THE APPLICATION OF SPEECH SYNTHESIS AND SPEECH RECOGNITION TECHNIQUES IN DIALECTAL STUDIES

Maria-Pilar Perea
Universitat de Barcelona

Abstract

Speech analysis techniques open new perspectives in the processing of dialectal oral data. Speech synthesis can be useful to create or recreate voices of speakers for extinct languages, to re-edit dialectal material using new technologies or to reconstruct utterances of informants that only were registered in notebooks. Speech recognition, applied to sound dialectal sequences, can make easier automatic transcription of oral texts. In this paper the possibilities of speech analysis techniques in their application to the dialectal studies is described. The presentation is illustrated with the results obtained in different projects.

Key words: *Dialectology, linguistic variation, text-to speech systems, speech recognition, automatic mapping.*

1. Introduction

Given that contemporary dialectology draws heavily on statistical and data processing resources as well as the methods applied in corpus linguistics in the edition and interpretation of its data, it is hardly surprising then that when working with oral texts it should want to take advantage of speech analysis techniques —after all, voice is the raw material obtained by the dialectology researcher from her informant's responses—. Speech analysis techniques —both of synthesis and recognition— are evolving rapidly and are being put to use in many areas of everyday life. Speech synthesis is being used in programs where oral communication is the only means by which information can be received, while speech recognition is facilitating communication between humans and computers, whereby the acoustic voice signals changes in the sequence of words making up a written text.

Speech analysis techniques are opening up new avenues of research in the processing of oral dialectal data. Speech synthesis is useful for creating or recreating the voices of speakers of extinct languages, for re-editing dialectal material using new technologies and reconstructing utterances of informants that have only been recorded in notebooks. Speech recognition, when applied to sound dialectal sequences, can facilitate the automatic transcription of oral texts, so that, first, a phonetic representation can be obtained, and, then, an orthographic representation.

In this paper the possible applications of speech analysis techniques in dialectal studies are described. The discussion is illustrated with results from various finished and on-going research projects.

2. Text-to-speech system (TTS)

In present-day dialectology, notebooks have become archaeological artefacts, replaced by digital and video recording systems. These modern techniques of data collection allow data to be preserved and reproduced as often as required in order to obtain the most faithful transcriptions possible. Earlier data, collected in notebooks, are wonderful dialectal treasures, but lack sound. The development of voice synthesis techniques, which have many useful applications including audiobook production for the disabled, can now provide sound for these dialectal data.

Speech synthesis (Keller 1994) is the process of converting written text into machine-generated synthetic speech. In general, there are three approaches concerning text-to-speech (TTS) systems: *a*) *formant:* this employs a set of rules to synthesise speech using formants, which are the resonance frequencies of the vocal tract; because it does not use human voice the result is a robotic voice; *b*) *concatenative*: this is based on the idea of concatenating pre-recorded human speech units in order to construct the utterance; and *c*) *articulatory*: this tries to model, analogically or digitally, the human articulatory system, i.e. the vocal cords, the vocal tract, etc.

To date, dialectologists have not used speech synthesis resources. In this paper we explain our attempts to put sound to a body of dialectal data for which, due to their age, we have only written documentary records. The dataset is "La flexió verbal in els dialectes catalans" (from now on "Verbal inflection"), by A. M. Alcover and F. de B. Moll, a corpus of almost half a million forms that describes the complete conjugation of 80 verbs from 149 towns of the Catalan linguistic domain. The data were gathered between 1906 and 1928 and published between 1929 and 1932. In 1999 the computerisation process of the materials started: first, a database was created and, later on, methodologies of automatic mapping were applied. Recently, the data were associated with the voice (male or female voices, since the informants belonged to both sexes).

Speech synthesis was possible because the materials do not only include the orthographic answers, but also the corresponding phonetic transcriptions. Without this sort information, the application of these techniques had been impossible.

Next, we explain the steps given in order to obtain the corresponding sounds of the numerous registers of "Verbal inflexion" by means of the speech synthesis techniques, the difficulties surpassed, and the types of syllabic forms that have permitted to obtain a better quality of speech (cf. Perea 2008). The result has been applied to the existing maps, and a sound atlas of Catalan verb morphology with data of the beginning of the 20th century has been obtained.

2.1. The original database and the new computer treatment

In the original edition of *La flexió verbal* ("Verbal inflection"), the sixty-seven verbs studied were classified by conjugation. Different verb tenses (infinitive, gerund, participle, present indicative, past indicative, etc.) of each verb were shown. To

develop the verb paradigm, each verb form was related to a morphological variant, which helps to determine its dialectal scope, and to a phonetic form. Alongside the phonetic variant there was a list of the localities in which this answer was recorded. The localities were represented by numbers (Figure 1).

96 *Segona conjugació*

18. — CREURE

INFINITIU

Creure: krę̟urə 1-31, 34-37, 39, 41-55, 57-58, 60-62, 132, 137-138, 140, 142. krę̟urę 30, 32-33, 36, 38, 40, 58-59, 87, 137. krę̟urę 56. krę̟urí 56, 75, 86, 89. krę̟urę 63-86, 88-107, 109, 111-117. krę̟urər 108, 110. krə́urə 118-121, 123, 126-131, 133-136, 139-141, 143-147. krəurę 119, 122. krə́urǭ 121, 125. krę̟urō 124. krę̟ura 132. krę̟urę 138-139, 141. krę̟ura 148.

PARTICIPI PASSAT

Cregut: -út 1-3, 5-148. -ę̟t 4. — **Cres:** krę̟s 84.

PARTICIPI PRESENT

Creyent: krəɹę̟n 1-2, 5, 11-13, 15-17, 22-24, 26-29, 33-41, 44-57, 59-62. krię̟n 6. krę̟ɹę̟n 63-67, 71-73, 75-76, 78-80, 82-87, 89, 91, 94. krəɹę̟nt 133, 141. krəɹę̟nt 148. — **Creguent:** krəgę̟n 1, 3-5, 8-12, 14, 16-19, 21-23, 25, 29-32, 39, 41-45, 48-49, 53, 57, 59, 142-147. krę̟gę̟n 68, 70, 73, 75, 81, 85, 94. krę̟gę̟nt 101-107, 111. krəgę̟nt 118, 120, 122, 124-125, 128-131. krəgę̟nt 119-121, 123, 125-127, 132-141. — **Creuren:** krę̟urən 7, 18. — **Crevent:** krəbę̟n 19-22, 27, 30-32, 40. — **Creent:** krəę̟n 58. krę̟ę̟n 88, 90, 92-93, 95-100, 109, 112-113, 115. krę̟ę̟nt 101, 103, 105, 107-108, 110, 114, 116-117. krəę̟nt 133, 135-136, 138, 141. — **Cregüent:** krę̟gu̟ę̟n 77.

PRESENT D'INDICATIU

1.a sg. — **Crech:** krę̟k 1-2, 5-14, 16-62, 87, 124, 132, 137-142. krę̟k 63-86, 88-117, 148. krə́k 118, 128-131. krək 119-123, 125-127, 133-136, 141, 143-147. — **Creui:** krę̟u̟i 3, 5, 15-16. — **Cresi:** krę̟zi 4.

