LEXICAL FILTERING BY MEANS OF PROSODIC INFORMATION

F. Béchet (*), P. Langlais (**), H. Méloni (*)
(*) Laboratoire d'Informatique d'Avignon (LIUAPV)
33, rue Louis Pasteur - 84000 Avignon - France
(**) IDIAP
Case Postale 592 CH 1920 - Martigny - Suisse

ABSTRACT

We present in this article a study on the integration of prosodic information in a lexical access module. Our approach consists, in a first step, to verify the pertinence of microprosodic features contained in fundamental frequency, duration and intensity parameters. We have realised, in order to do that, an inventory of relevant features which have been the subject of many studies by the past. Then, we describe their performance (in filtering and/or verification) when they are measured in an automatic way. As a conclusion, we will present an implementation of an efficient filter using the fundamental frequency parameter.

INTRODUCTION

Many studies point out the essential role of prosody in the linguistic code and mainly in emotional, pragmatic, semantic and syntactic areas. That does not mean however that prosody is not involved into lower levels. Di Cristo [1] has shown the importance of microprosodic information in French already underlined for other languages by earlier studies. Even if our language has no functional stress pattern at word level as it is the case in English, studying the integration of some prosodic features - such as duration, intensity and fundamental frequency variation - in a task of lexical filtering appears fully justified to us.

This study takes place in the general framework of word recognition processes developed in LIUAPV. So as to determine the pertinence and robustness of the prosodic features studied, we plan to add a prosodic lexical filter to the lexical access module of the SPEX system [2]. We deal here with problems posed by the recognition of isolated words from very large vocabularies. All modules developed in this project have a "knowledge based" approach. One of the main purposes is to propose an alternative to systems using statistical methods with a large training corpus. In our system the training needed for every new speaker is reduced to the realisation of 30 words chosen for containing all the French phonemes in various contexts introducing little distortion.

As we already have a lexical access module made of filters progressively reducing the dictionary of possible words, the questions we want to answer, at the beginning of this study, are :

- Is taking into account additional prosodic information likely to improve the process described ?
- In the affirmative, what is accurately the relevant information ?
- Is the latter robust enough to allow its efficient integration into our lexical module ?

We will first briefly present the lexical access module and the prosodic features studied.

THE SPEX SYSTEM

The SPEX system operates on two levels:

- a set of phonetic modules including an Acoustic Phonetic Decoding process (APD) [3] producing from speech signal and speaker references a lattice of valuated phonetic hypotheses
- a lexical access module which filters the global lexicon to propose a subset of candidates to the evaluation process.

The goal of a lexical access module is to find a correspondence function which links the phonetic units recognised with the lexicon. The first operation is choosing the kind of unit suitable to make the link with the lexical items. The phenomena of insertion, deletion and substitution which appear in the phonetic lattice lead us to think that the phoneme unit is not a realistic choice because of the insufficient performance of the bottom-up Acoustic Phonetic Decoding system. In fact there are too many uncertainties to allow a direct access to some parts of the lexicon.

Therefore the choice has been made to use macro-sets representing sounds which have distinctive acoustic features. The advantage of such a representation is to put a structure on the information contained in the phonetic lattice. The number of possible paths inside the lattice for identifying a word is then reduced.

Our lexical access method consists in representing the lexical data and the phonetic hypotheses in a common structure. This structure will allow us to select, in a bottom-up way, some lexicon items. By making this structure more accurate till the phonetic description of each item is complete, we progressively reduce the number of lexical hypotheses in order to give a cohort of valuated items as a probable solution.

The last step in the recognition process consists in sharply evaluating the hypotheses left. To this purpose, by means of spectral distances and contextual articulatory features, we confront the calculated phonetic decomposition of each item of the cohort remaining with the effective realisation of the sounds by the speaker.

PROSODIC FEATURES STUDIED

All the prosodic features chosen have been studied on large test corpora (from 500 to 1000 words pronounced by several speakers).

