Automatic Segmentation of Speech Units

Jorge A. Gurlekian and Humberto M. Torres

Laboratorio de Investigaciones Sensoriales CONICET
Hospital de Clínicas. Neurociencias. UBA. Buenos Aires

Abstract

An automatic segmentation method is tested here, which uses a combination of entropy coding, continuous multiresolution analysis, and Kohonen's self organized maps. The method considers that there are no limits imposed by any linguistic unit. Resulting waveforms represent phone chains dominated by spectral dynamic structures. Each obtained acoustic unit could be composed of a variable number of phones or a segmented part of them at their boundaries. The amount of units, unit structure and unit repetitions are speaker dependent, i.e. rate, segmental and suprasegmental distinctive features affect them as dynamic structure varies. Results obtained from two databases -one female, one male- being the female speaker with slower speaking rate than the male speaker - of 741 sentences each show this dependence, presenting both different number of total units, and occurrences for each speaker. Nevertheless, both speakers show a high occurrence of two and three phone sequences in overall measurements considering repetitions. For two phone sequences Consonant-Vowel syllables, which are phonemically frequent in Spanish (58%), are the most frequent type for the speaker with low speaking rate (22.4%) than for the speaker who speaks faster (9%). Three phone sequences show that Vowel-Consonant-Vowel is the most frequent type with similar percentages for both speakers (7.8-9.86%). When we count only different sequences without repetitions the most frequent sequences are composed by three phones and again Vowel-Consonant-Vowel sequences dominate the figures. Most frequently segmented phones are vowels, unvoiced stops and fricatives. The relevance of half phone segmentation is verified given that for both speakers 71% and 57% of the total units start and end with a segmented phone. These results are relevant to the design of text to speech systems and to further explore in speech perception and automatic speech recognition systems.

1 Introduction

Concatenation of natural speech is the standard technique for synthesis in text to speech systems (TTS). Speech quality is improved when phonemes and syllables are segmented at places where spectral variations are small. TTS systems improve in intelligibility and quality when linguistic units are segmented to form diphones or demisyllables to be employed in synthesis by concatenation. The goal is to preserve natural coarticulations between phones and then solve the mismatch of the stationary half portions before concatenation. These main assumptions to obtain natural speech have already been stated by Peterson (1958): “if synthesized speech is to sound natural, the normal dynamics of speech production must be maintained”. This author proposed dyads, actually diphones, and predicted misalignments of steady states during concatenation. Considerations about both fundamental frequency and overall energy matches between segments were also mentioned in their work in order to prevent clicks. Further improvements were obtained when searching for units that meet specific concatenation demands of F0, energy and duration values to reduce errors (Campbell, 1996; Black, 1995; Hunt, 1996).

The study of prosody contributed to improve sentence intonation and relative duration. More recently, some authors have proposed to choose those units that best approximate tonal accents in intonation contours. Campbell (1997) proposed a prosody-based synthesis unit selection process to
generate high definition speech. Trying to approximate natural speech, some practical contributions consider larger contexts than the diphone when accessing the database. For this purpose, context dependent phone units, such as triphones or quinphones, have been considered (Hon, 1998).

As a general rule the larger the context, the better the quality on concatenation to produce naturalness, and the greater the number of units on the database. The use of large units would normally require larger databases which are tedious to create as is the case with speaker dependent synthesis. Recent work (Prudon, 2001) showed that searching for a variable phoneme chain improves the results when compared with diphone concatenation. The authors confirmed that taking adjacent phones whenever possible is a better concatenation criterion than any other. Recent work for Spanish also confirms that the selection of intonation contours is highly improved by the presence of consecutive diphone units (Diaz, 2006). The question that arises is where and how to stop in this phone search.

In this work we apply a novel segmentation method to extract units from a speaker dependent database before synthesis takes place. Unit size will depend on each particular speaker, i.e. if the speech is fast, units will tend to be large, because acoustic dynamics will mainly reflect transitions between phones, rather than stationary segments.

Signal representation is based on wavelets, which gives a detailed frequency description at low frequencies and detailed time discrimination at higher frequencies. In this way we expect that the F0 contours will be well represented and its dynamic contribute to describe the tonal accent when present in each unit.

The speech signal representation is evaluated by a combination of 1) Continuous Multiresolution Entropy (CME) criteria (Torres, 2000), useful for detecting parameter changes in non-linear dynamic systems, and 2) Kohonen's Self Organized Maps (SOM's), for its capacity to cluster and visualize high dimensional data. As a consequence of this approach, units could be composed of a variable number of phones, starting and ending according to its acoustic dynamics.