2.a sg. — **Creues:** krę̟u̟əs 1, 9, 11-12, 16-17, 19, 21. — **Creus:** krę̟u̟s 2-3, 5-8, 10, 13-62, 87, 124, 132, 137-142. krę̟u̟s 63-86, 88-117, 148. krə́u̟s 118-123, 125-131, 133-136, 141, 143-147. — **Creses:** krę̟zəs 4.

3.a sg. — **Creu:** krę̟u̟ 1-62, 87, 124, 132, 137-142. krę̟u̟ 63-86, 88-117, 148. krə́u̟ 118-123, 125-131, 133-136, 141, 143-147.

1.a pl. — **Creyem:** krəɹę̟m 1-3, 5-8, 11-15, 22-24, 26-29, 31, 33-39, 41-42, 44-62, krəɹę̟m 55. krę̟ɹę̟m 63-67, 69, 71-76, 78-80, 82-86, 89, 91, 94. krę̟ɹę̟m 87. krəɹę̟m 148. — **Cresem:** krəzę̟m 4. — **Creguem:** krəgę̟m 9-10, 14, 16-19, 21, 30, 40, 43. krę̟gę̟m 68, 70, 73, 81, 85, 101-106, 108, 111. — **Crevem:** krəbę̟m 16-17, 19-23, 25, 27, 30-32, 36, 40, 51. — **Cresevem:** krəzəbę̟m 17, 25. — **Creuem:** krę̟u̟ę̟m 77. — **Creem:** krę̟ę̟m 88, 90, 92-93, 95-100, 104-105, 107-110, 112-117. — **Creym:** krə́ɹ̟m 118-123, 125-131, 133-136, 141, 143-147. krę̟ɹ̟m 124, 132, 137-142.

2.a pl. — **Creyeu:** -ę̟u̟ 1-3, 5-8, 11-15, 22-24, 26-29, 31, 33-39, 41-42, 44-62, 87, -ę̟u̟ 55, 63-67, 69, 71-76, 78-80, 82-86, 89, 91, 94, 148. — **Creseu:** -ę̟u̟ 4. — **Cregueu:** -ę̟u̟ 9-10, 14, 16-19, 21, 30, 40, 43. -ę̟u̟ 68, 70, 73, 81, 85, 101-106, 108, 111. — **Creveu:** -ę̟u̟ 16-17, 19-23, 25, 27, 30-32, 36, 40, 51. — **Cresevęu:** -ę̟u̟ 17, 25. — **Creueu:**

Figure 1

These data, along with the unpublished material included in the original notebooks, were introduced in a database. Once systematised and completed, they formed a suitable corpus for the creation of a computerised linguistic morphological atlas (Figure 2).

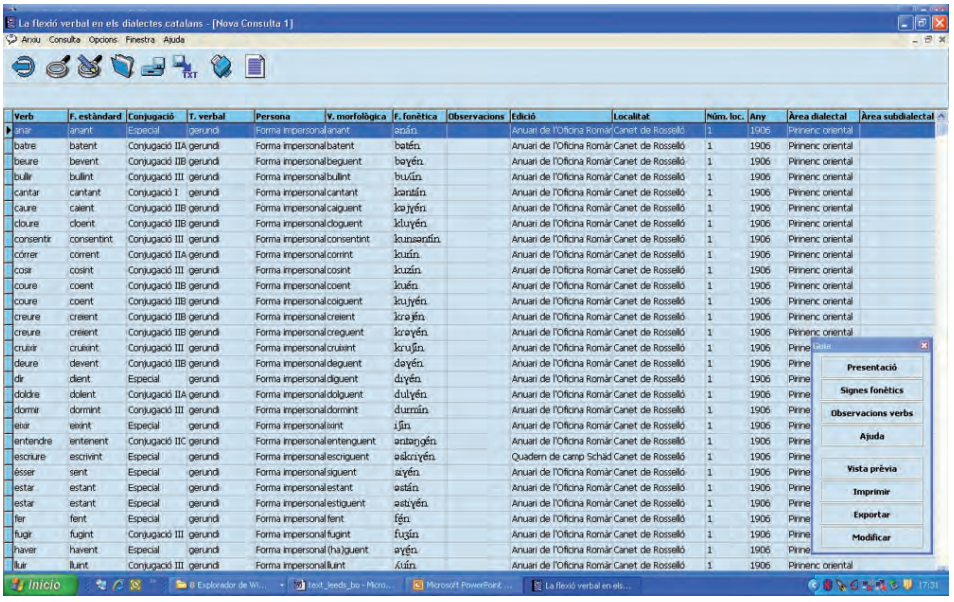

Figure 2

The process of ordering and completing the materials generated a morphological corpus of 470,255 entries (adapted to the IPA alphabet) —cf. Perea (2004)—. In fact, to be synthesised, from the total amount of entries, 1,080 were removed from the verbal database, because, in spite of being included in the notebooks, were forms impossible to be pronounced. They did not adapt to the rules of Catalan pronunciation (i. e. [krúʃɣi], it has to be [krúʒɣi]). Probably, they were written wrongly during the process of data gathering or data transcription. The used registers correspond to 28,161 different single verb forms.

The linguistic atlas (Perea 2005) created by the computer program is a collection of 6,000 potential maps that can be updated as the user wishes. Each map places the phonetically transcribed dates in the various localities surveyed, represented by points. The result (Figure 3) is a linguistic atlas that accomplishes three goals: *a*) it presents a synchronic, morphological and phonetic description; *b*) it shows the formation of different linguistic areas through the distribution of coinciding forms; and *c*) it provides representative material for subsequent study or interpretation of the data.

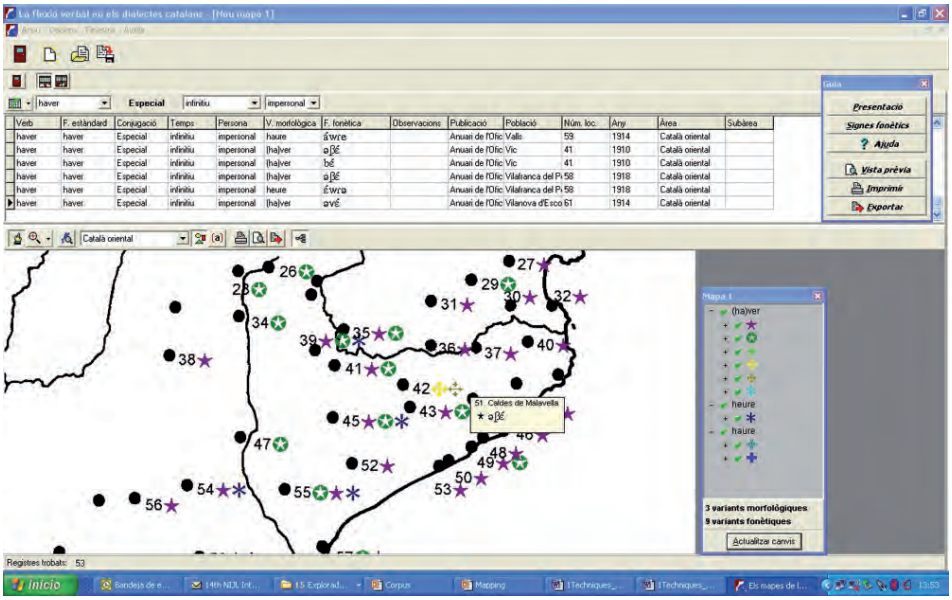

Figure 3

2.2. The stages of speech synthesis

As we have said before, there are three approaches concerning text-to-speech (TTS) systems (*formant, concatenative*, and *articulatory*). Here, we used a concatenative speech synthesis system (Lewis & Tatham 1999) because it gave more natural-sounding output. The system uses a corpus that stores a large inventory of units to be concatenated. To obtain pre-recorded speech units it is necessary to record human voices.