Duration

Thanks to numerous studies concerning the temporal aspect of speech, we know that the acoustic symptoms of this parameter are governed by multiple factors. Consequently, we are immediately confronted with two difficulties when we want to use them in an automatic process :

- How can we segment a speech signal into discrete units (in our case phonemes) ?
- What precision can we expect from our measures?

We have to study the variations of vowel duration by taking the duration measures produced by the lexical module. A precise study of this parameter can be found in [4]. Here is a summary of this study.

Among all the factors which influence intrinsic vowel duration, it seems that only a few can be observed - at least by our techniques - beyond average values. One can however retain that :

- an oral/nasal distinction can be made, at least partially, with a low error rate,
- the position of the vowel strongly influences its length, but the phenomenon is in no way easily localizable at the end of the word,
- the influence of the consonant to the right is not easily measured,

As a conclusion, it seems that intrinsic vowel duration is not reliable enough to be used in our system (except oral/nasal vowel distinction) because of its lack of robustness - largely due to its bad automatic extraction. The total output of these features in a top-bottom utilisation is rather weak. Therefore it does not seem judicious to integrate them for the moment.

Variation of fundamental frequency

It is often argued that fundamental frequency can be used with great benefit in a segmentation process. In spite of this opinion, it is rarely used in recognition systems for several reasons, the principal one being the lack of reliability of *f0* measures.

All intrinsic and co-intrinsic variation factors of vowel fundamental frequency can be reduced to the articulatory model of the vowel and to the voice/voiceless characteristic of the previous consonant.

About consonants, the shape of the *f0* curve can provide indications for the distinction between obstruent and non obstruent consonants. We have integrated in our lexical access module a filter based on a Bayesian decision of the obstruent/non obstruent nature of inter-vocalic consonants. This decision, calculated from the distributions of fundamental frequency measured on our corpora, allow us to filter about 15% of our word cohorts with an average gain of two position (when the word pronounced is not classified first). However the rejection rate is about 7%.

Intensity

Intensity is without doubt the least studied parameter of prosodic research, although it is by far the easiest to extract from the speech signal. The few studies on this parameter [5], however, point out the following result : the global intensity of a vowel generally seems weaker when it is preceded by a voiceless consonants ; low vowels generally have a higher specific intensity than high vowels (with a minimum for vowel /i/ and a maximum for vowel /a/ and /ɔ/).

We have thus developed a filter based on a decision made from the distribution of initial vowels /i/ and /a/ in our corpora. The error rate is low, but the filtering rate is not very efficient.

Voice/voiceless discrimination

The fundamental frequency parameter allows us to distinguish voiced consonants from others. We know, however, that a voice/voiceless distinction from the speech signal is difficult to obtain in all conditions. We measure this parameter with an algorithm based on the Amdf method, which gives good results. The voice/voiceless decision is taken according to the shape of the Amdf curve calculated on every signal frame. The results obtained on our test corpora allow us to consider efficient the use of this parameter for our lexical filtering task.

The lack of robustness of most of the microprosodic features examined leads us to implement, at first, a lexical filter using the voice/voiceless decision curve.

IMPLEMENTATION

We break up the problem in two stages :

- A first filter works before the recognition process in a bottom-up way. To this end it uses the phonetic chain associated to each word of the lexicon. This chain represents the "usual" pronunciation of these words. We have now to eliminate the candidates whose "theoretical" voice pattern does not match with those measured on the speech signal.
- The second step consists in filtering in a top-bottom way the resulting cohort produced by the lexical process by eliminating the words whose calculated voice pattern is not compatible with those of the speech signal. At this step we have a number of hypotheses about the phonetic chain and its temporal position.