We expect that the proposed segmentation method obtained off line would favor the posterior unit selection process -based on the information produced by both the intonation and duration models-during text to speech conversion.

The outline of this paper is as follows: In section 2 we present the database definition. In section 3 we briefly describe the segmentation tool (CME+ SOM's). Section 4 presents the results obtained from this segmentation tool. In Section 5 we discuss these results in order to evaluate our proposal. Finally, in section 6 we present our conclusions.

2 Database definition

The database consists of 741 declarative sentences extracted from Buenos Aires Argentine newspapers. The sentences contain 97% of all Spanish syllables, in both stress conditions and all possible syllabic positions within the word. Two native speakers female (slow speaking rate) and male (with fast speaking rate), read the sentences in a sound proof chamber. Recordings were made with an AKG dynamic microphone and 16 kHz/16bit conversion. The speakers were instructed to read the sentences with natural tonal variations.

Each sound file was manually labelled twice, by musically trained speech therapists who distinguished prosodic occurrences as intonational groups and accents. The files were labelled in different layers: phonetic, orthographic, pause levels between words, and tonal marks according to a phonetic method (Gurlekian, 2004). Parts of speech and syntactic layers were also indicated.

2-1 Intonation labelling

The tonal accent set called Sp-ToBI (Beckman et al, 2002) were used to define a set of phonetic tonal labels to characterize the intonation contours. To build up this phonetic set, additional information was added to H&L tones in order to capture variations that must be preserved to have high quality synthesis and eventually to study region variants and other variations by means of synthesis and perception. The included information is the F0 level and adjacent slopes. F0 level information was introduced by translating the lineal frequency values to the more compact ERB rate scale which has
been effective to measure equal prominence at different F0 registers. The ERB rate scale range from 1 to 12. Slope information was given by the syllable position of the complementary tone in bitonal accents –given by the number of syllables- apart from the star tone. Number of syllables range from 0 to 7. Table II shows examples of the phonological to phonetic transformation.

<table>
<thead>
<tr>
<th>Sp-ToBI</th>
<th>Associated Phonetic Transcription</th>
</tr>
</thead>
<tbody>
<tr>
<td>H*</td>
<td>6.42 H*</td>
</tr>
<tr>
<td>L*</td>
<td>4.23 L*</td>
</tr>
<tr>
<td>H* + L</td>
<td>5.86H* + 4.76L0, H*+4.3L0</td>
</tr>
<tr>
<td>L + H*</td>
<td>4.8L0+5.72H*, 5.22 L1 + 8H*+8.3H0</td>
</tr>
<tr>
<td>L* + H</td>
<td>(L*+!H) 3.2L* + 6.3 H1</td>
</tr>
</tbody>
</table>

**Figure 1.** Set of Sp-ToBI tonal accents (left), and some examples of phonetic labels (right). Shadow examples are used in Figure 2. Phonetic labels can be converted back to the original labels removing the added information.

**Figure 2.** Fundamental frequency contour of the natural sentence: “Informes e inscripción en la misma oficina”. Straight lines are drawn by linear interpolation of circles defined by the phonetic tonal labelling (“Ton” line marked at bottom). Over the contour, tonal labelling according to Sp-ToBI is shown, in correspondence with the phonetic labelling.

### 3 Segmentation Tool

In this section we will describe the method used for definition and segmentation of the acoustic units, denominated CME+SOM. The capacity to segment speech using entropy has been tested previously (Wokurek, 1999). Recently, Torres et al (2006) used a combination of continuous multiresolution entropy criteria and Kohonen’s self organized maps to detect and cluster slight
parameter changes in the dynamics of natural and synthetic vowels in diphthongs. Below we briefly describe the steps corresponding to the CME+SOM algorithm. For a detailed description see Torres (2006).

1. The speech signal was sampled by a regular sampling of 16 KHz and 16 bits resolution.
2. For each analysis window, the corresponding wavelet coefficients were obtained. Wavelet analysis was chosen because of its close properties to human-inner ear resolution-analysis (see figure 4). A suitable wavelet, as the Mexican Hat was evaluated, which is acknowledged to have the best temporal localization properties (Daubechies, 1992).

We have computed the Continuous Wavelet Transform (CWT) using the time-frequency toolbox developed by Auger et al (1996).

3. For each scale of the CWT, the Shannon's entropy (Shannon, 1948) was evaluated by sliding windows of length \( L = 800 \) and a shift of \( m = 160 \). In this way the Continuous Multiresolution Entropy (CME) was calculated. (See: Torres, 2000, for the computational details). With each of the CME matrices the total matrix CME was formed by concatenation.