The stages in the process of creating the speech synthesis system were as follows:

b.1) recording human speech;
b.2) developing a speech synthesis system;
b.3) applying the concatenation process;
b.4) editing sound and solving problems.

2.2.1. Recording human speech

The speech corpus was created by recording different real and unreal words pronounced by a number of speakers (male and female). The words contained all the syllables required by "Verbal inflexion" (3,526). Because the original data represent all the Catalan varieties (Eastern Catalan, Western Catalan, Valencian, *Alguerés*, *Rossellonés*, and Balearic), we surveyed different speakers of the four main dialects (Eastern Catalan, Western Catalan, Valencian, and Balearic), especially as regards their idiosyncratic pronunciation. In some cases, and in order to obtain specific sounds, we also surveyed speakers from small towns with particular vowel or consonant pro-

nunciation; for example, speakers from Felanitx or Santa Coloma de Queralt, concerning vowel endings, or from Benavarri, relating to the production of the consonant group [pʎ + vowel].

The survey covered about 1,000 words (monosyllables and polysyllables) in order to gather stressed and unstressed syllabic samples. Each informant pronounced only the part related with his/her own dialect. However, the questionnaire had a main section that coincided with the general pronunciation of Catalan read by two speakers (male and female).

Figure 4 shows a sample of the questionnaire eliciting words that contain syllables with [ɲ] and [ʎ]. On the right we can see the word read; on the left, the syllable obtained.

ɲút	bəɲút
ɲin	báɲin
ɲis	báɲis
ɲi	báɲi
ɲu	báɲu
ɲun	báɲun
ɲus	báɲus
ɲut	báɲut
ɲuk	báɲuk
ɲik	báɲik

ʎən	báʎən
ʎəs	báʎəs
ʎə	káʎə
ʎuk	káʎuk
ʎus	ʎuskám
ʎut	káʎut
ʎui	ʎuiré
ʎúu	
ʎəw	káʎəw
ʎóis	

Figure 4

The speech corpus was recorded at a frequency of 44,100 KHz, using one channel (mono) and 16 bits. The words of the questionnaire were read by the speakers without any specific context in order to obtain a homogeneous pitch (in spite of the differences between speaker voices) and similar frequencies. All the recorded speech that showed emphatic pronunciation, a sort of marked modulation or other anomalies was rejected.

We sought young speakers with similar pitch. This was because at the beginning of the twentieth century, the informants used for Alcover's "Verbal inflection" were young people, both men and women. In the case of female voices the process of homogenising differences was easier; in contrast, the male voices recorded showed a wide variety of low and high pitches.

We used the tools of the free software WavePad in order to homogenise different pitch voices, change the volume of the recordings, improve its quality, and reduce noise, etc. (see Figure 5).

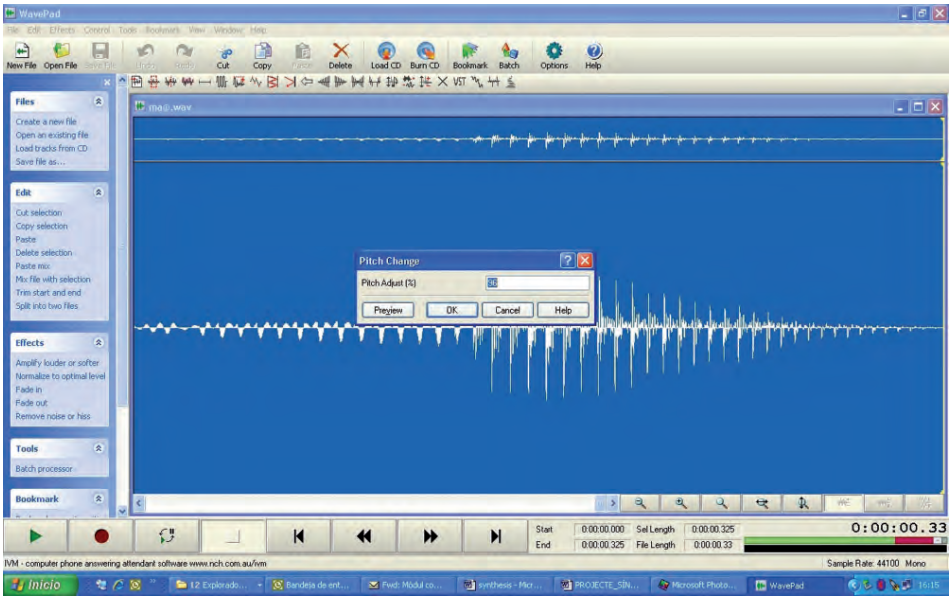

Figure 5

Several different variables may affect the recording of read text, and these need to be taken into account. For instance, the use of different microphones can produce unexpected results related to the pitch.

2.2.2. Developing a speech synthesis system

The speech synthesis system is based on the concatenation of sound units. In our system the syllable was chosen as the main unit for generating synthesised voice. Sounds for which syllables present some problems were used as supplementary units.

We used the syllable as the minimal sound unit, because, after different verifications, we observed that this yielded the best results as regards speech fluidity. Using phonemes as units also requires an intermediate basic unit called a "diphone" (based on the concatenation of each pair of recorded phonemes). In our system, and after applying certain algorisms, use of the syllable was observed to yield —with some exceptions— transitional results that were generally natural enough. Moreover, the syllable is a controlled unit. In Catalan there are about 5,000 different syllables combining consonants and vowels. As regards our corpus, the database of "Verbal inflection" required 3,526 syllables.

Catalan has a lot of monosyllabic words. Concerning stressed syllables most of the words read were real monosyllabic words. The syllable inventory created was stored in a speech database (DBISam database format). In this database each syllabic unit was labelled according to its phonetic representation (Figure 6).

