Diphone-based Unit Selection for Catalan Text-to-Speech Synthesis

Roger Guaus i Térmens, and Ignasi Iriondo Sanz

Secció de Tecnologies de la Parla, Dept. de Comunicacions i Teoria del Senyal
Enginyeria Arquitectura La Salle, Universitat Ramon Llull
Pg. Bonanova 8, 08026 Barcelona, Catalunya
{rguaus, iriondo}@salleURL.edu
http://www.salleURL.edu

Abstract. This paper describes a Unit Selection system based on diphones that was developed by the Speech Technology Group of the Enginyeria Arquitectura La Salle School, Universitat Ramon Llull. This system works with a PSOLA synthesizer for Catalan language which is used in an Oral Synthesized Message Editor (EMOVS) and Windows applications developed using Microsoft SAPI. Some common questions about Unit Selection are formulated in order to find solutions and achieve a better segmental speech quality.

1 Introduction

Latest investigations in our Speech Technology Group have been focused in Catalan language text-to-speech conversion based on unit concatenation (i.e. diphones and triphones).

Catalan is the native language of Catalonia, the Valencian Country and the Balearic Islands (central east and north east part of the Iberian Peninsula), which is spoken by more than 6 million people. Nevertheless, these are bilingual areas where Spanish is also spoken, being, in fact, the dominant language used for communication (news, TV, radio, ...) in the region. So Catalan is a minority language in front of Spanish influence. Like other languages, Catalan has several varieties (Central, North-occidental, Valencian, Balearic and others) that are spoken in different areas of the territory. We have focused our studies in the Central variety which is spoken near the city of Barcelona.

1.1 Previous studies

Unit concatenation has been the focus of our work during the last years. As it is well known, the segmental quality of synthetic speech is limited by the number of joins that can be encountered in an utterance. In order to resolve these join mismatches and to improve the segmental quality of the speech, different systems based on different ways of coupling diphones were developed, [1], [2] and [3]. The synthetic speech obtained with these systems had good intelligibility but it was not natural enough to be used in most applications (e.g. broadcast news, ...). Some experiences based on phone unit selection [4], [5] and [6] were developed to improve the generated speech quality but since poor quality was achieved, new systems based on half phone unit selection [7] came into consideration.

We propose an approach for adapting a text-to-speech system based on diphone and triphone concatenation architecture (using a Catalan pre-designed database) to a unit selection structure. Therefore, there is a basic difference between a traditional diphone- (i.e. half-phone-) based traditional system of unit selection [7] and ours: we chose to use diphones and triphones as basic units, as opposed to half phones (or half diphones).

2 Unit Selection based on Diphones

While a phone-based unit selection system has about 37 basic units for Catalan language, a unit selection system based on diphones has to consider all possible combinations among phones (i.e. more than 1200 units[1]).

Since the basic units in the present method are diphones and triphones, this approach will provide, at least, the same speech quality than a diphone- and triphone-based synthesis system (without using unit selection). To increase the segmental quality of the generated speech using unit selection, more units have to be added to the basic corpus. So, a unit selection database for diphones has to be divided into two halves. The first half contains only one version of each unit and it has to be designed for each language. The second half is created by adding a large set of texts. Different versions of the units described in the first half of the database are extracted from these texts.

2.1 The Designed part of the Database

This database half is specially designed for Catalan language and it consists on 895 diphones and 312 triphones (1207 units on the whole). This total amount of units ensures all possible combinations of phones and they are enough to create any utterance in Catalan.

Since not all the possible combinations among phones are allowed (i.e. aleatory combinations), the phonetic transcription module should only generate pairs included on these 1207 units. Other non-pronounceable combinations are not allowed. This is not a hard objective in Catalan, since a phonetic transcription module can be implemented with a limited set of rules and a few exceptions (Catalan phonetic transcription is more difficult than Spanish and quite easier than English). Notice that phone-based and half-phone-based systems permit all possible combinations between phones because the phone pairs are not defined. This could generate combinations among phones that could not be actually pronounced.

[1] Let mention 'units' as a synonymous of 'diphones and triphones'.

This already-designed database is the same one used for a simple diphone concatenation system (without using unit selection) previously developed for Catalan and described in [3].

2.2 Non-Designed part of the Database

This half of the database is generated by means of a large set of non-designed texts in order to ensure the availability of different versions of each unit defined in the first part of the database. Therefore, in order to generate this half of the database, the system searches the units described in the first part and includes the new version on the non-designed database. Usually, different versions of a unit are caracterized by the prosody (pitch, duration, energy, ...) and the articulation context.