For each scale, the corresponding CME was statistically normalized to zero mean and unit standard deviation, resulting \( Z \) the statistically normalized matrix associated to CME. The matrix

![Figure 3](image)

**Figure 3.** Main steps of the method used for segmentation. (1) A/D: Analog to Digital converter. (2) CWT: Continuous Wavelet Transform, (3) CME: Continuous Multiresolution Entropy, and (4) SOM: Self Organized Maps.

![Figure 4](image)

**Figure 4.** Time-frequency resolution for the Short Time Fourier Transform (top), and Wavelet Transform (bottom).
Z qualitatively reveals the occurrence of a slight parameter change in the underlying nonlinear dynamical system by means of a jump (up or down) at all the scales.

4. Each column of the matrix Z is used like an input vector to train a Self Organized Map (SOM). The SOM-PAK (Kohonen, 1995a, 1996b) software was used in order to do the experiments. At this stage, we needed to select the topology, dimensions and training parameters of the map. In general, there is not an a priori rule to fix these parameters and for each case there is an optimal set of parameters. The dimensions of the map and neuron disposition in the arrangement are critical items. In this study, a linear arrangement of neurons was defined, since it allows an easy visualization and analysis of results. The number of neurons was empirically determined and fixed in four, considering two steady states and transitions between them. The training parameters are chosen to make the SOM to converge. In general, it is accepted that this is obtained with small values of learning coefficients and a great number of iterations.

Once the SOM is trained, temporal evolution of the winner neuron is considered as the system output. In such a way, changes in the winner neuron represent variations in the system parameters and, for parameters of similar values, neighbour winner neurons are obtained. We must emphasize that, for different training data sets and/or SOM parameters, different output configurations will be obtained, since these depend on the training, although the data represents the same configurations of the system.

In Figure 5 we present the speech waveform corresponding to the phrase “Cada camión carg(a)...” ("Each truck loads..."). SAMPA phoneme labels are also displayed (dashed line). Solid lines represent the evolution of the winner neuron and the triangles indicate the limits of the obtained units. Here, we took the neuron number 1 as indicator of a point to segment. When this neuron number 1 remain the winner by a time interval, the interval's midpoint is taken as the point to segment.

4 Results

The results obtained from two databases originated by one male speaker (M) and one female speaker (F), of 741 sentences each, are shown in Figure 4. Acoustic units are composed of a variable number of phones. Both speakers show a high occurrence of two, three, and four phone sequences,
where the first and/or last phone could appear segmented. Each unit could appear segmented as the following four types: with the initial and final phone segmented called initial-final (IF), with only the final phone (F) segmented; with only the initial phone segmented (I) and preserving entire phones at the boundaries (0).

**Figure 6.** Left: Percentages of occurrence of units grouped according to segmented types. (IF: initial-final phone segmented; F: final phone segmented; I: initial phone segmented; 0: no phone segmentation). Right: Percentages of occurrence of units grouped according to the number of phones for all segmented types. Sp1(female, low speaking rate), Sp2(male, fast speaking rate).

As shown in Fig. 6, around sixty percent of total acoustic units, distributed as 71% and 57% for female and male speakers respectively, start and end with a segmented phone (IF). The most frequent units are composed by two, three, and four phones.

**Figure 7.** Results for unit type (IF). Overall number of appearances are compared with the number of different units. Left: Speaker1(female), Right: Speaker2(male)
Overall occurrences of vowel and consonant combination sequences for the female and male speaker. The letter “S” on top indicates that the sequence is also a syllable. Total number of sequences are 16169 and 14875 for the Speaker1 (female) and Speaker2 (male) respectively.

Within two phone sequences the consonant-vowel (CV) and the vowel-consonant (VC) are the most frequent types. For phone sequences without repetitions the results show a dominance of three phone sequences as shown in Figure 9, where vowel-consonant-vowel (VCV) sequences have the higher occurrences. Two and four phone sequences complete the most significant percentages. Higher number of phones sequences have percentages less than 1% and are omitted in the figure.

Figure 8. Overall occurrences of vowel and consonant combination sequences for the female and male speaker. The letter “S” on top indicates that the sequence is also a syllable. Total number of sequences are 16169 and 14875 for the Speaker1 (female) and Speaker2 (male) respectively.

Figure 9. Different occurrences of vowel and consonant combination sequences for the Speaker1 (female) and Speaker2 (male).
Vowels are natural candidates to be segmented because of their length and steady states. Unvoiced fricatives -noises- \([s, f, x, H, Z]\), which are somewhat independent of vowels because of their length and source, present the \(/s/\) sound as the most frequently segmented, both preceding and following a vowel. Vowel-like sounds as \([l]\) and nasals \([m, n]\) perceptually characterized as steady and independent sounds, are also often segmented. Unvoiced stops \([p, t, k]\) which are more dependent on the following vowel than the preceding one, appear integrated to the following vowel as in CV sequences and are often segmented at the silent interval. See Figure 10.