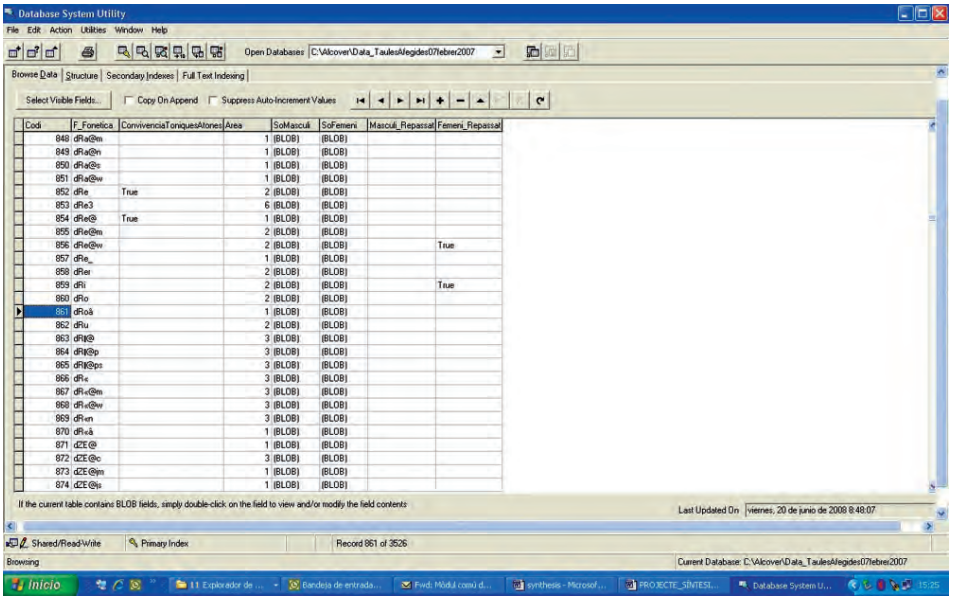

Figure 6

In order to obtain the most natural sounding speech, the corpus contains both stressed and unstressed syllables. The former require a higher pitch and frequency, while the latter require a more muted and constant pitch.

2.2.3. Applying the concatenation process

Firstly, a system was applied to the verbal database so that when we chose a word to be pronounced it automatically divided each word, transcribed phonetically, into syllables according to the specific rules of the Catalan language and taking into account diphthongs, triphthongs and hiatus (Figure 7 shows an example with *acudir* 'to attend').

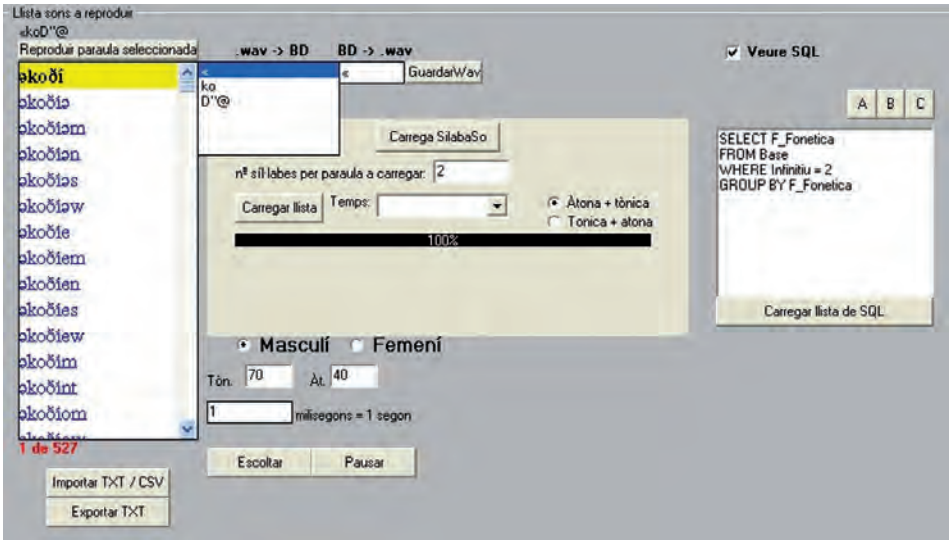

Figure 7

After the syllabic division, the corresponding syllabic sound from the speech database was associated with the phonetic transcription. This sound, comprising a series of numbers corresponding to its wave signal, was stored in a temporary file. Next, the sound of the following syllables (if the word is polysyllabic) was also associated. When a blank space showed the end of the word the program concatenated the different temporary files and the sound of the chosen word was obtained (Figure 8).

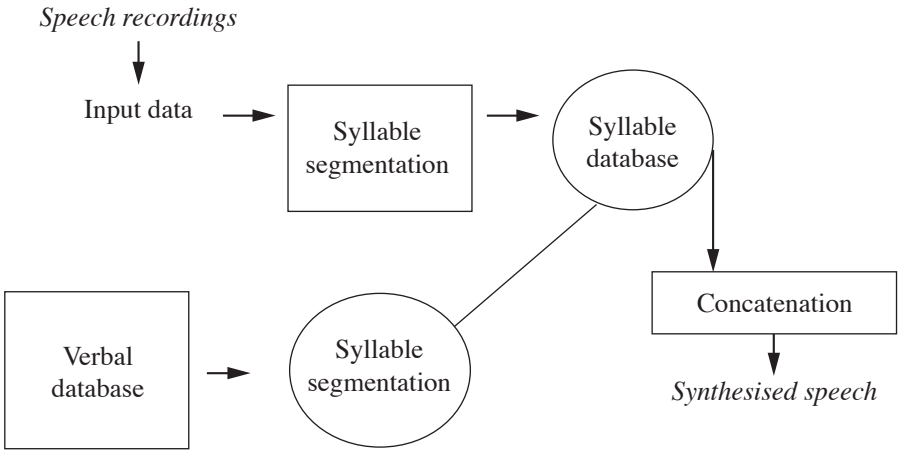

Figure 8

In the process of syllable concatenation we had to apply different algorisms:

c.1) If the first sound of the syllable was a voiceless plosive ([c],[1] [k], [p] or [t]) or if it fell in the middle of the word, the program added a silence of 20 milliseconds (ms) at the beginning, because these sounds are produced in a very short time. The silence helps to distinguish them clearly.

c.2) If the last sound of the last syllable of the word was not a voiceless plosive ([c], [k], [p] or [t]), an effect of reducing the sound progressively was applied to the second half of the syllable.

c.3) If a word contained a hiatus the transition of the vowels was included in the speech database. In this situation the transition produces a more natural sound. Thus, the process adapted to the word [koém] 'we cook' would be as follows: this word originally has two syllables: [ko] + [ém], but a new sound [oé] will be created, which will be associated to [k] and [m], so: [k] + [oé] + [m]. It is thus a sort of "diphone".

2.2.4. Editing the sound and solving problems

We used the WavePad program to edit the different syllables and, in some cases, to modify the sound qualities or to change a part of a syllable (Figure 9). In this case a complete synthesised sound is produced.

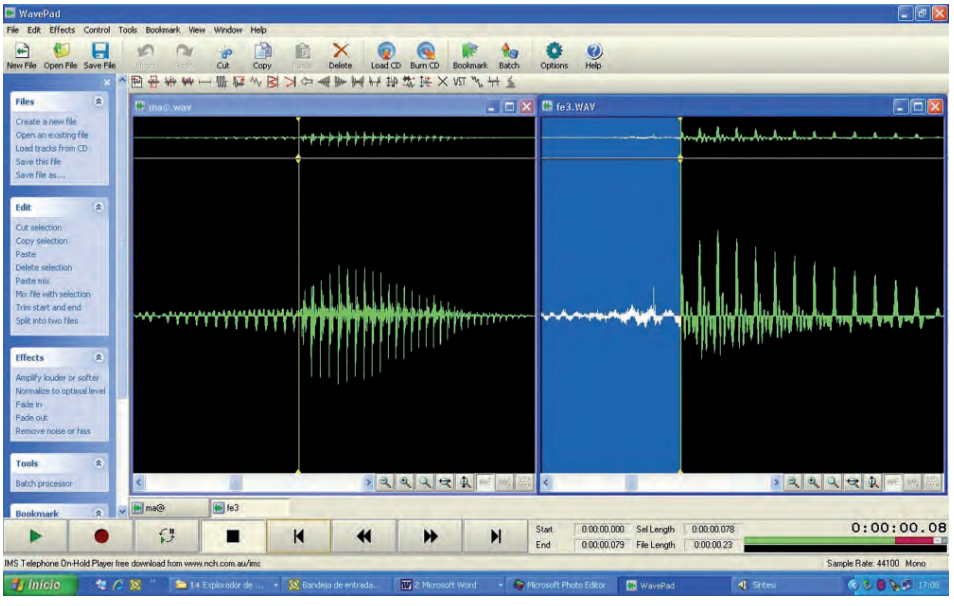

Figure 9

[1] The sound [c] is a velar voiceless plosive palatalised, pronounced in some Majorcan villages. It appears in very predictable contexts.

