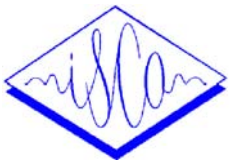

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

EVALUATION OF A SEGMENTATION SYSTEM BASED ON MULTI-LEVEL LATTICES

Jean-Luc HUSSON

LORIA UMR 7503
Bâtiment LORIA BP 239 54506 Vandœuvre-Lès-Nancy Cedex - France
Tel.: +33 (0)3 83 59 20 91 - Fax: +33 (0)3 83 41 30 79
e-mail: husson@loria.fr

10.21437/Eurospeech.1999-121

ABSTRACT

This paper addresses the problem of the evaluation of an automatic continuous speech segmentation system based on multi-level lattices called *dendrograms* previously described in [2] [3]. After a short overview of the segmentation framework, we briefly discuss the difficulties inherent to such an evaluation. We first state quantitative results based on the usual *quality coefficient* and *oversegmentation rate* measures obtained for the automatic paths selection yielded by our system and we estimate the potential improvement margin of our approach. These quantitative results are completed by a qualitative analysis of the system performance by examining the detection accuracy of all possible transitions between 6 broad phonetic classes. The results are judged encouraging given the restricted number of phonetic likelihood criteria used at this stage of development and demonstrates the feasibility of the segmentation process.

1. INTRODUCTION

This paper is devoted to the performance evaluation of a phonetic segmentation system based on multi-level lattices. The lattice, called a *dendrogram*, is an arborescent structure of acoustic segments which potentially contains all temporal segmentation of the signal, from coarse to fine and whose construction uses only acoustic criteria [1].

1.1. Overview of the segmentation system

The segmentation system we aim at evaluating searches for the N most reliable segmentation solution in a exponentially growing search space. This search is made possible by a specific search algorithm based on dynamic programming principles [3] which allows the system to incrementally build the optimal set of most probable solutions. The major advantage of our iterative strategy is that the required processing time required is proportional to the utterance duration even if the number of paths increases exponentially with respect to the global duration. The paths selection relies on a maximum likelihood approach which combines two complementary criteria:

- the acoustic homogeneity criterion which evaluates the probability for one acoustic segment to correspond to a unique phonetic segment, according to its spectral variations;
- the phonetic duration criterion which computes the probability of getting one phonetic segment with regards to its duration using phonetic gamma duration models.

To make these criteria more significant, a prior broad classification stage is applied which assigns to each acoustic segment a probable phonetic class (oral vowels, nasal vowels, fricatives, stops, semi-vowels or sonorants). This allows us to focus on the few models corresponding to the assigned broad phonetic class. Precise information about the learning and probability computing techniques can be found in [7].

The efficiency of our search algorithm is also due to two automatically generated constraints, which dynamically prune the search space. The first one is a local voicing constraint, which forces the paths to fit at best the voicing clues. This constraint rests on a hybrid time/ frequency domain pitch determination algorithm for robust voiced/unvoiced segmentation [4]. The second constraint is a global duration constraint which estimates the probable number of segments in the form of a confidence interval, according to the signal duration. A previous study showed that a very high successful prediction rate (greater than 99.6 %) is achieved for a 5.6 segments mean width, which allows the system to filter more than 99.98 % of the candidate paths. The number N of provided best paths is also dynamically adjusted with regards to the whole signal duration and to the likelihoods assigned to the paths.

1.2. Objectives

In the following sections, we first show that the objective evaluation of a segmentation system is a particularly

difficult task. The usual evaluation techniques based on a few quantitative measures are not informative enough to guide the possible improvements of the system.

The next section deals with the evaluation of our own system. We first state the results obtained for the usual performance measures. These quantitative results are completed with a qualitative analysis of the yielded segmentation solution. We also aim at evaluating the optimal potential performance of our system, which allows us to discuss objectively the obtained results and to evaluate the margin of possible improvement of the chosen approach, according to the fundamental characteristics of our system.

2. EVALUATION OF OUR SYSTEM

The precise evaluation of a segmentation system is a complicated task which provokes many questions: How is a significant test corpus built? Which is the best segmentation reference to use? The reference which is usually taken into account is the manual expert segmentation. But the manual segmentation task has never been precisely and definitely specified. One has to cope with the probable arbitrary local decisions of experts, especially in vocalic regions. Which are the most suitable measures and interpretation rules to be used?

Furthermore, many system properties may be very difficult to evaluate though they directly determine the usability and the performances of a system, as the implementation cost, the dependence to learning and using conditions, the behavior coherence of the system, the possible learning load, the inherent possibilities of interaction of the system with other speech decoding stages.

As a global evaluation is impossible, several quantitative measures have been proposed, each of them evaluating a particular characteristic of the system segmentation compared to the available segmentation reference, like the *oversegmentation rate* (cf. eq. 1) and the *quality coefficient* (cf. eq. 2) [6].

$$OSR = n / N \qquad \textit{(Eq. 1)}$$

where n is the number of automatically detected temporal boundaries and where N is the number of manual temporal boundaries set.

$$QC_\delta = n_\delta / N \qquad \textit{(Eq. 2)}$$

where n_δ is the number of manual temporal segmentation boundaries which are less than δ ms far from an temporal boundary set by the system and where N is the number of manual boundaries.

