THE UPC TEXT-TO-SPEECH SYSTEM FOR SPANISH AND CATALAN

Antonio Bonafonte, Ignasi Esquerra, Albert Febrer, José A. R. Fonollosa, Francesc Vallverdú
{antonio|ignasi|febrer|adrian|sisco}@gps.tsc.upc.es

Universitat Politècnica de Catalunya
C/Jordi Girona 1-3 08034 Barcelona, SPAIN

ABSTRACT

This paper summarizes the text-to-speech system that has been developed in the Speech Group of the *Universitat Politècnica de Catalunya* (UPC). The system is composed of a core and different interfaces so that it is compatible for research, for telephone applications (either CTI boards or standard ISDN PC cards supporting CAPI), and Windows applications developed using Microsoft SAPI. The paper reviews the system making emphasis in the parts of the system which are language dependent and which allow the reading of bilingual text (Spanish and Catalan). The paper also presents new approaches in prosodic modeling (segmental duration modeling) and generation of the database of speech segments, which have been introduced last year.

1. SYSTEM OVERVIEW

The UPC Text to Speech System (UPCTTS) is a bilingual system able to read text in Spanish and Catalan. Catalonia is a bilingual region of Spain, where Catalan and Spanish are spoken almost equally by a large part of the population. These two languages are similar in their set of phonemes. The system has been developed sharing as many resources as possible between the two languages, with the minimum differences in the language-dependent modules. The user can choose the desired language or, using a mark-up language, parts of text can be read in one or the other language, for instance for proper names.

The system is being used in several platforms (UNIX, MacOs and Windows32) with different objectives (research, broad applications, and specific applications). To do that, the system is composed of a core, which implement the TTS system and several interfaces. The UPCTTS system can be tried via Internet [1].

1.1. The Core of the System

The core of the TTS system is composed of four modules: text normalisation, phonetic transcription, prosody generation, and speech synthesis.

Text normalisation module. This module performs the normalisation of the text. It expands into full orthographic form acronyms, abbreviations, Arabic and Roman numerals, time expressions and dates. The input to this module is plain text, which can be optionally marked with the SABLE language [2]. SABLE is used by the text producer (a human user or an application) to control the system (language, speaker, prosody, speed, emphasis, mode of text normalisation, etc.). The use of a mark-up language is a very convenient way of controlling synthesisers: new features can be added without requiring any effort in the interface between modules.

The UPCTTS system also uses *customised* SABLE to communicate different modules. For instance, the normalisation module introduces in the plain text labels to mark the boundaries between sentences. The module also introduces marks to delimit the spelling parts of the text, which are read more slowly, etc.

Phonetic transcription module. This module makes the conversion from letters to phonetic symbols following deterministic language-dependent rules. The full orthographic text is first segmented in syllables, then stress is assigned to syllables and finally converted to phonemes. The marks of syllabification and accentuation are of key importance for a correct prosody assignment. The application of few rules is sufficient for standard Spanish. However, Catalan presents some ambiguities in the transcription: each orthographic form has a unique phonetic transcription, but in some cases, it cannot be derived from rules and a dictionary is needed to determine the transcription. For instance, vowels "e" and "o" in a stressed syllable but with no orthographic accent are impossible to distinguish whether they correspond to mid-closed sounds /e/,/o/ or to the mid-open sounds /E/,/O/.

A total of 31 allophones are considered for Spanish [3] and 36 for Catalan [4]. However, in the synthesis unit database the four affricate phonemes /ts/, /dz/, /tS/ and /dZ/ have been simplified to a combination of a plosive+fricative. The main differences are that the Catalan language has three vowels /E/, /O/ and /@/ in addition to the five Spanish vowels (*/a/, /e/, /i/, /o/* and */u/*). Furthermore in Iberian Spanish there are the sounds /jj/, /x/ and /T/ which are not present in Catalan, whereas the sound /Z/ appears only in Catalan. Finally, in Catalan word endings are not as limited as in Spanish, where only a few consonants can be found in the last position of a word.

Prosody generator. This is the principal agent in obtaining a natural sounding quality of synthetic speech. This module transforms the result of transcription into a string of allophones, each one having associated a value of pitch and duration.

The intonation model is hierarchical, being the result of the interaction of different levels. At this moment, only the sentence level and the tonic-group level are used. The sentence patterns are composed by straight lines between inflection points and

This research was partially supported by the CICYT under contract TIC95-1022-C05-04 and by CIRIT, Generalitat de Catalunya, through the Centre de Referència en Enginyeria Lingüística (CREL)

describe the evolution of the pitch along the time axis. They are represented in a parametric form so that the number of inflection points, position and frequency values of each point can be adjusted to model the intonation characteristics of each language and speaker or to represent prosodic variations. The basic intonation patterns considered are declarative, exclamation, interrogative and open (or not finished) sentences. Furthermore, different patterns can easily be included for specific applications provided that they are marked in the input text. For instance, specific patterns can be applied to directory entries.