An important difference between this system and a phone-based one is that, in the former, mid-phone joins are allowed, while in the latter only boundary joins were permitted.

3 Database evaluation: Super-Unit Mean Length parameter

Some questions usually arise when unit selection systems have to be designed: How do units interact in a unit selection database [8]? How does speech quality improve in front of database size increase? How large the database has to be? How does the database style influence the segmental speech quality?

In order to answer these questions we have defined a new parameter to estimate the segmental quality of the speech: Let L_{su} (Super-Unit Length) be the number of units consecutively chosen from the database to generate a synthesized utterance. These consecutively chosen units constitute a super-unit. Then, a super-unit contains at least one unit.

The segmental quality of the speech is directly related to the mean length of the chosen super-units. So, let us define $\bar{L}_{su}$ (Mean Super-Unit Length) as

$$\bar{L}_{su} = \frac{1}{N_{su}} \sum_{i=1}^{N_{su}} L_{su}(i) = \frac{N_u}{N_{su}} \quad (1)$$

where N_u is the number of units and N_{su} is the number of super-units, both in the generated utterance.

Let us generate the Catalan utterance 'Un hivern plàcid' ('A quiet winter'). Table 1 shows the units and super-units that are chosen by the system[2]:

To generate the desired phrase, ten units are selected from the already-designed half of the database. Then the unit selection process searches in the non-designed database the longest units that match with the input transcription. In this case five super-units are found in the process output. It can be easily seen

[2] SAMPA notation is used

Table 1. Unit selection example for a Catalan utterance

Text	*Un hivern plàcid*
Transcription	[uniBERmplasit]
Units	[un] [ni] [iB] [BE] [ERm] [mp] [pla] [as] [si] [it]
Super-units	[un] [niBERm] [mp] [plas] [sit]

that, for example, $L_{su}([niBERm]) = 4$ (this super-unit is composed by 4 units). If we apply equation (1) it can be seen that, in this case, $\bar{L}_{su} = 2$.

$\bar{L}_{su}$ increases when super-units are longer (and then the number of super-units decreases) and the super- unit length depends on different parameters such as database size, database style vs. target style correlation and database utterance mean length ($\bar{L}_{DB}$).

4 Experiment

A mixed database divided into two parts is used. The already-designed half of the database is extracted from a previous text-to-speech converter for Catalan language. This Catalan corpus consists on 1207 units (including diphones and triphones).

In order to evaluate the style influence, two different non-designed databases have been created. Politics-style and Society-style non-designed databases containing 86,000 units each have been extracted from the Catalan newspaper AVUI [9].

On the other hand, three different style test texts (politics, society and literature) have been chosen from books and newspapers to test politics-style and society-style databases. These three different texts allow us to evaluate the style influence on the database and they are large enough to ensure that the value of $\bar{L}_{su}$ for one of the databases will not change when more test text is added (i.e. to ensure reliable statistics).

For evaluating the behavior of $\bar{L}_{su}$ in front of database size variation, the value of $\bar{L}_{su}$ has been measured using a database from 2,000 to 86,000 units in steps of 2,000 units. This process has been performed 6 times (two database styles against three test text styles). The value of $\bar{L}_{su}$ is calculated as explained before (see Sec. 3) using a Viterbi search algorithm.

5 Results

The results of $\bar{L}_{su}$ are obtained depending on database size . On figure 2 it can be seen that the obtained curve works as a permanent convex growing function. Therefore, it can be said that an increase in the size of a small database is fairly more noticeable than the same increase in a larger database.

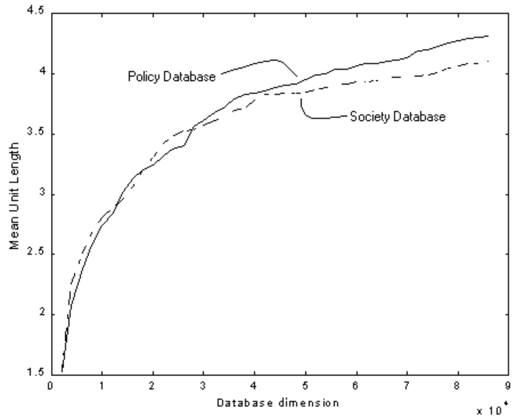

Fig. 1. Super-unit mean length vs. database dimension

Figure 1 shows curves which are generated by checking the politics-style database against the politics-style test text and the society-style database against the society-style test text. The behavior of $\bar{L}_{su}$ is similar in both cases.

