ON USING A PRIORI SEGMENTATION OF THE SPEECH SIGNAL IN AN N-BEST SOLUTIONS POST-PROCESSING

T. Moudenc, D. Jouvet & J. Monné

France Télécom - CNET - LAA/TSS/RCP
2, route de Trégastel, 22301 Lannion - FRANCE

ABSTRACT

This paper describes a new method for the post-processing of N-best solutions based on stochastic modelling of the number of speech signal stationarity changes which occur within the phonetic segments of each solution.

The objective of this post-processing is to validate the presence of stationarity zones in the speech signal. This particular validation cannot be exploited using a centisecond approach. The signal stationarity changes are detected using an "a priori" segmentation algorithm.

Two phonetic models are calculated for each phonetic segment. One corresponds to correct solutions and the other one corresponds to incorrect solutions. These two models are used simultaneously in order to compute a post-processing score for each solution.

In the initial set of experiments, which was conducted on telephone databases, the use of this method resulted in a 9 % error rate reduction on the "Number" database, and a 15 % error rate reduction on the "Digit" database.

1. INTRODUCTION

Classic HMM speech recognition systems are based on the computation of cepstral parameters on centisecond frames. In order to improve HMM recognition performance, research has been conducted on the use of other parameters (such as acoustical event duration [1]). Further investigation has explored the use of a priori knowledge on the stationarity zones of the speech signal.

In terms of the latter, the traditional approach consists of conducting a variable-frame duration cepstral analysis on segments defined by two consecutive changes in signal stationarity [2, 3]. This paper proposes a new approach to the incorporation of automatic a-priori segmentation into an HMM based speech recognizer.

Initially, an N-best-solutions algorithm [4, 5] is used. Such algorithms provide the list of the N most likely solutions.

The post-processing method is designed to extract the correct solution from those mentioned above (i.e. N most likely solutions). Usually, the post-processing algorithms are designed to process information different from that ones used in the computation of the N-best solutions. The post-processing method presented herein is based on a discrete stochastic modelling of the number of stationarity changes detected within phonetic segments.

Stationarity changes are detected using an "a priori" segmentation algorithm [2]. This technique is based on the estimation of two statistical models of the speech signal on two different windows: a sliding window, and an expanding window. These models are recalculated for every new sample. A statistical divergence test is then computed in order to estimate the difference between the two models. When the computed difference is higher than a given threshold, a change is detected and a mark is placed on the signal.

Discrete densities are used in order to model the number of stationarity changes detected inside the boundaries of the segments. Because each segment is associated to a phoneme, the models are estimated for both context-dependent units and context-independent units.

Furthermore, two models are estimated for each unit. One is associated with the correct solutions and the other one with the incorrect solutions. The two models are used simultaneously in order to compute a post-processing score for each solution.

The speech recognition system used in the calculation of the N-best solutions is based on a Hidden Markov Modelling approach. The input vectors consist of the cepstral parameters and the first and second temporal derivatives, which are computed on a 5-frame window. The sorted list of the N-best solutions is provided with the HMM's scores [6].

Section 2 describes the phonetic modelling of the number of stationarity changes and Section 3 details the post-processing task. Experiments have been conducted on telephone speech databases. The results are provided in Section 4. The final section elaborates on the results and suggests further improvements.

2. PHONETIC MODELLING

Two classes of phonetic units are examined herein: context-dependent units, and context-independent units. The corresponding models are estimated on the training set.

2.1 Correct and Incorrect Segments

Each phonetic segment in the training set belongs either to a correct alignment ("the" correct solution) or to an incorrect alignment (one of the N-1 incorrect solutions).

0-7803-2431-5/95 $4.00

As such, the "correct" phonetic units and the "incorrect" ones can be modelled.

For example, Figure 1 shows the signal (top of Figure) corresponding to the word articulated in French, "annulation". The N-best solutions calculation provides a set of N words, among which the words "annulation" (the correct solution), and "Lannion" (an incorrect solution). The Markovian alignments associated with those two solutions are represented in the Figure below (middle and bottom part of Figure 1).