At times, in the process of concatenation of syllables pronounced by different speakers, we noticed stress and speed differences, even though the pitch was correct. When this happened, we selected the syllables, or a part of them, that presented the best phonetic quality. This process was applied to similar syllables, such as [bíu], [kíu], [díu], [fíu], [gíu], [jíu], [míu], [níu], [ʎíu], [líu], [ɲíu], [ɾíu], [ríu], [síu], [ʃíu], [tíu], [píu]. In this case, [íu] was the most representative part of the syllable and the initial consonant was added to this part.

The most important difficulty in terms of sound concatenation was joining vowels (cf. b.3.c.3) divided into two syllables ([ɾí]+[a] from [kan]+[ta]+[ɾí]+[a] *cantaria* 'I would sing'), because the contact produced a superposition or a sudden change in the pronunciation. To solve this problem we recorded this joining (or transition) independently.

The most conflictive vowels were the stressed [í] and [ú]. The vowel contact with these sounds led to the appearance of "ghost" sounds. Thus, in a form such as *plaïu* [pla-í-u] 'you please', in the coarticulation between [a] and [í], the listener does not hear this form, but rather a non-existent transition consonant: pla[β]íu, pla[ɣ]íu or even pla[ɾ]íu. Here, as happens with other sounds, there is also a perceptual question.

Other problematic sounds to be synthesised were the nasals, [m] and [n] at the beginning of the syllable or [ŋ] and [ɲ] between syllables.

In contrast, the synthesis was easier when the syllables start with the voiceless plosives [c] and [k] or end with [c], [k], [p] and [t]. As regards voiced and voiceless fricatives, in the middle of the word, [ʒ], [s] and [z] there is no problem of synthesis and perception. The same happens when [s] or [ʃ] are at the end of the syllable.

The lateral [ʎ] at the beginning or end of the syllable does not present any problems. However, younger generations tend to lose this sound changing it into [j]. Thus, the chosen speakers had to pronounce this consonant clearly. The sound [l], in the coda position, was sometimes perceived as [w], especially when it was followed by another syllable [val] + [dɾí] + [a] *valdria* 'it would cost'. The sound [w] can be explained by the fact that the lateral alveolar in Catalan has a strong velar secondary articulation.

Rhotic [r] is easier to synthesise than the intervocalic [ɾ].

Other sounds can be synthesised by starting from others. The approximants [ɣ], [β] and [ð] between vowels and between other contexts can be extracted from the second half of the corresponding voice plosives ([g], [b] and [d]).

When the end of a syllable was any strong vowel and the next one was an unstressed [i] or [u] it was necessary to make these weak vowels longer so that they could be clearly perceived. In contrast, when the joining was produced between the strong vowels of the end of a syllable and [j] or [w], the diphthong was distinguished by making the glides shorter and making the pronunciation faster.

A new difficulty has to do with sounds that existed at the beginning of the twentieth century but which are no longer pronounced today, especially among younger generations. This is the case of the sound [á], which can experience changes or a sort of diphthongation in [eá] or [eɛ́]. The diphthongation was pronounced in Son

Servera, a Majorcan village.[2] The transcription of this diphthong in the "Verbal inflection" was [ęǽ]. Because we have the recordings of the two sounds separately ([ę] and [ǽ]) it was possible to create this diphthong joining them.

Finally, because we synthesise only words, prosodic and intonation aspects were avoided. Here the main objective was the intelligibility, quality and naturalness of the utterance of each word pronunciation.

2.3. The new (sound)maps

Application of corpus-based concatenative speech synthesis systems to Alcover's data enabled us to produce a sound atlas, which not only shows the geographical different distribution of Catalan verb morphology but which can also reproduce the phonetic verbal form through a male or female voice (Figure 10).

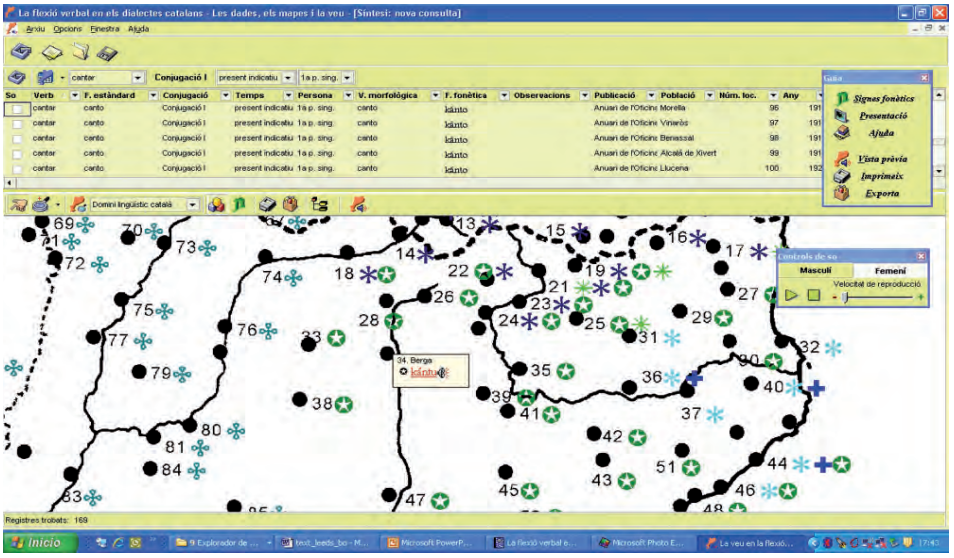

Figure 10

Moreover, the program maps the spatial distribution of the different phonetic and morphologic variants of a verb form, and enables the user to listen to the series of utterances, stopping when desired or repeating a form as often as is required. The speed of the pauses between two utterances can also be changed, making them slower or quicker (Figure 11).

[2] See about this sound, Veny (1983: 100) and Recasens (1991: 91).

