were then matched against the T11, T15, T25 and REFT15 template sets.

Table II summarizes the Word Recognition Rates (rank 1) obtained; the symbols in square brackets indicate the bootstrapping template case used.

<table>
<thead>
<tr>
<th>Speaker Set</th>
<th>Word Recognition Rate (1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GR (CV)</td>
<td>73.4  74.5  74.4  72.0</td>
</tr>
<tr>
<td>AC (GR153)</td>
<td>77.2  84.4  83.5  81.4</td>
</tr>
<tr>
<td>PP (AC151)</td>
<td>89.7  89.0  89.4  89.2</td>
</tr>
</tbody>
</table>

**TABLE II** - Recognition results for the 10,000 words vocabulary.

The analysis of such results reveals quite clearly the convenience for a multiple template description of the spectral states (T25 and T25 have always better results than T11). The use of 2 templates on transitional diphones does not show, in these experiments, significant differences with respect to single template case; this is in our opinion due to the fact that the training material contained a too low number of occurrences for many of the transitional diphones. In this situation, averaging of the occurrences to a single template works as or even better than the two templates case, as overtraining effects are avoided.

Note however, that T25 sets always got lower global distortion. A fact that confirms the statement that each template set ensures minimal distortion with respect to the training set and not to the other one.

Comparison of numbers from columns labelled T15 and REFT15 shows error rates reduction ranging from 0.2% on speaker GR to 1.0% on speaker AC. Although such improvements in the WER did not seem to be very large at a first glance, they became very encouraging after a detailed analysis of the errors was done. In fact, a large portion of the errors depended both on imprecise word boundaries detections due to the noisy conditions, and especially on speaker GR on poor pronunciation material. The last kind of error can be partly recovered by means of a more accurate and permissive language model.

**CONCLUSIONS**

We have presented an iterative optimization procedure for extracting representative templates of sub-word basic units from a training speech material. The procedure is based on previous methods of bootstrapping by means of forced re-recognition, followed by some clustering of localized unit occurrences. The location and clustering steps are iterated, according to the criterion of finding the "optimum" Continuance alignment and of selecting templates in order to reduce spectral distance with the given alignment. In this way the templates get closer to the optimum, i.e. they constitute a language model that minimizes the overall spectral distance from the training speech material. Experiments have shown that recognition performance does improve with "a few" (say 1,000) bootstrap iterations, and absolute performance has little significance here: recognition rates depend a lot on how well the training material is representative of the test one, and not only on how well templates represent the training material; besides, improvements in recognition rates depend on how bad was the initial bootstrapping, in the sense that with an almost perfect initial bootstrapping we cannot gain anything from the iterative procedure. The only important conclusion here is that the initial bootstrapping should need to be very good; errors are corrected through iterations.

**REFERENCES**


EUROSPRACH '89, Paris, France, September 1989 1082