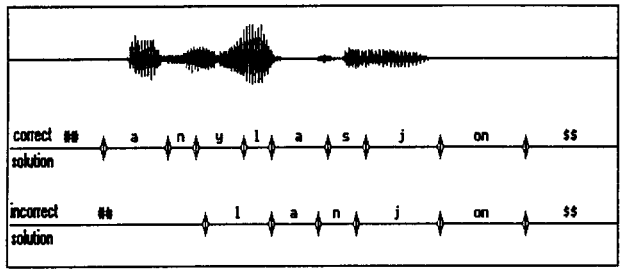

Figure 1 : Two of the N-best Solutions for the Word "annulation"

Between the two alignments, the segments "a", "j", "on" and "$$" (the final silence) of the incorrect solution, overlap the corresponding segments of the correct solution. As such, they are not regarded like "true" incorrect segments. Thus, segments of incorrect solutions which overlap a corresponding correct segment with an identical phonetic label by more than 75 % are not validated when estimating incorrect models.

On the contrary, in our example, the segments labelled "##", "l" and "n" of the alignment associated with the incorrect solution, "Lannion", are used to estimate the "incorrect" models of these units.

2.2 Variations in Markovian Boundaries

The implemented discrete distribution modelling consists of an estimation of the number of stationarity changes detected inside the Markovian segments. The positions of the Markovian boundaries may be moved according to a temporal parameter called the Tolerance Step T (specified in milliseconds). The Tolerance Step enlarges (or reduces) the phonetic segments as illustrated in Figure 2.

Figure 2 shows the signal, the HMM alignment (Markov segmentation) and the changes detected by an automatic a-priori segmentation (last line in the Figure). Letters A and B (bottom of the Figure) represent the Markovian boundaries for the "s" phoneme. If the boundaries of the "s" phonetic segment are validated, then only two changes are enclosed in the segment. However, when validating a 30 ms Tolerance Step (T=30), the enlarged segment contains 4 changes. Thus, the Tolerance Step is a modelling parameter.

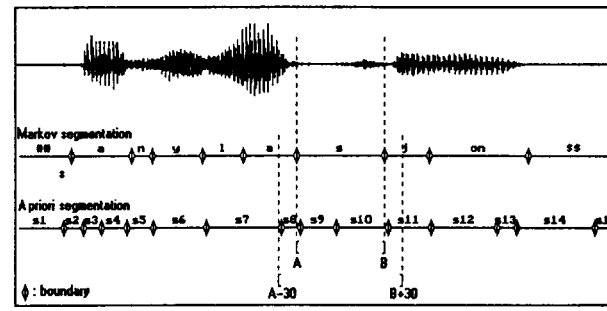

Figure 2 : Enlargement of the "s" Phoneme Segment for a Tolerance Step of 30 ms.

3. SEGMENTAL POST PROCESSING

3.1 Scoring

The segmental post-processing score ($Score_{SPP}$) for each of the N-best solutions is computed as follows [6]:

$$Score_{SPP}(X) = -Log\left(\frac{Pr(X \text{ is correct})}{Pr(X \text{ is incorrect})}\right)$$

where X represents the phonetic segmentation for the current solution.

For a given alignment and a specified Tolerance Step, the probability that the alignment will be correct is calculated as the product of the probability of each segment which constitutes the alignment, using the correct models. As such, segments are presumed to be independent from each other. The probability that the alignment will be incorrect is calculated in a similar way using the incorrect models. The probability for each segment is obtained using discrete probability distributions. The observation is the number of stationarity changes inside the specified Tolerance Step.

In addition, a reliability threshold R is used for context-dependent probability distributions. If the number of context-dependent examples used in order to estimate the corresponding model is lower than this threshold R, then a smoothing of the context-dependent model is carried out using the context-independent model. R specifies how much data is presumed to be necessary in order to obtain a reliable estimate for the discrete probability models.

3.2 Combination with the HMM Score

Once the post-processing score is calculated for a solution, it is linearly combined with the Markovian Score, thus providing the final score for the solution:

$$Score(X) = \alpha\ Score_{HMM}(X) + (1-\alpha)\ Score_{SPP}(X)$$

The solution which delivers the lowest combined score is regarded as the answer.