These measures are very simple to implement but their respective results are not individually significant. Both measures have to be jointly taken into account. The quality coefficient will be surely high if the oversegmentation rate is high. A system leading to an oversegmentation rate near 1 but in the same time to a low quality coefficient is not efficient, as is means that the automatic segmentation boundaries don't correspond temporally to the manual ones. A system will be seem as efficient if both values are simultaneously near 1. Such results exhibit that the system leads to the exact number of temporal boundaries and that these boundaries are good temporal approximations of the reference segmentation ones. The lower the temporal threshold δ, the more efficient is the segmentation system.

This pair of values gives us an idea about the global average quality of the system solutions. However they must be supported by an qualitative analysis of the observed error to allow the system designer to exploit them in order to improve the system. The qualitative study should concern the type of phonemes or phonetic classes which lead to good results or those which are, on the contrary, oversegmented or omitted by the system. One should also classify the types of phonetic transitions which are well or badly detected.

One must remember that the segmentation stage is only a preprocessing and that it's evaluation can't be done independently of the characteristics required by the algorithm which uses its results. The performance of a phonetic segmentation system can only be evaluated by the results obtained by a analytical speech decoding system which exploits these segmentation solutions. We could not manage to carry out such an experiment. Therefore, we first provide here some quantitative results obtained for the usual evaluation measures, and then, we focus on a qualitative point of view by examining the efficiency of our system for all possible transitions between the 6 broad phonetic classes used.

2.1. Quantitative evaluation

Our system is particular because it builds a set of several probable segmentation solutions. The classical quantitative measures are nevertheless still usable if they are applied on an individual path as the best one (judged as the most reliable solution) or the second one. However, it appeared rather difficult to compute from all these particular results a unique score which evaluates objectively the global quality of the set of N selected

paths. In order to permit the comparison of our system with other segmentation frameworks, we present:

- the values of the average quality coefficients CQ_{δ} for three particular values of the allowed temporal imprecision δ (8, 16 and 32 ms) in table 1;
- the average oversegmentation rates obtained by each of the three best candidate paths selected by our system in table 2.

However, these results jointly evaluate the likelihood selection function proposed and the capacity of the dendrogram structure for containing the correct segmentation solutions. In order to focus on the only effect of the likelihood selection function and to cope with the possible temporal errors due to the dendrogram construction technique, we decided to compare the results of the quantitative measures computed for the first three segmentation candidates proposed by our system and the results obtained for the three best segmentation solution contained in the dendrogram, given by a specific search algorithm. This DTW algorithm examines all the paths of the dendrogram and builds the set of the three paths whose alignment costs with the manual segmentation are the lowest. Tables 3 and 4 shows the quality coefficients and the oversegmentation rates obtained by this new reference solution. These results allow us to compare the performance actually reached by our segmentation system to the best results that could be reached with on optimal likelihood selection function.

The test corpus in made with about 60 seconds of clean speech and consists of data from 5 speakers. The reference solution taken into account for all the tests carried out here is the manual segmentation of a speech decoding expert and whose intrinsic coding precision is 8 ms.

Results obtained for the three first paths selected by our automatic segmentation system

	Path 1	**Path 2**	**Path 3**
QC_8	*0.82*	*0.80*	*0.75*
QC_{16}	*0.85*	*0.85*	*0.77*
QC_{32}	*0.90*	*0.89*	*0.79*

Table 1: Quality coefficient obtained for the first three paths selected by our system

	Path 1	**Path 2**	**Path 3**
OSR	*0.85*	*1.01*	*0.96*

Table 2: Average oversegmentation rates obtained for the first three paths selected by our system

Results obtained for the three best paths contained in the multi-level lattices

	Path 1	**Path 2**	**Path 3**
QC_8	*0.91*	*0.91*	*0.83*
QC_{16}	*0.92*	*0.94*	*0.88*
Q_{C32}	*0.93*	*0.95*	*0.91*

Table 3: Quality coefficients obtained for the three best paths contained in the dendrograms

	Path 1	**Path 2**	**Path 3**
OSR	*0.96*	*1.28*	*1.18*

Table 4: Average oversegmentation rates obtained for the three best paths contained in the dendrograms

Discussion

The comparison of the quality coefficients (tables 1 and 3) shows that the best segmentation solutions contained in the dendrograms are not systematically classified by our system within the set of the best three candidates. This does not necessarily mean the these best solutions are not retrieved by our system as only three candidates were taken into account in our tests. However this statement should question the efficiency of the proposed likelihood selection function to classify accurately the candidate paths. The provided results do not allow us to validate or disprove these hypothesis. Complementary tests were consequently carried out on the same corpus to evaluate more objectively the performance of our segmentation system. The results showed that for about 70 % of the dendrograms computed on the corpus, the best contained path is retrieved by our system and classified as the N provided candidate paths. For 83 % of such cases, it is classified as the most probable candidate path and for 98 %, the path appears in the set of best three candidates. These results prove that the proposed selection function, though it is quite simple, is efficient enough to select reliable segmentation solutions among the large number of candidates. Despite this, several improvements are obviously necessary. The improvement margin is estimated to be 30 % (for 70 % of the dendrograms the best contained path is retrieved by our system).