For three phone sequences the vowel-consonant-vowel (VCV) combination is the most frequent type, where voiced consonants at the middle of the sequence and formant transitions to adjacent vowels compose a dynamic structure which is preserved by the segmentation method. Transitional consonants like approximants \([B, D, G]\), and \([r]\) are seldom segmented with this method, because of their strong dependence on vowel context.

![Figure 10](image.png)

**Figure 10**: Percentage of occurrence of segmented consonants (C) and segmented vowels (V). Results are obtained from a total of 7297 initial phones and 8138 final phones for the male database and 13605 initial phones and 13806 final phones for the female database. Initial segmented phones are derived from initial (I) and initial-final (IF) segmented units. Final segmented phones are derived from final (F) and initial-final (IF) segmented units.

5 Discussion

The results shown in Figure 6 confirm that 64% of the units in average obtained from for the female and male subjects databases follow general diphone strategies regarding the adequacy of half phone segmentation at the juncture. Most of the repeated units are composed by two and three phone sequences (see Figure 8). These results also agree with standard diphone segmentation, but also suggest the use of a complementary set of triphones for certain consonants as is discussed below. That is, spectral dynamic information seems to be extended during more than two phones. This is due to certain voiced consonants, and also due to speaking rate. Particularly, the male speaker (age: 24), who speaks faster than the female speaker (age: 41), presents more quantity of long dynamic units and less
amount (half) of two phone units than the female speaker. Furthermore, the male speaker presents more unit segmentations at unvoiced stops \([p \ t \ k]\) and fricative \([s]\) and less in nasal and liquid voiced sounds when compared to the female speaker (see Figure 10). This may be due to the fact that stop silence and fricative duration are obligatory distinctive features for these sounds and although his speaking rate is fast he must preserve those features.

When units are counted without repetitions, different units composed by three and four phones, are more frequent as expected from increased phoneme combinations, but again three phone sequences are more frequent favouring VCV sequences due to the high number of vowel segmentations. (see Figure 9).

Large units of three to four phones tend to preserve the natural shape of the tonal accent compared to smaller units and they seem to be more adequate for further prosody modification. Inversely, a posterior modification of the intonation contour, starting from units with neutral accent as in regular diphones or from units with any F0, produces an unnatural sensation. This is due to the dichotomy between spectral information -with harmonics transporting information about F0- and the modified source information, which the ear can detect. We explored accented and unaccented units on the data base to see the distribution of tonal accents in relation to different size units. Figure 11 shows that 75% (compared to Figure 7 left) of the units are unaccented. From the remaining 25% of accented units, the proportion shows a decrease of two phone units or a relative increment of three and four phone accented units. According to these result some units are best associated to the presence or absence of tonal accents.

![Figure 11. Distribution of occurrences of both unaccented and accented units for the initial-final type. Groups of columns indicate different number of phones.](image)

For a complete strategy based on larger units with two or three tonal accents alternatives, the number of units available must be increased considerably. An intermediate solution could be to use a combination of diphones with the available larger units as was shown in this work. As an example, considering 30 phones of the Argentine SAMPA alphabet there are about 700 possible diphones. The size of this database is considered acceptable because diphone units could be available within it, by means of a conventional method, if necessary. In this work, 300 diphone units are already available after segmentation. Two phone inventory is completed with initial diphones (176), final diphones (198) and two entire phones (86) for a total of 760. If the database increases in size, more three- and four-
phone units will be available and more instances of specific tonal accents would be encountered in them, thereby increasing the quality of synthesized speech.

6 Conclusions

In this study, the effectiveness of a method to automatically detect dynamic changes and segment speech sequences at stationary portions was tested and validated for continuous speech. Resulting dynamic speech sequences determine units with a length that is a direct consequence of each speaker’s rate.

This method confirms that spectral dynamic information continuously goes beyond two phones for voiced consonants. Most frequent dynamic units have two, three and four phones. Text to speech conversion can be performed with a regular size database using a combination of diphones and these added dynamic units to improve quality of synthesized speech.

Future work related to text to speech synthesis considers the automatization of unit selection, taking into account other aspects related to rhythm, such as the incorporation of a duration model. We expect to explore the relevance of these units in both speech perception and automatic speech recognition, in an attempt to answer related questions such as: What is the perceptual meaning of these acoustic dynamic units compared to phonological units? Are the units signalling preferred triphone and tetraphone sequences, which must be weighted differently in our acoustic models for speech recognition?

References