

A NEW SEARCH ALGORITHM IN SEGMENTATION LATTICES OF SPEECH SIGNALS

Jean-Luc HUSSON, Yves LAPRIE

CRIN-CNRS & INRIA Lorraine, BP 239 - 54506 Vandœuvre-Lès-Nancy
Tel.: (33) 83 59 20 91 - Fax: (33) 83 41 30 79 - e-mail: husson@loria.fr

ABSTRACT

This paper describes a new segmentation system using a multi-level representation, called dendrogram [1]. We address the issues of estimating the confidence of one path, and finding the N most reliable paths in the segmentation lattice. Our approach rests on automatically trained criteria and on an efficient strategy to prune the search space.

1. INTRODUCTION

In a analytical speech recognition system (based on acoustic and phonetic knowledge), the efficiency of the phone identification stage strongly relies on the quality of the segmentation. This justifies the numerous researches devoted to the design of robust and reliable automatic segmentation algorithms. Most of those rest on the search of temporal or spectral continuity breaks and lead to the following drawbacks :

- They yield a unique segmentation solution.
- They come against the difficulty of locally computing the threshold which minimizes globally the rate of false segmentations (undersegmentation and oversegmentation cases).

To solve these problems, some algorithms which produce a multi-level segmentation lattice have been proposed, like the dendrogram [1] which uses the spectral representation of a speech signal. Fig. 1 shows an example of a dendrogram computed on a 600 ms voiced speech signal.

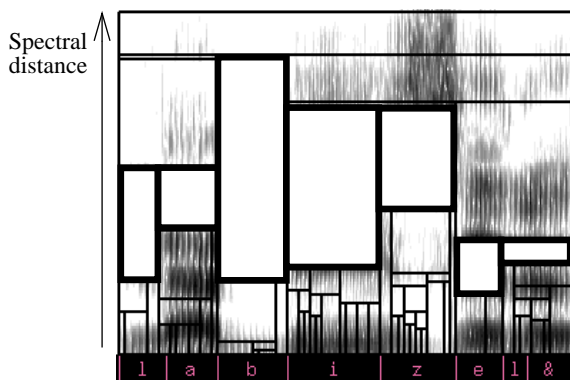


Figure 1: Example of dendrogram (75 segments, \approx 400 000 paths).

The advantages of this structure are numerous :

- It potentially contains all the decompositions (from coarse to fine) of a speech signal, in a uniform hierarchical (arborescent) structure and exhibits all the significant values of distance. A previous study [2] demonstrated that the structure always contains the correct segmentation (the one an expert would accept).
- The segments are hierarchically classified through the structure in accordance with their spectral homogeneity.
- The segmentation is independent of any level of segmental representation because all levels are potentially represented from subphonemic events to words.

In return to these advantages, a dendrogram presents a high complexity. The tremendous number of contained paths (in the order of 2 billions of paths for 2 seconds of speech) prevents the exhaustive examination of the whole set of solutions by the recognition algorithm. An automatic selection algorithm of the N-best paths of the dendrogram is thus required, but in spite of several previous attempts [3], this problem has not received any satisfactory solution yet.

In this work, we propose a general framework to solve this problem and to find automatically the most reliable paths in a dendrogram.

We address the issues of:

- estimating the likelihood of one path in the dendrogram,
- finding the best paths in a search space growing exponentially with respect to the utterance duration.

2. ESTIMATING THE LIKELIHOOD OF ONE PATH

We choose to compute the likelihood of one path C of the dendrogram as the product of its segment probabilities (under classical independence assumption). In practice, we use the logarithm of this likelihood :

$$\log[p(C)] = \sum_{S_i \in C} \log[p(S_i)]$$

The confidence $p(S)$ of one segment is obtained by combining two probabilities, each of them corresponding to a particular criterion of its phonetic likelihood.