The comparison of the oversegmentation rates obtained for the system paths selection (table 4) and for the best contained paths (table 2) clearly show that our system usually leads to undersegmentation cases, which may appear more serious for the decoding performance than in oversegmentation case. This phenomenon is essentially due to the way the likelihood of one path is computed given the phonetic likelihood coefficients assigned to each acoustic segment. The processing technique we use tends to favour the paths whose number of segments is low. A possible alternative solution may consist of the normalization of the path likelihood by the number of its

segments as proposed by Hajislam in similar conditions [5] (cf. eq. 3)

$$p(C) = \sqrt[n]{\prod_{k=1}^{p} p(S_k)} \qquad \textit{(Eq. 3)}$$

Where C is a path made of p segments S_k.

2.2. Qualitative evaluation

Test protocol

We state here the results of an experiment carried out to evaluate the qualitative behavior of our system. Our goal is to measure the capacity of the segmentation tool to locate on the signal all possible transitions between the chosen broad phonetic classes. The corpus used for this experiment is made up of continuous speech recorded from 15 different speakers (men and women). Half of this material consists of clean speech and the reminding part of news reports recorded from a French radio station. A particular phonetic transition of the manual segmentation is judged accurately detected by the system if a boundary for at least one candidate path among the best three is located in a 16 ms range.

Results

Spreadsheet 5 summaries the results obtained for the 6000 phonetic transitions of our test corpus. Beforehand, each segment has been assigned a phonetic class label according to the corresponding labelling phoneme in the manually-performed transcription. The abbreviations mentioned in table 5, *i.e. OV*, *NV*, *F*, *S*, *SV* and *SO*, indicate respectively the following broad phonetic classes: oral vowels, nasal vowels, fricatives, stops, semi-vowels and sonorants. The *SIL* abbreviation class identifies acoustic segments corresponding to silence or noise.

	OV	*NV*	*F*	*S*	*SV*	*SO*	*SIL*
OV	*48 %*	-	*97 %*	*97 %*	*65 %*	*68 %*	*93 %*
NV	-	-	*99 %*	*96 %*	-	*35 %*	*100 %*
F	*98 %*	*95 %*	-	*91 %*	*83 %*	*71 %*	*100 %*
S	*95 %*	*95 %*	*92 %*	-	*87 %*	*82 %*	-
SV	*72 %*	*65 %*	*100 %*	-	-	-	-
SO	*70 %*	*71 %*	*91 %*	*71 %*	*62 %*	*58 %*	*100 %*
SIL	*100 %*	*100 %*	*100 %*	*61 %*	-	*100 %*	/

Table 5: Percentage of transitions between broad phonetic classes accurately detected

3. CONCLUSION

The obtained results reflect the usual inherent difficulties of segmentation and decoding tasks. The segmentation of the phones which exhibit distinctive spectral characteristics (as stops and fricatives for example) is relatively easy whereas the segmentation of vocalic areas is much more complex. In this case, the acoustic and duration criteria used are not sufficient.

Nevertheless, we consider these first results as encouraging given the restricted number of phonetic likelihood criteria used at this stage of development of the system. The evaluation shows that the proposed segmentation process was feasible.

The development of these automatic evaluation tools appropriate for the system will be very useful since it will allow us to improve the system continuously through a back-and-forth process between a system modification stage and a test stage.

4. REFERENCES

[1] Glass J. R. and V. W. Zue (1988) *Multi-Level Acoustic Segmentation of Continuous Speech.* Proc. ICASSP-88, pp. 215-218

[2] J.-L. Husson and Y. Laprie. *A new search algorithm in segmentation lattices of speech signals*. Proc. ICSLP'96, vol. 1, pp. 2099–2102, Philadelphia, PA, Octobre 1996.

[3] J.-L. Husson and Y. Laprie. *Searching for the n-best segmentations in dendrograms*. Proc. SPECOM'96, pp. 101–106, St-Peterburg, Octobre 1996.

[4] J.-L. Husson and Y. Laprie. *Cooperation of frequency and time-domain methods for pitch tracking*. *Proc. SPECOM'98*, pp. 293-298, St-Peterburg, Octobre 1998.

[5] R. Hajislam. *Décodage acoustico-phonétique et robustesse en reconnaissance automatique de la parole*. Phd thesis, University Henri Poincaré, Nancy 1, 1994.

[6] H. Kabré, G. Pérennou and N. Vigouroux. *Automatic labelling of speech into events*. Proc. ICPhS'91, 1991.

[7] J.-L. Husson. *Une approche hiérarchique de la segmentation automatique de parole*, Phd thesis, University Henri Poincaré, Nancy 1, 1998.