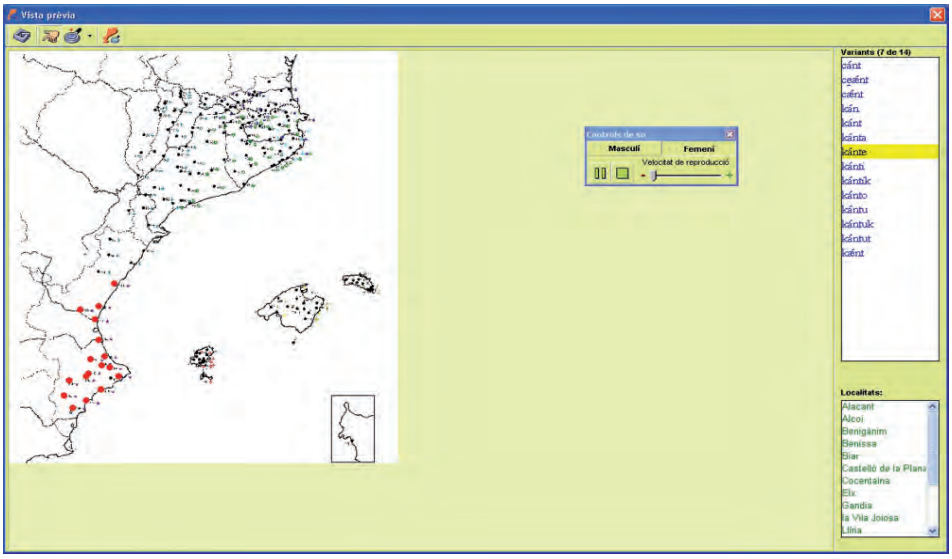

Figure 11

Developing this project has implied a small step in the use of speech synthesis techniques as applied to dialectology. Until now these techniques have been used when voice is the only way of receiving information: for blind people, phone applications (call centres) or in some experimental applications. However, they can also be useful to create voices of speakers for extinct languages, to re-edit dialectal material using new technologies or, as in this case, to recreate the utterances of the informants surveyed by Alcover one hundred years ago. Because at that time it was impossible to record dialectal material, the fact that data are transcribed phonetically helps in using the text-to-speech system.

There are two aspects of the original data which have made the synthesis procedure easier:

a) It is a closed corpus with a limited number of verbal forms — that is to say, it comprises isolated word sequences (in which emotions and prosodic elements have not been considered), and for that reason the speech database was limited. Only in a few cases is there a conjunction of two words in the periphrastic verb forms (*he cantat* 'I have sung' or *vaig cantar* 'I sang'). These utterances were more difficult to produce.

b) Verbal forms are already phonetically transcribed. Thus, we leave out what is called the "front-end" of the text-to-speech system: the input plain text that is transformed in phonetic transcription and the subsequent division into prosodic units.

There is an important element that has sometimes been associated with the intelligibility of an utterance. The rich Catalan verbal variety that was gathered by Alcover at the beginning of the twentieth century is now progressively reduced under the pressure of standard morphology. As regards the subjective evaluation, we

checked that extinct verbal forms or those that are disappearing are more difficult to perceive by speakers of standard Catalan, even though the speech output is of good quality. This means that it is easier to perceive a predictable or known word, as it is more intelligible. In the "Verbal inflection" the dialectal variation presented by a verb such as *obrir* 'to open' is of 919 forms; in contrast, the standard paradigm of this verb, or any other, has only 50.

From a critical point of view, the isolated utterance of a single word in a point on the map can result in some cases that are not much natural; however, the same happens with real voice that has been extracted from a given context, in order to create sound maps.

The knowledge acquired here will enable future research to apply this speech synthesis system to a corpus with similar characteristics. For example, the phonetic transcriptions of the *Diccionari Català-Valencià-Balear*,[3] also by Alcover in collaboration with Francesc de B. Moll (cf. Perea 2004), and which likewise forms a closed corpus, could also be used to generate speech.

3. Speech recognition[4]

Broadly speaking, speech recognition (Zue, Cole & Ward 1997) aims at permitting oral communication between human beings and computers. The challenge is to harmonize information coming from different fields of knowledge (acoustics, phonetics, phonology, lexis, syntax, semantics, and pragmatics) and to obtain, in spite of possible ambiguities and errors, an acceptable interpretation of an acoustic message.

Speech recognition can be applied at different language levels, presented here in increasing order of difficulty:

1. Isolated words.
2. Connected words.
3. Continuous speech.
4. Spontaneous speech.

At each of these levels the speech acoustic signal has to be transformed into the sequence of words from which the written text is configured. In general, the procedure involves the conversion of sound waves into variations in voltage through a microphone device, followed by the sampling of this voltage, so as to obtain, for example, 8,000 samples per second (i.e., each sample is encoded in eight bits; while greater fidelity requires a greater sampling frequency; in our case, 44,100 KHz encoded in 16 bits with which the answers of the informants were recorded). Having obtained the sample by means of the "digital processing of the acoustic signal", the data are analysed in order to extract the prominent information that enables words to be identified. Thus, the intonation of the sentence does not contribute to their identification, but rather it is the frequency of the acoustic signal as a sound is being

[3] http://dcvb.iecat.net/default.asp.

[4] This research is sponsored by the Spanish Ministerio de Educación y Ciencia and the FEDER (research project HUM2007 65531/FILO: "A dialectal oral corpus exploitation: analysis of the linguistic variation and development of computerised applications to automatic transcription" (ECOD))".

emitted that is the prominent information, since it indicates its phonic characteristics (pitch, means of articulation, etc.).

In theory, once the parameters characterizing the sign acoustics have been established, the search can be initiated. But prior to commencement, a dictionary of words has to be generated. The search is then conducted in this dictionary so as to identify the word that most closely resembles the word we wish to identify. However, problems arise when working with continuous speech (without any artificial pauses in the majority of words) and, in particular, when dealing with spontaneous speech, because part of the search process also involves determining the limits of the word, i.e., identifying where words begin and finish. Two types of external knowledge facilitate this search: the *acoustic* and the *language models* (Fosler-Lussier, Byrne & Jurafsky 2005). The former establish the distribution of the acoustic parameters of each phoneme (thus, /b/, for example, is a voiceless bilabial plosive), while language models establish the usual distribution of the words of a specific language (i.e., the sonorous sequences that are most likely to occur). Thus, based on the characteristics of the acoustic signal and the expectations set up by the acoustic and language models respectively, a hypothesis can be formulated regarding what was said.

Automatic speech recognition has many applications in daily life, including machine translation, speech recognition in cars, computers (dictaphones), GPS, S*peech-to-Text* programs (SMS text transcriptions of speech), robotics, etc. Yet, as is the case with voice synthesis, automatic speech recognition systems have not as yet been applied to dialectology.

The aim of one part of the project designed by the *Universitat de Barcelona*, "A dialectal oral corpus exploitation: analysis of the linguistic variation and development of computerised applications to automatic transcription" ((ECOD) [HUM2007 65531/FILO]) is to develop data processing tools that facilitate the automatic transcription —phonetic and subsequently orthographic— of interviews collected in dialectal surveys carried out in the county capitals (or equivalent centres) of the Catalan linguistic domain. We have gathered a total of 258 oral texts ranging in length from between five and ten minutes and recorded in 86 different places.

We are currently seeking to design a tool that transcribes the oral texts of the informants phonetically, regardless of the variety of their Catalan dialect. Subsequently, we wish to develop an application that can convert the phonetic version to the corresponding orthographic transcript by applying rules developed from the generalizations of each dialectal system.

