USING DECISION TREES TO CONSTRUCT OPTIMAL ACOUSTIC CUES

Sandrine Robbe[1], *Anne Bonneau*[1], *Sylvie Coste*[2], *Yves Laprie*[1]

1: CRIN-CNRS & INRIA Lorraine, BP 239
54506 Vandœuvre-lès-Nancy Cedex
2: CRIL & IUT de Lens, Université d'Artois
Rue de l'Université, SP 16, 62307 Lens Cedex
FRANCE
{robbe,bonneau,laprie}@loria.fr coste@lens.lifl.fr

ABSTRACT

This paper presents an approach to the optimization of acoustic cues used for stop identification in the context of an acoustic-phonetic decoding system which uses automatic acoustic event extractors (a formant tracking algorithm and a burst analyzer). The acoustic cues have been designed on the basis of acoustic studies on stops and spectrogram reading experiments. This ensures that these cues have a certain amount of discriminating power but we do not know either the optimal thresholds nor which combination of cues are the most efficient.

Therefore, we propose to use the decision tree theory [4] to choose the most discriminating cues and to improve their discrimination power. Considering the stop occurrences of a training corpus, the best cues are those which allow the decision tree leading to the best partition to be constructed. We have considered all the cues derived from the ones provided by the phonetician on formant transitions and burst characteristics. The improvement of the cues has been achieved on a corpus of 941 stops.

1. INTRODUCTION

Numerous acoustic studies and spectrogram reading experiments [5, 1] contributed to find acoustic cues which characterize the stop place of articulation. By exploiting these results and automatic extraction algorithms for formant trajectories and burst characteristics [7] we developed an acoustic-phonetic decoding system, called Daphné, adapted to sentences constructed from vowels and stops [6].

The basic acoustic cues provided by the extraction algorithms, such as formant frequencies or spectral peaks of burst, are combined into more elaborated cues, called complex cues, which have been designed with the help of a phonetician. More precisely, two classes of cues have been distinguished [3]: **strong** cues whose high discriminating power allows direct identification or rejection of the place of articulation without considering any other information, and **weak** cues of lower discriminating power.

These cues involve a number of thresholds on energy and frequency ranges that have been chose to obtain the expected behavior. Nevertheless this does not ensure that the thresholds are set to the optimal values, or that slightly different cues would not be better. The purpose of this work is to ensure that the set of cues retained is optimal at least with respect to corpora on which the Daphné system can be trained. Note that the problem is not to construct the absolute best cues, because it is hard to imagine all the possible acoustic correlates, but rather to optimize the cues proposed by the phonetician. As the strong cues have turned out to be of great interest because they allow our system to maintain the consistency of the decoding process we want to favor their emergence.

We propose to use the decision tree theory to improve the discrimination power of acoustic cues. The general idea is to consider the possible cues as questions that allow the training corpus to be split up with respect to the value returned by the cues for each of the stop in the corpus. The best cues will be those which allow the decision tree leading to the best partition to be constructed.

This approach raises the following issues:

- Which set of complex cues to accept for constructing the decision tree, i.e, which are the cues which will be derived by varying thresholds and other parameters ?
- Which is the most appropriate goodness-of-split evaluation criterion ?
- Which could be the strategy favoring the emergence of strong cues at the top of the tree ? This question is closely linked to the previous one.
- Last but not least, this work is intended to aid the phonetician in his work. This means that he should be able to interact with the system during the tree construction, for example by obliging the system to accept a given cue even if it is not the most discriminant at the current node. This may be the case when the user considers that a acoustic cue can be exploited in a wider phonetic context than that under investigation. Consequently, this means that the user accepts a final result

which slightly differs from the best according to the evaluation criterion.

In this paper we first review the acoustic cues defined for the identification of the place of articulation of French stops followed by back vowels. Then we present our approach to control the tree construction. Then we describe experiments carried out with voiceless stops followed by back vowels, and discuss the results.

2. OVERVIEW OF THE ACOUSTIC CUES

These cues were previously detailed in [2].

2.1. Cues provided by the burst

The following cues are designed only for voiceless stops followed by back vowels. These cues could be adapted for voiced stops.