3.3 The Selection of Diverse Parameters

The parameters which must be optimized are: the Tolerance Step T, the reliability threshold R, and the linear combination parameter α. The set of parameters which yields the smallest error rate on the training set is selected.

4. TESTS AND RESULTS

The N-best solutions are provided using a context-dependent phoneme-based HMM. The acoustical input vectors are formed by 8 cepstral parameters plus the energy, and the first and second temporal derivatives, which are computed on a 5-frame window.

The statistical models of the speech signal in the "a priori" segmentation are estimated using the auto-correlation method for the sliding window (of 20 milliseconds), and the Burg trellis algorithm for the expanding window. The model order is 16.

4.1 Databases

In these experiments, 3 speech databases were used, comprising 800 speakers, collected over the telephone network (mainly long distance calls), covering different regional accents. Each database was split into 2 parts: the training set for training model parameters, and the test set for evaluating recognition performance.

The 3 databases are comprised of: the "Digit" database (ten digits), the "Trégor" database (36 French words) and the "Number" database (numbers between 00 and 99).

4.2 Definition of the Parameter Variation Range

The different context-dependent and context-independent models have been estimated for 16 Tolerance Steps T in the range [-25 ms; 50 ms]. Furthermore, several values were evaluated for the reliability threshold R (from 50 occurrences, up to a few hundreds). The best linear combination parameter α, wich was the last parameter to be estimated, was established in the range [0; 1], and more precisely in the sub-range [0.8; 1].

4.2 Segmental models

The following Table identifies the number of correct context-dependent and correct context-independent units that were modelled for each database.

The first line indicates the number of utterances contained in the training set. An identical amount of data, from different speakers, is present in the test set.

	Digit (0 to 9)	Trégor (36 words)	Number (00 to 99)
Number of utterances on the training set	3600	7300	12700
context-dependent units	45	303	236
context-independent units	22	37	28

Table 1 : Number of Modelled Units

4.3 Tests Results

The following Table cites the error rate reductions obtained on the test set for the parameter values wich provided the smallest error rate in the training set. The first tree rows of the Table give the optimal parameter values. The first row in the second part of the Table cites the error rate obtained with the HMM alone, the second row reports the error rate obtained after the post-processing, and the last row indicates the reduction in the error rate .

	Digit (0 to 9)	Trégor (36 words)	Number (00 to 99)
Reliability threshold R	150	400	225
Tolerance Step T	35	5	-10
Linear combination α	0.92	0.92	0.91
Error rate HMM alone	1.14 %	0.83 %	3.58 %
Error rate HMM+Post-Processing	**0.97 %**	**0.79 %**	**3.25 %**
Error rate reduction	**14.91 %**	**5.64 %**	**9.22 %**

Table 2 : Results Obtained on the Test Sets

The error rate reductions varies from 5.6 % for the Trégor database to 14.9 % for the "Digit" database. However, the optimal values obtained for the training set are not those which will provide the lowest error rate for the test set. To illustrate this fact, the following figures represent the error rate evolution for the 3 databases, according to one variable parameter (the tolerance step) while the two others are maintained fixed (values corresponding to the optimum reported in the above Table).

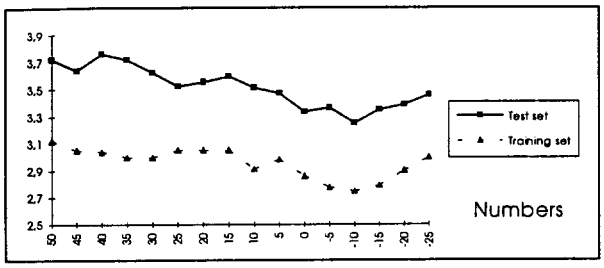

Figure 3 : Error Rate Evolution for Tolerance Step Variations on the "Number" Database