3.1. From "WavSons" to "WavFonemes"

Our speech recognition system has been applied initially to isolated words from eastern Catalonia. Here, as our point of departure comprises oral texts and a phonetic database, we need to segment out data in isolated words. To do this, we developed a sound processing program, "WavSons 1.0", which enables us to segment, process and store a range of syllabic units in files. As with voice synthesis techniques, the unit of analysis is the syllable (cf. Adda-Decker, de Mareüil, Adda & Lamel 2005), although diphones have also to be taken into consideration. Note, however, that we do not work with isolated sounds or complete words.

"WavSons 1.0" parameterises and stores syllables according to their defining acoustic and phonetic properties. The program proposes an initial segmentation of the sequence under analysis, but subsequent modifications have made certain adjustments to this. The empirical base drawn upon is an oral corpus known as the C*orpus Oral Dialectal* (COD), which provides specific syllable samples.

Samples were collected from informants resident in a number of towns in western Catalonia, including Granollers, Terrassa, Mataró, Santa Coloma de Farners, Vic and Berga. Then, the ability of the program to recognise syllables was tested by analysing these samples. The program was found to distinguish plosives and vowels more readily than liquids, nasals and approximants followed by a vowel, and consonant sounds in the same block. As for the elements of analysis, the collocation of syllables comprising a voiceless plosive plus a vowel was given particular emphasis. Next, the sounds of the segmented syllable were associated with their orthographic and phonetic transcription. In this way, we were able to create a syllabic database. Last year, we replaced "WavSons 1.0" with an improved version, "WavFonemes". This program has been used to design a Catalan sound database and subsequently to normalise it using voice patterns. Our voice pattern database has been created by selecting significant groups of sounds obtained from the speech of our informants. These groups correspond to syllables, though they might also be considered to be diphonemes (understood as entities formed from two sounds).

When incorporating these voice patterns to the database we are concerned solely with the characteristics of the sound spectrum. Thus, first it is necessary to transform the voice signal recorded directly with a microphone (time axis) to sound frequencies (frequency axis), since the identifying and invariable information obtained from the vowels and consonants, independent of the informant, is encoded on the frequency axis. A mechanism must be applied, therefore, that can transform the information in the samples collected on the time axis into frequencies. To carry out this transformation, we apply the "Fourier Transform" operation, which works on the principle that each sound signal, even if it is complex, can be split into the sum of its simple periodic functions (sinusoidal and cosinusoidal) of different frequency. In this way we are able to obtain values that determine the importance of the frequencies, and the most significant are those that compose and shape the voice patterns that we wish to create.

3.2. Quality voice patterns

Obtaining quality patterns requires going through two prior stages:

a) Training patterns (pattern design): establishing consistent representations of minimally reliable sound patterns.
b) Comparison of patterns (obtaining speech recognition from the pattern designs): a direct comparison is made between the unknown signal (the one to be recognized) and all the possible patterns acquired in the training stage so as to determine which pattern gives the best fit.

The main difficulty encountered in adopting this method is that of creating a complete and correct speech pattern database. This was particularly true in our case where we worked with isolated words coming from a closed database. An accurate database is clearly not easily obtainable, because of the dimensions involved and the variations in

the signal sound, and given that the spectrum information it contains does not have a recognizable phonetic meaning. This variation in sound can be attributed to:

1. The variation presented by a single informant as she seeks to maintain a constant and consistent pronunciation in her word production.
2. The variation between informants due to length differences in the processing of vowels and consonants, and to differences in accents associated with dialects, etc.
3. The variation in microphones when recording speech in terms of volume, intensity, etc.
4. The variation in the recording settings (general background or specific noises).

3.3. The functions of the "WavFonemes" program

The "WavFonemes" program allows us to:

Segment parts of a wave file and assign to each part its corresponding phonetic syllable. Once segmented, the sequence can be stored in the database together with the file name referring to that segment and the phonetic transcription assigned to it (Figure 12).

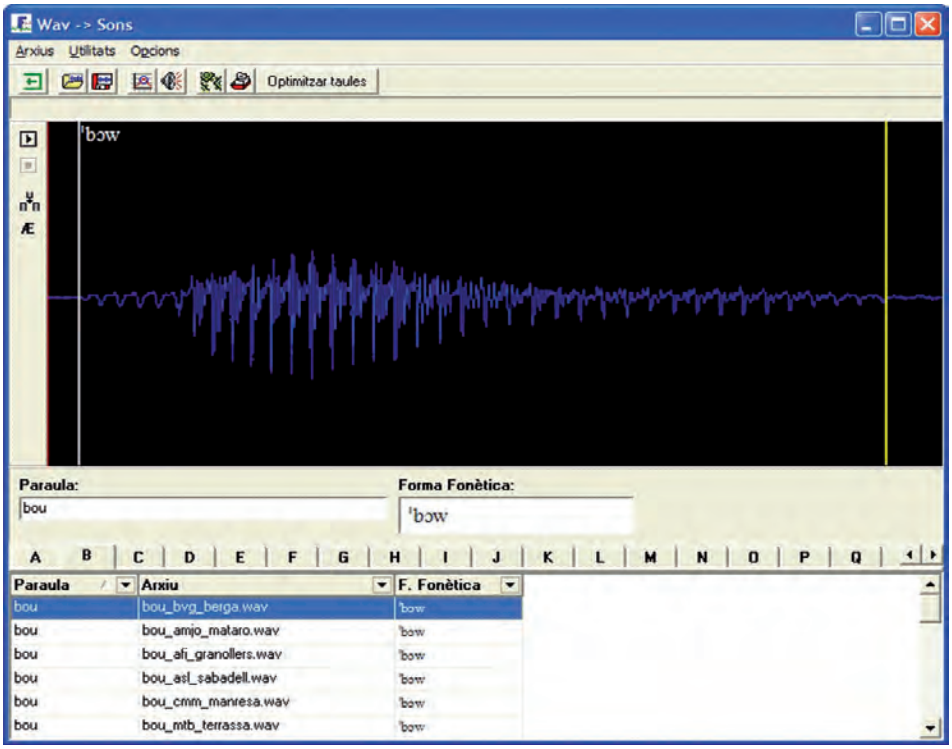

Figure 12

The program has two main tools:

a) Normalising (training stage) the phonetic syllables stored in the database. The aim here is to build a valid phonetic pattern for speech recognition, in which the most representative frequencies of each syllable have a particular importance, independent of the informant. These frequencies are those with the highest values. Various parameters are taken into consideration in obtaining this optimum normalization: first, maximum and minimum frequencies (in general the frequency interval corresponding to the threshold of the human ear, which oscillates between 20 and 10,000 Hz); and second, the average frequency among the higher values, fixed in this case at between 70 and 80%. Note fixing the value at 100% would be incorrect as this would include all sample values, and clearly the values that appear infrequently do not contribute to the construction of a valid pattern (Figure 13). What we seek to do is to generalize and avoid excessive diversification.

b) Recognising (comparison stage) phonetic syllables after they have been normalised. The aim here is to use the wave file to bring up on the screen the di-