This module also assigns the duration to each allophone. Different options have been considered as are discussed in section 2.

Speech synthesiser. This module generates the output waveform. Synthesis is performed by concatenation of recorded units consisting mainly of diphones, plus some longer units *(plosive+{r|l}+vowel)*. The TD-PSOLA algorithm is used to adapt the characteristic prosody of the stored units to the values assigned by the prosodic model. The frequency is linearly interpolated along each allophone.

For some speakers of UPCTTS system, the speech database has been extracted from non-sense carrier words. However, during last year a new Catalan female speaker has been incorporated from a generic speech database. The selection criterion used to choose the diphones from the speech is presented in section 3. One voice is bilingual which means that the speaker can read text either in Spanish or in Catalan. For this voice, the diphones and polyphones that are present in Spanish and Catalan are recorded only once achieving a very significant reduction on the required memory.

1.2. The Interfaces to the Core System

The system presented has been encapsulated in a generic library that allows defining the interfaces needed for several applications. This library is independent of the operative system and has been tested in Unix, Windows and MacOs. It allows direct control of the system functionality (in addition to the commands embedded in the input text), callback notification functions, and multiple instances running in parallel but sharing the required resources. In Windows, different interfaces have been developed for different applications. First, for telephone applications in either Computer Telephony Integration (CTI) boards (Dialogic, Natural MicroSystem, Teima Audiotex) or ISDN PC cards supporting CAPI. Second, for using the system from Internet. Third, for generic Windows applications, a SAPI compliant interface has been developed. Finally, a SSIL driver is going to be developed so that the system can be acceded by some actual applications.

2. MODELING SEGMENTAL DURATION

The previous version of the system [5] used a factorial model to model vowel duration based on phonetic studies. The evidence of the suitability of other models for modeling duration found in literature and the aim of modeling also the consonants motivated the study of several models.

2.1. The Corpus

The experimentation of this work is based on a Catalan female voice obtained from a professional radio speaker on a high-quality recording. The corpus contains 3600 short sentences with neutral intonation and constant speech speed. The number of phones is about 72000: around 54000 are used to estimate the models and 18000 are reserved for testing purposes.

2.2. The Descriptor Vector

The duration of a phone is predicted depending on different factors, which are considered to affect it. The definition of the possible factors is done taking into account the effects considered in other languages. All this factors are joined in a descriptor vector. The components of the descriptor vector associated to each phone in the corpus are phone identity, stress, phrasal position, surrounding phones, syllable length, syllable position, etc.

An analysis of the data with a Classification and Regression Trees (CART) system is used to define precisely the description vector and to evaluate the most relevant factors on phone duration. Vowels and different groups of consonants (depending on manner of articulation) are studied independently. Some already known considerations are observed, such as stress or prepausal position lengthening as of first importance, but it is also noted the importance of postvocalic vs. prevocalic phones.

CART does not generalize results benefiting from interactions of different parameters and abstracting properties which could help in estimating duration. This is why there is a need of a model that captures all the known properties that are observed in this specific domain. When evaluating the model used, different experimentation was done to evaluate those described advantages in front of sparsity and generalization.

The effects of different parameters can be additive or multiplicative. The interactions of different parameters are difficult to model in independent terms. For example, the effect of syllable coda position in phone /s/ increases its duration, an effect that is hardly accentuated in prepausal position.

2.2. The Models

Different models have been considered:

List-like approach. For each phone, the vectors with the same values define a cell and the cell is modeled by the mean duration.

Factorial model. The duration of a phone is assumed to be a product of several factors. For instance, for vowels, the duration is assumed to be

$$D(v,a,p,c,t) = F_1(v)\, F_2(a)\, F_3(p)\, F_4(c)\, F_5(t)$$

where v means the vowel identity, a the stress, p the sentence position, c means voicing, and t stands for the manner of articulation of the post-vocalic consonant.

Sum-of-products models. Sum-of-products models capture the phenomenon of directional invariance [6], which was observed in the experimentation procedure. The effects of a factor, like

stress or prepausal position, have always effects on the same direction. Observing the mean values of two vowels, average duration of non-prepausal /O/ is longer than /o/. Holding all else constant, the same vowels in prepausal position had longer duration values but holding /O/ longer than /o/.

