, # KNOWLEDGE BASED LEXICAL FILTERING : THE LEXICAL MODULE OF THE SPEX SYSTEM

F. Béchet, H. Méloni, P. Gilles

Laboratoire d'Informatique de l'Université d'Avignon et des Pays de Vaucluse

ABSTRACT - We present the lexical component of a speech recognition system (SPEX) which connects an acoustic-phonetic decoding process with a large lexical data base. The lexical level is made up of some coding modules of the phonetic and lexical data and of a lexical access stage using a set of various filters to reduce the number of hypotheses.

INTRODUCTION

Within the framework of isolated word analytical recognition on large vocabularies (more than 10 000 words), we present the lexical access techniques developed for the SPEX system [Béchet, 1994]. This module makes the link between the information given by the acoustic-phonetic level - mainly a lattice of phonetic units [Gilles, 1993] - and a lexical database. This paper describes the common representation of the phonetic hypotheses and of the lexical data before presenting the various filters which use them. We will also describe the test protocol which has allowed us to evaluate all our modules.

METHODS

The goal of a lexical access module in an Automatic Speech Recognition System (ASR) 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 (APD). 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 identifiying 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 phonological 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.

INFORMATION CODING

Acoustic-phonetic information coding

The information available at this step of the system are :

- energy parameters on the speech signal,
- a measure of the spectral distance between each frame of the signal and the various phonetic references corresponding to a speaker ,
- a lattice of phonetic hypotheses [Gilles, 93]

The suppression of consonants and the large number of insertions for each sound pronounced do not allow us to look for the phonetic strings of the lexicon items directly in the lattice. To deal with these problems we use an intermediate representation between the lattice and the phonetic description of each word. This representation is based on a segmentation of the phonetic lattice into various areas called *families* of phonemes. The idea is to bring together all the phonetic units proposed by the APD during the realisation of a sound.

Once the segmentation into families is done, the system calculates various parameters (energy, duration, spectral distances, etc.) in order to label each family according to several symbolic criteria of likelihood. The goal of this labelling operation is to generate a set of hypotheses (called *patterns*) on the vocal/consonantal (V/C) structure of the spoken word. These patterns are obtained by applying different derivation rules on the configuration of the families of the speech signal. Each pattern represents a possible interpretation of the phonetic lattice according to the structure of the spoken word. These V/C patterns are the intermediate steps which make the link between the phonetic lattice and the lexicon. We will describe now the coding of the phonological data before explaining the access processes which make this link.

Phonological data coding

The lexicon used by the system is made of a text file where each line describes a different item by giving its orthographical form followed by its phonological form. The system generates some models for each word of the lexicon. These models are the V/C patterns described above. They are obtained automatically by appliying two kinds of rules : the suppression rules of the consonants by the APD and the phonological rule of the ellision of /ə/ in French. These patterns are all the different forms a word can take in the hypotheses built from a lattice representing the realisation of this word.

The link between the hypothesis deducted from the phonetic lattice and the words of the lexicon is made. To permit optimised access to the data of the lexicon, the latter is coded in a table where the access keys are the different patterns deducted from the phonetical transcriptions.

LEXICAL FILTERING

We now present the lexical filters which have been developed for the recognition of isolated words. There are one bottom-up filter (Fi1) and two top-bottom filters (Fi2, Fi3). To optimize the link between all the data, we put together in the same structure the hypotheses on the V/C patterns of the spoken word and the data stored in the phonetic lattice. This structure is an acyclic graph whose states are the families of phonemes. All the paths linking the families represent the different V/C patterns already calculated. All the lexical processes use this graph instead of the phonetic lattice. The insertion and suppression phenomena are taken into account by calculating the possible paths in the phonetic graph. The lexical access is then improved by a global treatment of these problems.

Filter 1 (Fi1)

The first bottom-up filter uses the information produced by the analysis of the phonetic lattice to have access to a sub-lexicon compatible with the spoken word. This sub-lexicon is directly obtained by the comparison between the patterns generated from the lattice and the key-patterns of the access table of the global lexicon. The lexicon is then reduced to the words which can be represented by one of these patterns. The first process in the lexical access is to cover the phonetic graph with the V/C patterns of the lexicon. For each pattern all the possible paths in the graph are searched. The comparison and evaluation algorithm is now able to work with a limited set of structured hypotheses where each sub-graph obtained represents all the words corresponding to the pattern examined.

Filter 2 (Fi2)

With Fi2 begin the top-bottom filters testing the validity of the lexical information at the acoustic-phonetic level. Filter Fi2 makes a phonetic comparison according to the V/C patterns chosen between the items kept by Fi1 and the families of phonemes. The phonological transcription of the items kept are reduced to the V/C pattern considered, then the presence of every phoneme of the items so reduced is checked in the list of the *n* best phonetic propositions of the corresponding family. Every missing phoneme in the reduced phonetic transcription of a word leads to its suppression from the list of candidates. The sub-lexicon produced contains all the compatible words with, for each of them, the string of phonemes found in the lattice with all their characteristics (beginning, end, most stable frame and value).