Filter 1

The voice pattern of the signal is obtained from the voice curve calculated with the fundamental frequency variation curve. A "theoretical" voice pattern for every item of the lexicon is made according to the following technique :

- Very few words include a sequence of three consonants. We do not take into account their possible consonantal assimilations.
- When we have a sequence of two consonants, the voiceless consonantal assimilations are ignored. The temporal alignment is not yet known, so these phenomena cannot affect the voice pattern of the word.
- We then consider the possible voice consonantal assimilations which can affect the voice pattern.
- Finally we take into account the fading of the /ə/ at the end of a word.

We have tested the efficiency of our filter on the corpus AviLex (700 words pronounced by six speakers) [2] previously used for testing the lexical access module, with the same lexicons of 15 000 and 20 000 words. The results of table 1 show a filtering rate of 60% for an error rate of 3%. This filtering rate is good, compared to the simplicity of the techniques used.

Table 1 : result of filter 1

speaker	filtering	errors
fb	60%	2.9%
hm	60%	2.1%
ts	59%	3.4%
pg	59%	1.2%
lc	59%	3.7%
si	59%	2.7%
all	59.3%	2.6%

We wish to specify that these results are obtained with a speaker independent algorithm. An error rate under 1% - for the same filtering rate - can be reached when we determine a voice threshold specific for each speaker. Although we think that it is possible to automatically determine this threshold during the training phase of our system, we did not proceed further in this direction because of the efficiency of the global algorithm.

Filter 2

This filter is used at the end of the recognition process whose output is a valuated cohort of 50 to 150 words containing the pronounced item.

Unlike the first filter, it operates with the phonetic chain supposed and its temporal alignment. Thanks to this information, it is then possible to sharply filter the word cohort by integrating all the consonantal assimilation and /ə/ elision rules.

Table 2 presents the results obtained for all the speakers of the AviLex corpus. We obtain an average filtering rate of 20% on the word cohort with an error rate of about 3%. The errors can be explained by a bad voice decision (30% of the errors) or by a mistake in the phonetic alignment proposed by the lexical level of the system. The average gain - for all the speakers - is about three positions up when the pronounced word is not first in the cohort.

Table 2 : result of filter 2

speaker	filtering	errors	gain
pg	20.9%	2.1%	3.5
fb	17.5%	2.2%	2.7
ts	21%	2.7%	3.7
si	16.9%	2.2%	3
lc	17.8%	1.5%	2
all	18.8%	2.1%	3

CONCLUSION

The study of the integration of microprosodic features in a lexical access module suggests the following comments.

Most of the various filters realised had insufficient robustness. Without denying the important role of the microprosodic features in the discrimination of sounds, it seems that the automatic extraction of these features leads to a loss of precision in their measurements. Because of this situation, we use average values for all the features and the filtering rate obtained is rather disappointing.

Nevertheless the results obtained by our filter using the voice/voiceless detection curve justify the use of robust features in a lexical access process (in a top-bottom or bottom-up way). Even if it is pretentious to talk here of a prosodic treatment, this method has the advantage of presenting good results which can be easily integrated into an automatic speech recognition system.

REFERENCES

[1] A. Di Cristo ; *De la microprosodie à l'intonosyntaxe* ; Université de Provence, 1985.

[2] F. Béchet, *Système de traitement de connaissances phonétiques et lexicales : application à la reconnaissance de mots isolés sur de grands vocabulaires et à la recherche de mots cibles dans un discours continu*, Thèse de l'Université d'Avignon et des Pays de Vaucluse, 1994.

[3] P. Gilles, *Représentation et traitement de connaissances acoustiques et phonétiques pour la reconnaissance de la parole*, Thèse de l'Université d'Avignon et des Pays de Vaucluse, 1993.

[4] P. Langlais, *Promenade légère dans les allées prosodiques du jardin de la parole* ; Thèse de l'Université d'Avignon et des Pays de Vaucluse, 1995.

[5] M. Rossi, *Interaction des glissements d'intensité et des glissements de fréquences*; XIVth International Conference on Acoustics, 1976.