- The acoustic homogeneity probability
- The segments duration confidence

2.1. The acoustic homogeneity probability

Intuitively, the less homogeneous a segment is, the less it may belong to the final solution. To exploit the arborescent structure of the dendrogram, which implicitly codes the spectral consistency of segments, we choose to estimate the spectral homogeneity of a segment S by a distance measure between the representative vectors of its sons in the dendrogram. The representative vector of a segment S is its average spectrum (a 12 MFCC vector) and is noted $Vr(S)$.

Instead of using an arbitrary spectral distance which produces no probability, we decided to use the hand labelled segmentation and the difference vectors (between the representative vectors of the sons of the segments) for the 20 000 segments of the training dendrograms, to modelize a gaussian random variable. The homogeneity probability of a segment in the dendrogram is the value of the random variable given the difference vector of its sons. A segment without son (at the bottom of the dendrogram) is given an arbitrary homogeneity probability of 1. If a segment has more than 2 sons, a special computation is needed that will not be reported here.

2.2. The segment duration confidence

We intend to estimate the confidence in segment S whose duration is d_S in the form of the probability $p_{duration}(S)$. Our first attempt was to estimate this probability from the training of the duration distribution of all the segments of our corpus, without class distinction. This coarse criterion proved to be not very selective on account of the high duration disparity. In order to make this criterion more discriminating, we first use a broad phonetic classification of the segment and the final probability is the highest

probability to observe the given duration for the phones of the predicted class $c_{max}(S)$:

$$p_{duration}(S) = \max_{h \in c_{max}(S)} [p\langle d_S|h \rangle]$$

This requires to build a reliable **broad phonetic classifier** (cf. **The broad phonetic Bayesian classifier**) and to **modelize the duration distribution** for each of the french phonemes (cf. **The phone duration criterion**).

The broad phonetic Bayesian classifier

The aim of the classification method is to predict the phonetic class a given segment S most probably belongs to. We did choose a standard partition (θ), dividing the french phonemes in 6 broad phonetic classes :

$\theta = \{oral_vowels, nasal_vowels, fricatives, stops, semi-vowels, sonorants\}$

We assign to each segment S of the dendrogram the most probable class hypothesis $c_{max}(S)$:

$$c_{max}(S) = \operatorname{argmax}_{c \in \theta} [p_{class}\langle c|Vr(S) \rangle]$$

with

$$\begin{cases} p_{class}\langle c|Vr(S) \rangle = \frac{p\langle Vr(S)|c \rangle \cdot p(c)}{p(Vr(S))} \\ p(Vr(S)) = \sum_{c \in \theta} p\langle Vr(S)|c \rangle \cdot p(c) \end{cases}$$

Each of the 6 classes has been modeled by a 12 dimensions gaussian random variable (the number of MFCC of the representative vectors).

The computation of the parameters of each gaussian is based on a training algorithm using a french corpus of about 8500 phones. The probabilities $p(c)$ of the classes observation issue from a statistical exploitation of a 300 000 phones french corpus [4]. Table 1 shows the results of the test of our broad classifier on a 3300 phones corpus in the form of a confusion matrix. The classifier may be improved but the preliminary results appeared satisfactory enough to implement the phone duration criterion.

	O_V	N_V	Fr	St	S-V	So
O_V	89.38	2.34	1.06	1.70	3.82	1.70
N_V	0.22	93.95	0.0	0.90	3.81	1.12
Fr	0.0	0.0	91.22	6.34	0.98	1.46
St	1.30	0.53	6.33	57.09	11.58	23.17
S-V	6.09	1.63	7.28	10.85	65.53	8.62
So	2.07	1.38	5.52	10.34	4.14	76.55

Table 1: Confusion matrix in %.

The phone duration criterion

The probability $p\langle d_s|h \rangle$ to observe a phoneme h of duration d_s has been modeled by a gamma density function [5] for each french phoneme, using a training corpus of about 8500 hand labelled segments.

The obtained modelization is not context-dependent and does not take the phone position in the rhythmic group into consideration. However, an automatic duration normalization is currently under development. This improvement may increase the duration modelization efficiency and then the selective ability of this criterion.