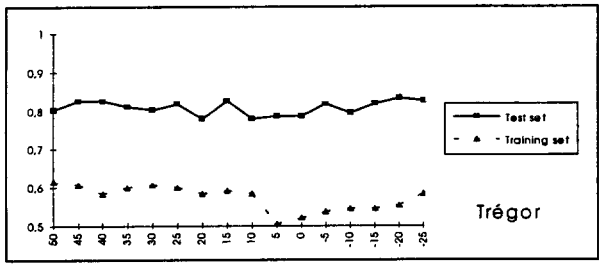

Figure 4 : Error Rate Evolution for Tolerance Step Variations on the "Trégor" Database

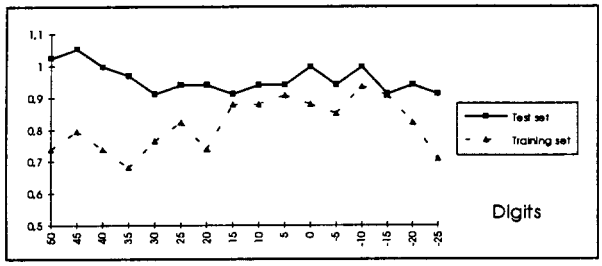

Figure 5 : Error Rate Evolution for Tolerance Step Variations on the "Digit" Database

These curves show that the calculation method for the Tolerance Step may be improved. Indeed, the behaviour of the error rate for the "Number" database (Figure 3) seems to be very similar to the training set and the test set. However, the behaviour of the error rate for the "Digit" database is almost chaotic (Figure 5).

Approximately the same difference may be observed for the different databases for the other parameters.

5. CONCLUSION AND DISCUSSION

In this paper, a new method for improving the recognition performance of an HMM based system which exploits the detection of abrupt changes in signal stationarity has been presented.

In this approach, the number of stationarity changes within a phonetic - Markovian - segment is modelled by discrete probability distributions. In the current implementation, this modelling is performed in the post-processing stage of a classic HMM-based system.

Three parameters were validated and estimated on the training set. The first parameter modifies the position of the Markovian boundaries. The second parameter corresponds to a reliability threshold of the context-dependent models. The third and last parameter represents the linear combination between the HMM score and the post-processing score.

The results obtained at this stage are encouraging and show that the measure provided by the detection of stationarity changes in the speech signal is relevant.

As mentioned, the calculation method for the different parameters is not optimal. Thus, gains in the test error rate reduction can be anticipated while applying a better optimizing method. For example, the Tolerance Step may be chosen by a computation of the overlapping of the "correct" and "incorrect" models. Likewise, the reliability threshold may be optimized using the Estimate Maximise algorithm.

ACKNOWLEDGEMENTS

This work was partially subsidized by the Conseil Régional de Bretagne. The authors would like to express their appreciation to Régine André-Obrecht for providing the "a priori" segmentation programs and to Chafic Mokbel for very fruitful discussions about this work.

REFERENCES

[1] S. E. Levinson, L. R. Rabiner, M. M. Sondhi, "*Continuously Variable Duration Hidden Markov Models for Automatic Speech Recognition*", Computer Speech and Language, Vol. 1, No 1, pp. 29-45, 1986.

[2] R. André-Obrecht, "*A New Statistical Approach for the Automatic Segmentation of Continuous Speech Signal*", IEEE Trans. on ASSP, Vol. 36, No 1, pp. 29-40, January 1988.

[3] R. André-Obrecht, "*Reconnaissance automatique de parole à partir de segments acoustiques et de modèles de Markov cachés*", 18th J.E.P, Montréal, Canada, pp. 212-216, June 1990.

[4] F. K. Soong, E. Huang, "*A Tree Trellis Based Fast Search for Finding the Nbest Sentence Hypotheses in Continuous Speech Recognition*", Proc. ICASSP, Toronto, Canada, pp. 705-708, May 1991.

[5] R. Schwartz, Y. Chow, "*The N-best Algorithm: an Efficient and Exact Procedure for Finding the N most Likely Sentence Hypotheses*", Proc. ICASSP, Albuquerque, USA, pp. 81-84, 1990.

[6] M. N. Lokbani, D. Jouvet, J. Monné, "*Segmental Post-Processing of the N Best Solutions in a Speech Recognition System*", Proc. EuroSpeech, Berlin, Germany, pp. 811-814, September 1993.