The interaction between different test text styles and database styles is shown in figure 2. The best result of $\bar{L}_{su}$ for the politics-style database is obtained when the politics-style test text is applied. Society-style test text has a quite lower result than the policy-style one. Literature-style test text has the worst result because it includes a large amount of unusual words.

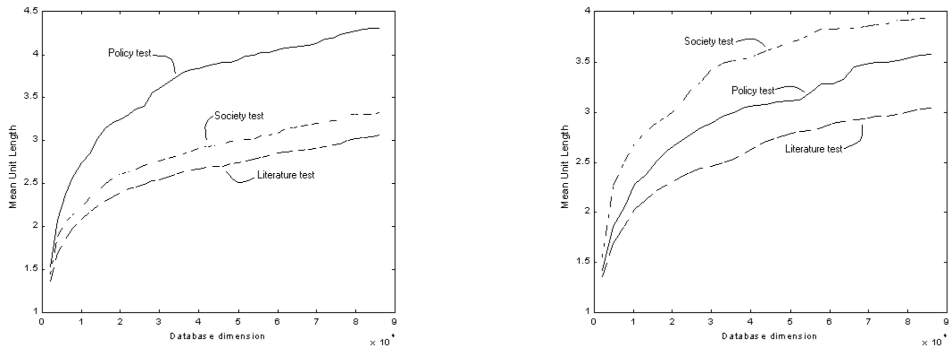

Fig. 2. Super-unit mean length vs. of politics-style and society-style database size respectively, using three different test text styles.

6 Conclusions

As it can be seen in the curves, about 45,000 units for the size of a database is an acceptable value in order to obtain a $\bar{L}_{su}$ close to 4 units per super-unit. A larger database would generate a high computation load without achieving any clear improvement.

Database style is an important parameter which determines the final value of $\bar{L}_{su}$. Notice that the difference between politics and society styles in figure 3 is around 0.5 units per super-unit. Unlike what could be expected, when database size is increased, the influence of the style remains equally strong.

Which is the relation between $\bar{L}_{su}$ and the segmental speech quality? Since $\bar{L}_{su}$ is a mean value, generated speech can have a signal portion without transitions (consecutive units from the database) and another signal portion with unit transitions (units separated from the database). This produces quality variations, so that global quality is not uniform. The main disavantage of this system lies in the fact that it depends on a designed database, so that new voices are difficult to create because more than 1,200 units have to be specially recorded. However, this already-designed part of the database does not add design costs.

On the other hand, the text extracted from Catalan newspapers contains foreign words, specially from Spanish, that are transcribed as if they were Catalan although they should be pronounced in the original language. In this respect, it might be convenient to consider the use of text sources other than the mentioned ones.

We hope that all of these ideas help to achieve a high quality text-to-speech synthesizer in Catalan language.

References

1. Guaus, R., Oliver, J., Moure, H., Iriondo, I., Martí J.: Síntesis de voz por concatenación de unidades: Mejoras en la calidad segmental. Tecniacústica 98, Lisboa (1998) 123–125
2. Guaus, R. Oliver, J., Gudayol, F., Martí, J.: Síntesis de voz utilizando difonemas: Uniones entre vocales. SEPLN 97, Madrid (1997) 234-456
3. Guaus, R.: Implementació i millores d'un sistema de síntesi de veu d'alta qualitat utilitzant PSOLA. Projecte final de Carrera, ETSETB, Universitat Politècnica de Catalunya, Barcelona (1999)
4. Black, A.W.: Optimizing Selection of Units from Speech Databases for Concatenative Synthesis. EUROSPEECH 95, Madrid (1995)
5. Conkie, A.: Robust Unit Selection System for Speech Synthesis Joint Meeting of ASA, EAA and DAGA, Berlin (March 1999)
6. Beutnagel, M., Conkie, A., Schoeter, J., Stylianou, Y., Sydral, A.: The AT&T Next-Gen TTS System. Joint Meeting of ASA, EAA and DAGA, Berlin (March 1999)
7. Beutnagel, M., Conkie, A., Sydral, K.: Diphone Synthesis using Unit Selection. 3rd ESCA/COCOSDA Workshop on speech synthesis. Jenolan Caves, Austalia (November 1998)
8. Beutnagel, M., Conkie: Interaction of Units in a Unit Selection Database. EUROSPEECH 99, Budapest, Hungary (September 1999)
9. Avui Catalan newspaper URL: `http://www.avui.com`