But the effects of sentence position do not affect in the same percentage to all vowels. So the use of factorial models can not model properly this situation. Neither the use of an additive model can model the interactions between factors. A combination of sums and products are more capable of reflecting the properties of duration, as directional invariance and interactions. Several sum-of-products models have been considered to model either the duration or the logarithm of the duration, as proposed in [7], getting similar results. For instance, for vowels, the best results are obtained with

$$D(v,a,c,p,t) = S_{1,1}(v) + S_{2,1}(v,a) + S_{3,1}(v)S_{3,2}(p)S_{3,3}(c)S_{3,4}(t)$$

Consonants. The difference in the nature and properties of different groups of consonants motivates the division of the model for consonants in a set of subsystems based on manner of articulation. The capability of a sum-of-products model in extrapolating and generalizing should be more effective when the set of possible descriptor vectors is restricted to similar contexts. The analyzed subsystems were nasals (m, n, N, J), voiceless plosives (p, t, k), voiced plosives (b, d, g), fricatives (S, s, f, Z, z) and liquids (l, L, r, rr). The descriptor vector of consonants includes syllable position (onset or coda).

The comparison of the results obtained with all the models showed that the sum-of-product model obtains consistent results and more broad coverage.

From the observation of the parameter values of the model, it was noticeable that diphthongs were candidates to form a subsystem in a specific model. In the first evaluations, semivowels appeared as a parameter applied in the post-vocalic phone term. The difference of numerical order between semivowel terms and other phones confirmed the need of studying diphthong behavior independently.

Referring to consonants, from the observation of the parameters, a higher influence of accent is manifested in nasals /m/ and /n/, while the other consonants are not so affected. The effects of sentence position are highly accentuated in fricatives, especially in /s/ prepausal and in coda phone.

The observation of the behavior of the subsystem of liquid consonants suggested a different treatment. Groups formed by a plosive consonant, a liquid consonant and a vowel should have a particular treatment. The coarticulation in these cases makes difficult to determinate the correct emplacement of phone boundaries, as these liquid consonants are highly contaminated by the following vowel.

3. DATABASE GENERATION

Concatenative systems using TD-PSOLA require several speech databases to synthesize different voices. In previous work, specific speech databases were recorded. Each unit (diphone and polyphones) was recorded in non-sense carrier words. A new Catalan female voice has been added using a generic speech database. This makes possible to choose the best diphone unit from several examples. The database used is the same than the one described to study segmental duration. The total number of diphones in the database was 76194 obtaining 707 different diphones. Some synthesis units appeared frequently, the 75% of them more than 10 times. Not all the 796 possible diphones in Catalan appeared; the 89 remaining diphones are substituted for other similar units applying phonetic rules or artificially created from existent demiphones.

3.1. Automatic Labeling

The speech is aligned using a HMM tool using context dependent demiphones. These units are taken as the half of a phone and have been introduced recently giving better performance than triphones [8]. The main advantage of using demiphones for labeling synthesis databases is that the alignment provides not only the phoneme boundaries, but also a consistent point to split the phone into two parts. This point is needed to define diphones. In previous databases, the diphone boundaries were defined from the phone boundaries using a simple rule: 2/3 of the first phone plus 1/3 of the second one. In some cases, this criterion produced inconsistencies when two diphones were concatenated. If demiphones are used, the criterion to split a phone is based on the statistic of the signal (modeled by HMM) producing much better results.

An automatic epoch detection algorithm is applied to the speech database [9]. Post-filtering and error detection algorithms complete the pitch synchronous labeling of diphones, with an estimation of the F_0 mean value. This value is used to supervise possible errors in the automatic labeling procedure by rejecting out of range values.

3.2. Diphone Selection

The synthesis databases of our system include only one instance of each possible phone. The diphones are chosen to minimize distortion in output speech, which can be produced by *i)* concatenation of synthesis units, *ii)* changes in F_0 and in duration introduced by the synthesis algorithm to adapt the speech segment to the values given by the prosodic model, and *iii)* errors in the labeling (phone boundaries or F_0 labels).

The diphone is selected using a cost function that combines several information about all the instances of each diphone.

F_0 value. Large modifications of the fundamental frequency produce high distortion in the speech signal in a synthesis method like TD-PSOLA. The criterion of selection must consider the mean F_0 value in two senses. First, because extremely high or low values can indicate incorrect pitch detection. Second, because the selection of a unit with a F_0 close to the mean F_0 of the speaker implies that modifications will be reduced in most cases during the synthesis, resulting in lower distortion of signal.