3. THE SEARCH ALGORITHM

A dynamic programming algorithm, has been developed to select the N-best paths of the dendrogram. This algorithm estimates the confidence of each path of the structure as described before and builds step by step the N-best paths according to a maximum likelihood criterion. The processing time required by our iterative strategy is proportional to the utterance duration even if the number of paths increases exponentially with respect to the global duration.

However, dendrograms usually contain many improbable paths, whose probability is nevertheless estimated. That is why two complementary constraints have been added to the initial algorithm to prune the search space and limit the hypothesis set to the most probable sub-set.

The first constraint is a *local voicing constraint* (cf. 3.1) which removes the segments which are not phonetically realistic. The second constraint is a *global duration constraint* which removes the paths whose number of segments is too high or too low (cf. 3.2).

3.1. The local voicing constraint

Theoretically, the best paths must fit the voicing boundaries, because a phone cannot be simultaneously voiced and unvoiced. One of many ways to ensure that the voicing constraint is fulfilled is to split up the dendrogram D of the entire utterance into the set of adjacent sub-dendrograms of D which fit the voicing boundaries at best, the voicing boundaries being automatically defined by the pitch determination algorithm of [6].

This strategy removes from the global dendrogram the less probable segments, especially the very long ones. This reduces the structure complexity and proportionally the time of the processing. More, splitting the dendrogram in smaller sub-dendrograms improves the efficiency of the global duration constraint presented in the next section (cf. 3.2).

However, this required to modify the selection algorithm to reconstruct the best global paths from the results obtained locally on the sub-dendrograms. This work is not reported here.

3.2. The global duration constraint

Given the utterance duration, this constraint indicates the number of segments expected, in the form of a confidence interval. Instead of using the constraint once on the whole utterance, it is generated and used dynamically during the search on each voiced or unvoiced region of the sentence. This allows us to prune step by step the search space. We describe now how the confidence interval is computed.

Beforehand, we have represented the duration distributions of all regions (fully voiced or fully unvoiced) from 1 to 15 segments with a gaussian. The maximum number of 15 segments is the result of an automatic training process. This allows 99.8 % of the regions to be well represented with the set of the 15 gaussians. There was not enough observations to represent accurately the remaining 0.2 % of regions, whose number of segments is greater than 15, with a gaussian. We can explain the observation of this small number of implausible regions by some labelling errors and pitch tracker failures.

Then, we are able to compute the probability $p\langle k|d \rangle$ to have a region of k segments given its global duration d :

$$\begin{cases} p\langle k|d \rangle = \frac{p\langle d|k \rangle \cdot p(k)}{p(d)} \\ p(d) = \sum_k p\langle d|k \rangle \cdot p(k) \end{cases}$$

$p\langle d|k \rangle$ is given by the k^{th} gaussian and $p(k)$ is the proportion of regions of k segments in the training corpus.

For each region (of duration d), a confidence interval I_d is built on the following way.

Initially, this interval is :

$$I_d = [k_{max}, k_{max}]$$

with :

$$k_{max} = \operatorname{argmax}_{k \in [1,15]} [p\langle k|d \rangle]$$

... and it is iteratively extended up to satisfy simultaneously the two following conditions :

$$\begin{cases} \sum_{k \in I_d} p\langle k|d \rangle \geq \varepsilon \\ \text{Width}(I_d) \text{ is minimal} \end{cases}$$

We have distinguished voiced and unvoiced regions, because their duration probability are very different. We carried out tests to study the effect of ε on the constraint efficiency. Fig. 2 shows (in the voiced case) the evolution of the successful prediction rate (R) according to the mean width (\bar{I}_d) of the confidence intervals.