This filter is efficient and fast because the structure of the graph permits to overlook the temporal index of each phonetic hypothesis. The filtering rate is high even with reduced phonetic transcriptions.

Filter 3 (Fi3)

Filter Fi3 uses the phonetic strings corresponding to the reduced transcriptions produced by Fi2. Its goal is to order the cohort of candidate words with a phonetical score (called SP) defined by the following formula :

- M : a word described by the phonetic units U_1, U_2, ..., U_n,
- i_k, j_k : the frames of beginning and end of unit U_k,
- $Ph(U_k)$: the phoneme represented by unit U_k,
- $Dis_p(t)$: the distance of the spectrum of frame *t* to the reference spectrum of phoneme p.

The score $SP(U_k)$ of unit U_k and the score SP(M) of word M are calculated as follows:

$$SP(U_k) = \frac{\sum_{t=i_k}^{j_k} Dis_{Ph(U_k)}(t)}{j_k - i_k + 1} \qquad SP(M) = \frac{\sum_{k=1}^{n} S(U_k)}{n}$$

The words are ordered in an array containing the *p* best scores. This constant *p* can be modified according to the application required, it represents the compromise desired between the number of words proposed to the recognition module and the missing rate considered as acceptable.

This filtering stage permits to progressively improve the hypotheses made on the spoken words by going from a vocalic/consonantal representation to a reduced phonetic transcription. The next step consists in complementing these transcriptions by looking for the missing phonemes in order to have a complete evaluation of the hypotheses produced.

ASSESSMENT OF THE LEXICAL HYPOTHESIS

The result of the previous phase is an ordered cohort of likely words. Before giving the final cohort we have to evaluate all the phonemes of the items, even those unrecognized by the APD.

We then complement the reduced phonetic transcriptions of the items by looking for the missing phonemes. Some of these phonemes may have been identified by the APD without being in the families. It is often the case for a slightly closed consonant by a voiceless occlusive, for example the sequence /kl/ of the word *cyclone* will be represented by only one family, even if the /l/ is present in the lattice.

In this example, the algorithm will check the presence of the /l/ in the area comprised between the most stable frame of the phoneme /k/ and that of the phoneme /ɔ/. If the phoneme is in the lattice, it will be added to the partial transcription of the item. If not, a top-bottom process will look for it at the acoustic-phonetic level.

The result of this step is the complete phonetic string of each item kept. All the cohort is then evaluated again with the phonetic score SP. As before, only the *p* best scores are kept.

CONTEXTUAL TOP-BOTTOM VERIFICATION

This module is the last step in the recognition process. Its goal is to reduce the cohort produced by the previous steps by means of the phonetic context supposed by the lexical treatment. This phonetic context is necessary to discriminate the items phonetically close which have been sorted with a good score. Actually only top-bottom contextual processes are efficient to separate items which cannot be distinguished only through spectral distances.

Presentation of the acoustic-phonetic features

The acoustic-phonetic features we use come from the work of Jakobson for the classification of phonemes by mutual opposition based on a more or less direct correlation between the acoustics and the articulatory features of a phoneme [Jakobson, 1963 ; Méloni, 1982 ; Gilles, 1993].

Thus, in our system, the evaluation of a phoneme consists in validating its own features in the phonetic context where it is spoken, by means of a set of binary rules associated with each feature. To evaluate all the rules which define the features, we have verified them with the corpus of tests which will be presented in the next paragraph. We have now a confidence threshold for each rule. These thresholds permit us to define different strategies depending on the uncertainty of the verification.

Top-bottom verification strategy

Once the features are defined, we have to determine a strategy integrating all the rules. The goal of this module is to discriminate the items of the resulting cohort. We do not need to score every hypothesis again but have to filter the cohort by suppressing the items incompatible with our top-bottom rules. Our strategy is based on the two following principles :

- a phoneme is kept if all its features have been confirmed by the top-bottom rules,
- a word is eliminated from the cohort of results if the percentage of its rejected phonemes in relation to its total number of phonemes is bigger than a given threshold *S*.

This percentage *S* permits to adapt the "strictness" of the top-bottom filtering according to the application required. The integration of the confidence threshold to balance the rules used for the phonetic verification also allows us to elaborate various strategies adapted to the reject rate accepted.

EVALUATION OF THE LEXICAL MODULES

To validate our choices, the lexical modules have been implemanted in an isolated word recognition system on large vocabularies called SPEX. This system permits us to obtain a spoken word, to calculate its acoustic and phonetic parameters and to recognize it from a chosen lexicon. There is no high level linguistic modules, the complexity rate of the recognition is equivalent to the size of the lexicon.

Testing conditions

In our tests we use the French lexical database BDLEX [Pérennou, 1992]. Because of the lack of high level linguistic treatment, we have kept in our tests only the orthographic and phonetic fields of each word. We have extracted from the database a lexicon, called SPLEX-1, containing all the terminal words of BDLEX-1 (roughly 21,000 items). The tests were carried out with six cooperative speakers (6 males, 1 female). The corpus of tests (C1) contains 700 items from SPLEX-1.