- *Strong preference cue for velars:* the energy is concentrated around a prominent peak situated in front of the F2 of the subsequent vowel.
- *Weak cue for velars:* the energy maximum is situated in front of the F2 of the subsequent vowel.
- *Strong exclusion cue for velars:* the lack of a peak in the vicinity of the F2 of the following vowel.
- *Strong preference cue for dentals:* a relatively prominent peak in high frequencies.
- *Weak cue for dentals:* the energy maximum is situated in high frequencies. A maximum situated in the locus frequency region is not taken into account [1].
- *Weak cue for labials:* the energy maximum is situated in low frequencies.

We did not define a strong exclusion cue for the dental bursts since it would be relatively redundant with the strong preference cue of velars. We do not believe labial bursts provide any "strong" evidence of the labial place, consequently we defined no "strong" labial cue.

2.2. Cues provided by the transitions

The following cues are designed for voiced and voiceless stops followed by back vowels.

- *Weak cue for velars:* the transition of F2 is relatively flat.
- *Weak cue for dentals:* the transition of F2 comes from the dental locus.
- *Strong exclusion cue for dentals:* the transition of F2 certainly does not come from the dental locus.
- *Weak cue (1) for labials:* the transition of F2 is rising.
- *Weak cue (2) for labials:* the transition of F3 is rising.
- *Strong exclusion cue for labials:* the F2 transition of the vowel following the stop is clearly falling, although the vowel preceding the stop is not more anterior than it[1].

2.3. Choosing the thresholds

Our cues involved the use of numerous thresholds, in particular:

- the limits of the frequency regions to be investigated, notably for the burst cues (we do not consider the whole spectrum), and for the dental locus region;
- the frequency limits of the high and the low frequency regions, for the weak dental and labial cues;
- the emergence thresholds used to determine whether the peak is very prominent (strong burst cues) and whether a peak dominates the spectrum (weak burst cues);
- the slopes of the transitions ...

In a first time, we have chosen the thresholds empirically. The aim of the present work is to validate the cues and optimize the thresholds.

3. CONSTRUCTING THE DECISION TREE

The idea is to use decision trees techniques [8] in order to infer the discrimination power of the acoustic cues described before and also to choose between several versions of the same cue the most discriminant one (in order to optimize the choice of the thresholds present in the implementation of the acoustic cues for a given corpus). We consider all the stop occurrences of a corpus. The principle is to derive a decision tree with the best partition of the corpus, choosing at each level the acoustic cue and its version which lead to the best partition. In the following, we explain which criterion we use for choosing a cue and a version among its various ones. Then we give the global construction strategy.

3.1. Which goodness-of-split evaluation criterion ?

The choice of the next cue partitioning the corpus is made by selecting the cue (and its version) which enables the greater information gain.

The evaluation of the goodness-of-split could amount to measure the entropy reduction at every node of the tree. Hence, the cue (and its version) chosen at a node is the one which

[1] This condition is necessary to take into account the following coarticulation phenomenon: If one vowel of the / V labial V / context is far more anterior than the other and if the two vowels are strongly coarticulated, F2 of the less anterior vowel does not follow the expected direction.

maximizes entropy reduction. The information contained by the corpus (the set of stops present at the node) is defined by : $M(Corpus) = -plog_2p - tlog_2t - klog_2k$ where p is the probability for an item (a stop of the node) to be a labial, t the probability of an item to be a dental and k the probability to be a velar. The probability is approximated by the relative frequency in the corpus (p is the proportion of labial in the considered set - hence present at this node).

For every cue tested at a node (containing a set of stops or corpus), the entropy reduction is calculated : $M(Corpus) - B(Corpus, cue, version)$ with $B(Corpus, cue, version) = P(N_{yes}) \times M(N_{yes}) + P(N_{no}) \times M(N_{no})$ where $P(N)$ is the relative population of the node with respect to the population of the corpus and $M(N)$ is the information contained at node N (this version of this cue is detected or not). The choice of the next cue (and its version) is made in considering the cue (and its version) which maximizes the information $M(Corpus) - B(Corpus, cue, version)$, hence the cue and version which minimizes $B(Corpus, cue, version)$.