Duration of the segments is another parameter to be considered in the selection of units. Too short diphones would systematically degrade the synthetic speech. Furthermore, this parameter can also indicate possible elisions or phones highly contaminated by coarticulation. Other considerations, like deviation respect mean duration of the segment, were evaluated. However, as the synthesis technique allows modifications of duration with low distortions, the final cost function only penalizes segments with short duration.

Spectral characteristics in the boundaries. In concatenative synthesis quality is degraded when acoustic distances between concatenation points are large. Spectral distortion in segment boundaries seems to be a more adequate criterion to evaluate the adequacy of a diphone to concatenate with another diphone. Because only one diphone is kept in the database, the idea is to choose the one whose boundaries are more *neutral*, i.e., more *standard*. This has been implemented using a measure between the boundaries of the diphone and the mean of all the boundaries of the database: first, for each phone present in the database, the Mel-cepstrum is computed in the point were the phone is split to form the two diphones. Then, for each phone, the mean of the cepstrum is computed and is stored as a reference. Finally, for each diphone, the penalization value is computed as the sum of the Mel-cepstrum distance of the two boundaries and the correspondent references. For instance, the penalization value for an instance of the diphone *xy* is the sum of the distance from the beginning boundary of the diphone to the reference of phone *x* plus the distance from the ending boundary of the diphone to the reference of phone *y*.

This value ensures that the diphones are *standard* and therefore *i)* they are not a particular instance and they are not too contaminated by the neighbors phones and *ii)* the segmentation point is also *standard*, avoiding possible errors or deviations in the automatic segmentation described at the beginning of the section.

Cost function. The choice of the selection method has been based on subjective evaluation. As a consequence it is no feasible to propose a generic function and maximise the weights to optimise some criteria. Instead, a very simple rule has been used giving priority to small spectral discontinuities, small F0 deviation (quantified in three levels) and duration larger than a minimum.

The effect of the selection method has been evaluated subjectively, hearing synthetic speech. This method has shown to be very effective to eliminate errors on the automatic labelling and discontinuities in the concatenation of the speech segments.

4. SUMMARY

In this paper, the text to speech system developed in the UPC has been described. The system allows the reading of text in Spanish and Catalan from several platforms. There is a bilingual voice, which shares all the common diphones and polyphones between Spanish and Catalan, with a substantial reduction of the memory needed to store the speech database.

Different interfaces have been developed so that the system can be used from different applications: research, telephone applications using CTI boards or ISDN PC cards, Internet, and any Windows application that uses SAPI.

The paper also presents a study about segmental duration. The sum-of-products model has shown to be very effective for modelling vowels and most clusters of consonants. Diphthongs and liquids require a posterior study.

Finally, the paper has presented a method to build synthesis databases from generic speech corpus. The database is segmented into phones using a HMM tool. The sublexical units of the tool are not phones but context dependent demiphones. In this way, not only the phones boundaries but also the diphones boundaries are provided automatically avoiding the errors which are produced when the diphone boundaries are determined by a simple rule from the phones boundaries. The units of the database are selected taking into account the mean pitch, the duration of each unit (a minimum value for each unit is required) and the distance between the boundaries of the unit and the mean of all the boundaries of the database. The method has proven to be very useful for automatic database generation, although a final supervision of the labelling improves the quality of the synthetic speech.

The UPCTTS system can be evaluated from the Internet site [1]

5. REFERENCES

1. URL: http://gps-tsc.upc.es/veu/demos.html
2. URL: http://www.cstr.ed.ac.uk/projects/ssml.html
3. J.B. Mariño, *"Reglas para la transcripción fonética aplicadas en RAMSES"*, Research Report, UPC, June 95
4. A. Pujol, I. Esquerra, "*Regles de transcripció fonètica del català*", Research Report, UPC, October 1996
5. A. Bonafonte, I. Esquerra, A. Febrer, F. Vallverdú, "A bilingual text-to-speech system in Spanish and Catalan", *Proceedings of EuroSpeech'97*, pp. 2455-2458, Rhodes 1997.
6. *Multilingual Text-To-Speech Synthesis: The Bell Labs Approach*, Richard Sproat, editor, Kluwer Academic Publishers, 1998.
7. J. van Santen, "Assignment of segmental duration in text-to-speech synthesis", *Computer Speech and Language*, 0:95-128
8. J.B. Mariño et alA. Nogueiras, A. Bonafonte, "The Demiphone: An Efficient Subword Unit for Continuous Speech Recognition", *Proceedings of EuroSpeech'97*, pp. 1215-1218, Rhodes 1997.
9. J.L. Navarro, I. Esquerra, "A Time-Frequency Approach to Epoch Detection", *Proceedings of Eurospeech'95*, pp 405-408, Madrid, 1995