Results of the tests

- Table 1 presents the filtering rates of all the filters and the percentage of errors introduced by each of them. Filters 1, 2 and 3 are filters Fi1, Fi2 and Fi3 presented in a previous paragraph. Filter 4 corresponds to the evaluation phase of the items after the research of the missing phonemes in the lattice. Filter 5 represents the selection of the *p* best hypotheses after the last evaluation of the items.

	% Resulting lexicon	% Errors produced
Filter 1	50 %	1,3 %
Filter 2	24 %	0,75 %
Filter 3	6,5 %	0,3 %
Filter 4	5,8 %	0,3 %
Filter 5	1 %	2,3 %

Table 1 : filtering rate and errors produced at each step of the lexical access process.

- Figure 1 presents the relation between the recognition rate of the SPEX system and the size of the cohorts considered.

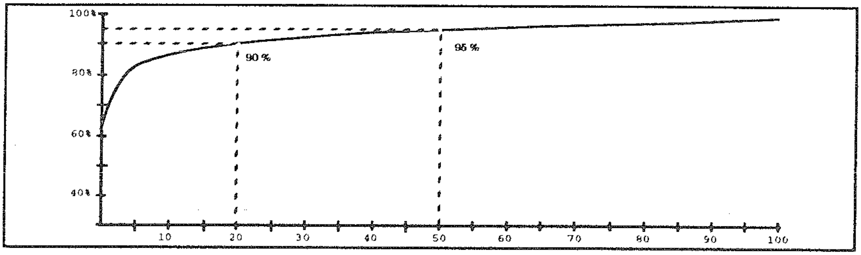

Figure 1 : relation between the recognition rate and the size of the cohorts considered for the corpus C1 on the lexicon SPLEX-1 (21,000 items). The abscissa axis represents the number of elements of the cohort. The ordinate axis corresponds to the recognition rate obtained in the cohort considered.

The results of the filtering phase shows the performance of each filter. The first bottom-up filter is able to filter between 40% and 50% of the global lexicon. This average rate, rather inefficient, largely depends on the word considered. In fact the distribution of the lexicon's items in the access table is not well balanced. This points out the difficulties of the recognition process : on average half the items of the lexicon are compatible with the realisation of a word. The error rate after this first filter is roughly 1.5%. These errors correspond to a bad segmentation of the lattice into families. There are two main reasons : the suppression of a phoneme in the lattice and the presence of an unexpected phenomenon (for example, the realisation of a word with a phonetic form different from those coded in the lexicon).

Filter 2 keeps only 20%, on average, of the items. The error rate of 0.7% is due to the elision of phonemes by the APD.

Filters 3, 4 and 5 are based on a more and more accurate evaluation of the hypothesis kept. The filtering rate is directly linked to the number *p* of items kept at each step of the recognition. In the tests presented, the final cohort has been limited to 100 items and the constant *p* is equal to 12 for filter 3 and 16 for filter 4. By increasing this constant's value and the final number of hypotheses proposed, the recognition rate goes up to the limit value represented by the error rate of filters 1 and 2, i.e. roughly 98% in our tests.

The recognition results show the necessity of including high level linguistic modules in our system to reduce the complexity of the recognition. In fact a cohort of 50 items is necessary to obtain an acceptable recognition rate (95%).

The other tests carried out point that a reduced number of the possible words permits to greatly bring down the size of the cohort to consider (for example, with a lexicon of 500 items, a cohort of 8 words permits to obtain 95% of good recognition).

CONCLUSION

The work presented here is about the conception of lexical modules allowing to make a link between the phonetic decoding processes developed at the LIA and a lexical database. All the modules have an analytical approach using phonetic and lexical knowledge.

One of the main purposes of this study 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 34 words chosen for containing all the French phonemes in various contexts introducing little distortion.

The complexity factors of recognition, pointed out by all the tests performed, and the results obtained with large lexicons shows the limits of our modules and leads us to integrate them into systems with other levels of knowledge.

We now have to think up other algorithms in order to improve our contextual rejection module of the hypotheses proposed in the final cohort. At present we are studying how to improve the phonetic rules of the top-bottom APD and how to integrate prosodic and phonologic knowledge. The binary evaluation of the phonetic features used in the top-bottom APD could be replaced by a continuous evaluation of the quality of the phonemes, meaning a definite improvement.

REFERENCES

Béchet F., (1994), "Un 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.*

Gilles P., (1993), "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.*

Jacobson R., (1963), *"Essai de linguistique générale"*, Édition de Minuit.

Méloni H., Gilles P. (1991), "Décodage acoustico-phonétique ascendant", *Revue Traitement du Signal,* Vol. 8, n° 2, pp. 107-114.

Méloni H., Gilles P., Béchet F. (1992), "Reconnaissance analytique de mots isolés d'un grand lexique", *19° Journées d'Étude sur la Parole*, Bruxelles.

Pérennou G., M. de Calmès, I. Ferrané, J.M. Pécatte, (1992) *Le projet BDLEX de base de données lexicales du français écrit et parlé*, actes du Séminaire Lexique, IRIT-UPS Toulouse, pp. 41-56.