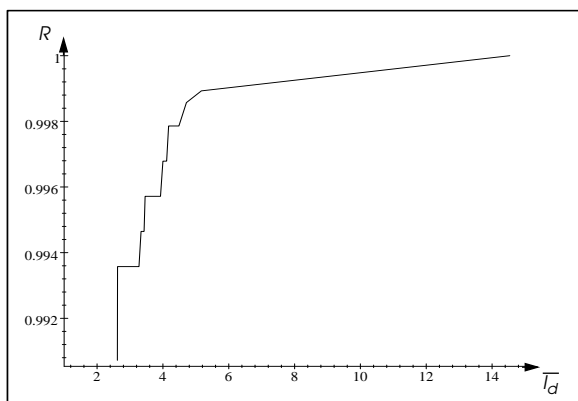


Figure 2: Evolution of the successful prediction rate

It appears that a particularly short mean width (2.6 segments) already gives a very high prediction rate (> 99 %). Furthermore, this graph shows that a very high successful prediction rate (99.8 %) is achieved for a 4.2 segments mean width, which is still much lower than the mean standard deviation of the segment number of the paths in the sub-dendrograms. The corresponding value of the threshold ε is 0.99. This demonstrates the reliability of the prediction algorithm and proves that the global duration constraint is very efficient. As an example, for the voiced region of Fig. 1, the global duration constraint predicts the confidence interval [5,12], which allows the system to focus on only 1491 paths among the initial 389 590 paths.

Though the risk ($1-\varepsilon$) is minimal, the case where no path satisfies this constraint is still possible. To avoid this irremediable error, we have implemented a retrieving technique which, in this case, selects the most probable paths independently of this constraint.

3.3. Constraint effect

The combination of the two presented constraints shows several major benefits. They remove the segments and the paths whose probability is too low from the search space. So, they strongly reduce the computational time and then improve the efficiency of the selection function.

More, tests show that adding these constraints to the algorithm does not introduce additional errors. The global duration constraint does

not remove correct paths. The pitch determination algorithm has been extensively tested so as to avoid serious voicing pre-segmentation errors.

4. RESULTS AND DISCUSSION

We address the issues of finding the N-best segmentation paths in a multi-level structure. In opposition to others algorithms which use heuristic rules, our approach rests on systematic training and on automatic adjustment of the system parameters.

We have initiated an automatic evaluation of the results of our system by using a dynamic time warping algorithm to align automatically the selected path and the hand labeled solution. First results demonstrate the efficiency of the search algorithm and validate the proposed framework. As an example, Fig. 1 shows the highlighted segments of the best path selected by our system. The omission of the phoneme /l/ in this path may constitute an error, but it may be easily omitted when the phonetic transcription is not given. This demonstrates the difficulty to define an objective evaluation method of a segmentation system. It may take into account a possible hand labeling mistake, the impossibility, even for an expert, to segment precisely some vocalic regions as "un voyageur", the fact that a 20 ms segmentation error is more serious between a oral vowel and an unvoiced stop than the same error between the phonemes /l/ and /&/ of our example (Fig. 1). The usual results given (the omission rate, the insertion rate and the temporal mean error) may consequently lead to misinterpretation. Our work in this area will be reported elsewhere.

Our algorithm appears very effective and turns out to be well suited to incorporate additional criteria and pruning constraints. It also may be adapted to other multi-level hierarchic structures [7].

5. REFERENCES

1. Glass J. R., V. W. Zue (1988) "Multi-Level Acoustic Segmentation of Continuous Speech", *Proc. ICASSP-88*, pp. 215-218
2. Hajislam R. (1994) "Décodage acoustico-phonétique et robustesse en reconnaissance automatique de la parole", *Thèse de doctorat de l'université Henri Poincaré-Nancy I*
3. Hübener K., A. Hauenstein (1993) "Controlling search in segmentation lattices of speech signals", *Proc. EUROSPEECH '93*, V. 3, pp.1763-1766
4. Tubach J.-P., L.-J. Boe (1985) "Un corpus de transcriptions phonétiques : constitution et exploitation statistiques", Rapport ENST 85D001
5. Burshtein D. (1995) "Robust parametric modeling of durations in hidden Markov models", *Proc. ICASSP-95*, pp. 548-551
6. Martin Ph. (1982) "Comparison of pitch detection by cepstrum and spectral comb analysis", *Proc. ICASSP-82*, pp. 180-183
7. Witkin A. P. (1983) "Scale-space filtering", *IJCAI-83*, pp.1019-1022