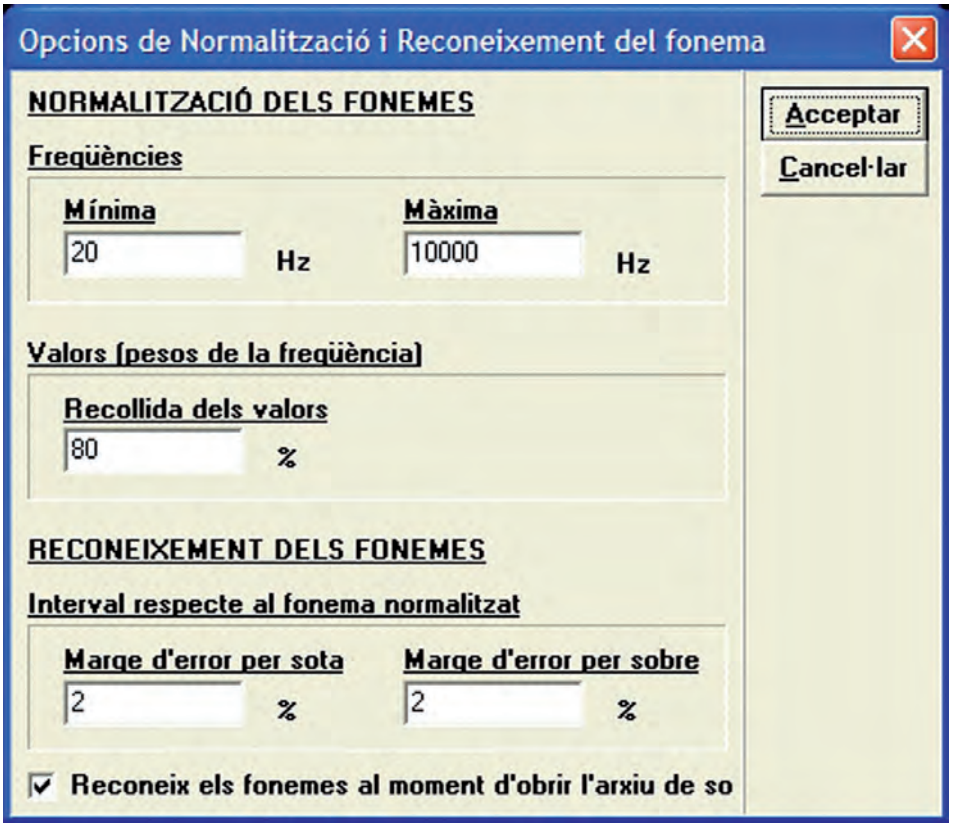

Figure 13

visions in the phonetic syllables and so the corresponding syllable can be identified from the patterns stored in the database.

In the first stage of the voice recognition process, a graph representing the voice signal on the frequency axis is taken as a reference. The samples of the signal are then divided into superimposed sections (windows) of the same size. On each set of samples within each window the "Fourier Transform" formula is applied to detect and group similar frequencies, before each syllable is divided. Once this division is complete, the most typical values are collected, according to the parameters of maximum and minimum frequencies, the average value (the frequency weighting), and the margins of error with respect to the normalized phoneme.

Only the values that coincide are collected. In figure 13, the values have a margin of error of 2%. If these parameters were to be expanded, we would obtain a broader list of candidates, but they would not be as exact. To obtain a progressively more precise technique, the mean values would have to be manipulated. Finally, the frequency values are compared with the previously stored database values and a list of possible candidates drawn up, with the phonetic syllables organised from greatest to smallest degree of accuracy.

3.4. Prospects

a) In the normalisation stage, more statistical parameters might be incorporated in order to condition and perfect the phonetic pattern (median of frequencies, standard deviations of the values that have to be collected, etc.).

b) Other methods of speech recognition might be explored so as to increase the degree of accuracy in the patterns. One such model, capable of providing satisfactory results at the practical level, is the Hidden Markov Model (HMM) (Knill & Young 1997). It should be borne in mind, however, that the model of the pattern can differ depending on how the patterns are obtained and on the particular speech sequence that we wish to recognise. In fact, there is no escaping the marked influence of the informant who recorded the particular speech sequence.

c) The database might be restructured in order to make the processes of normalisation and recognition more agile.

4. Conclusion

Dialectology must take advantage of new computer technologies that can facilitate the processing of dialectical data. To date, the procedures developed in the text-to-speech (TTS) systems and in speech recognition programs have only been used in conjunction with standard language varieties, and so the challenge is now to apply them to the study of linguistic variation.

References

Adda-Decker, M., de Mareüil, P. B., Adda, G. & L. Lamel, 2005, «Investigating syllabic structures and their variation in spontaneous French», *Speech Communication* 46, 2, 119-139. http://dx.doi.org/10.1016/j.specom.2005.03.006.

Fosler-Lussier, E., Byrne, W., & D. Jurafsky (eds.), 2005, Pronunciation Modeling and Lexicon Adaptation. Special Issue. *Speech Communication*, 46, 2.

Jurafsky, D. & J. Martin, 2000, *Speech and Language Processing: an Introduction to Natural Language Processing*, Computational Linguistics and Speech Recognition, Upper Saddle River, New Jersey, Prentice Hall.

Keller, E., 1994, *Fundamentals of Speech Synthesis and Speech Recognition. Basic Concepts, State of the Art and Future Challenges*. Jon Wiley & Sons, Chichester.

Knill, K. & S. Young, 1997, «Hidden Markov Models in Speech and Language Processing», in S. Young & G. Bloothooft (eds.), *Corpus-Based Methods in Language and Speech Processing*. Kluwer Academic Publishers, Dordrecht (Text, Speech and Language Technology, 2), 27-68.

Lewis, E. & M. Tatham, 1999, «Word and syllable concatenation in text-to speech synthesis», in *Proceedings of the European Conference on Speech Communication and Technology* (http://www.cs.bris.ac.uk/Publications/Pa-pers/1000377.pdf).

Perea, M.-P., 2004, «New Techniques and Old Corpora: *La flexió verbal en els dialectes catalans* (Alcover-Moll, 1929-1932). Systematisation and Mapping of a Morphological Corpus», *Dialectologia et Geolinguistica*, 12, 25-45.

—, 2005, *Dades dialectals. Antoni M. Alcover*, Conselleria d'Educació i Cultura. Govern de les Illes Balears, Palma de Mallorca (CD-ROM edition).

— (to be printed), «Retrieving the sound: applying speech synthesis to dialectal data», paper read at METHODS XIII (Thirteenth International conference of Methods in Dialectology), Leeds, July, 2008.

Pols, L., 2001, «Acquiring and implementing phonetic knowledge», in P. Dalsgaard, B. Lindberg & H. Nemmer (eds.), *Eurospeech 2001. Proceedings of the 7th European Conference on Speech Communication and Technology*. September 3-7, 2001, Aalborg, Denmark. Vol 1. pp. K3-K6. In IFA Proceedings (Institute of Phonetic Sciences, University of Amsterdam) 24 (2001): 39-46.

Recasens, D., 1991, *Fonètica descriptiva del català*, Institut d'Estudis Catalans, Barcelona.

Veny, J., 1983, *Els parlars catalans*, Moll, Palma de Mallorca.

Zue, V., Cole, R. & W. Ward, 1997, «Speech Recognition», in R. A. Cole, J. Mariani, H. Uszkoreit, A. Zaenen, & V. Zue (eds.) *Survey of the State of the Art in Human Language Technology*, Cambridge U. P., Cambridge, 4-10.