We add a heuristic rule in order to discard cues concerning very few examples.

3.2. Construction strategy of the decision tree

The different versions of an acoustic cues are derived by varying the thresholds used in the cue implementation. The number of versions of a cue depends on the "complexity" of the cue defined with the phonetician.

If a version of a cue is selected at a level in the decision tree, all other versions of the cue are removed from the set of remaining cues (for the next step).

Obviously, if the partition were perfect, the leaves of the decision tree should only consist of stops corresponding to the same phoneme. Obtaining such a tree is not always possible. Stopping criteria are used in order to decide if it's worth continuing the partition. A first very simple criterion stops the construction if the error[2] obtained at a node is smaller than a threshold value. A second criterion does not make the partition if the error of the node is smaller than the average error of its sons weighted by the relative frequency. Furthermore, we do not split up a node if the population concerned contains less than 5% of the total set of examples.

Pruning the tree after the construction is also possible if the error of a branch (defined as the average error of its leaves weighted by their relative frequency) is greater than the error of the current node.

Interactive construction is also considered in the following way : the phonetician controls the choice of a cue (but not its version) at each level in order to guide the overall construction. Note that this control must respect more or less the decision of the automatic process. The phonetician may only select a cue among cues giving similar results for the partitioning. In this way, he can put strong cues at the top of the tree. Hence, he eliminates the item which are "easy" to classify and concentrate on the others.

[2] number of examples misclassified / number of stops in the corpus set

4. EXPERIMENTS

We have tested the construction of the decision tree on two different French corpora. The first one, VERLOC, was recorded in an office without any special care and contained 17 sentences spoken by 16 male speakers. The sentences were pronounced 3, 4 or 5 times. The second one, recorded in an anechoic chamber, is constituted of 22 read sentences made up of stops and vowels, where each sentence was repeated 3 times by 4 male speakers. From these corpora two files were extracted, the first one with 779 voiceless stops followed by back vowels, and the second one with 162. We have then constructed two decision trees, one for each file of stops, using the 12 cues derived from the basis cues provided by the phonetician. In the construction of the tree, we have implemented one to three different version of each cue with a view to adjusting as well as possible the different threshold values.

5. RESULTS AND DISCUSSION

It is worth noticing that we have used the tree to study acoustic cues and not to identify stop place. We obtain two different trees, one for each corpus we have tested. An example of tree is given on figure 5. The leaves represent the best partition of the initial corpus we can obtain with the threshold values we decided to test.

One must keep in mind that these results are associated to two corpora made up of clean and carefully uttered speech. When dealing with spontaneous speech, it is probable that thresholds would be different and even that the hierarchy of cues could change.

Here come some interesting results, although somewhat expected.

- The cues which have the best discrimination power are cues based on the burst. This was quite normal for voiceless stops (tested here) and we used the discriminating power of burst cues to define strong positive cues. Nevertheless, with regard to voiced stops and stops in spontaneous speech the relative importance of transitions and bursts may change.
- Strong cues are not always chosen before weak cues due to the strategy used. Our strategy favors cues which discriminate a great number of stops, another strategy would be to favor strong cues, which discriminate less stops but make no error.
- The dental locus was found to be between 1300Hz and 1800Hz.

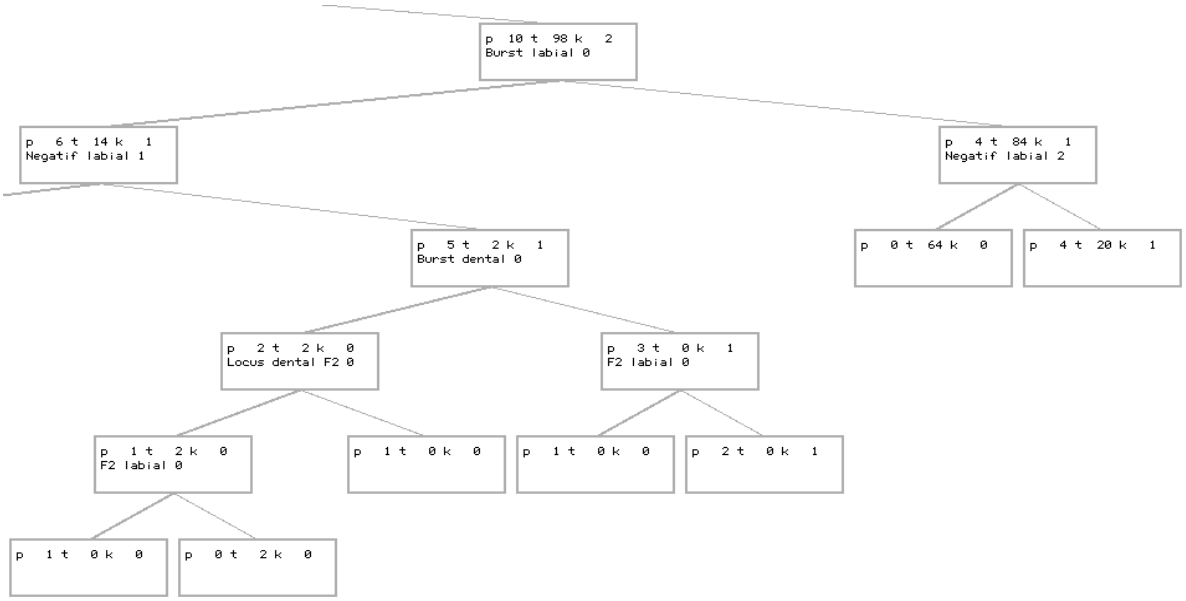

Figure 1: Part of a tree

Note that the robustness of the acoustic cue detectors is implicitly taken into account since a cue based upon a bad detector will be discarded during the tree construction. This behavior is particularly noticeable with the view of designing strong cues (which cannot be subjected to errors and must be robust), and also in the perspective of constructing a recognition system because of the elimination of too fragile cues.

6. CONCLUSIONS

We have presented an approach to the optimization of acoustic cues based upon decision tree theory. This approach has been tested on a corpus of 941 voiceless stops followed by back vowels. It has allowed us to determine the best thresholds, to validate a set of complex acoustic cues and to determine the best cues. Until now, we have used the tree for a well defined application, but it offers many other possibilities.

- One can use the tree to test basic cues, such as formant frequencies or spectral peaks of burst.
- For the moment, the tree construction is automatic, but later on, the phonetician must be able to interact during the construction stage.
- The tree could highlight different behaviors of the cues according to various speech styles.

7. REFERENCES

1. S. E. Blumstein and K. N. Stevens. Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. *J. Acoust. Soc. Amer.*, 66:1001–1017, 1979.
2. A. Bonneau, S. Coste, L. Djezzar, and Y. Laprie. Two Level Acoustic Cues for Consistent Stop Identification. In *Proceedings International Conference on Spoken Language Processing*, volume 1, pages 511–514, Banff (Alberta, Canada), October 1992.
3. A. Bonneau, S. Coste-Marquis, and Y. Laprie. Strong cues for identifying well-realized phonetic features. In *Proceedings of The XIIIth International Congress of Phonetic Sciences*, volume 4, pages 144–147, Stockholm, Sweden, 1995.
4. L. Breiman, J. H. Friedman, R. A. Olshein, and C. J. Stone. *Classification and Regression Trees.* Chapman and Hall, 1993.
5. R.A. Cole, A.I. Rudnicky, and V.W. Zue. Performance of an expert spectrogram reader. *J. Acoust. Soc. Amer.*, 65, Supp. 1:S81, Paper presented at the 97th meeting of the ASA, 1979.
6. S. Coste-Marquis. *Utilisation des techniques du raisonnement hypothétique pour la reconnaissance automatique de la parole.* Thèse de L'Université Henri Poincaré, Oct 1994.
7. Y. Laprie and M.O. Berger. A new paradigm for reliable automatic formant tracking. In *Proceedings International Conference on Acoustics, Speech and Signal Processing, Vol. II*, Adelaide, Australia, April 1994.
8. J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1986.