Models and Analysis of Vocal
Emissions for Biomedical Applications
(MAVEBA 1999)
Firenze, Italy, September 1-3, 1999

WAVEFORM MODELING OF NASAL TO VOCALIC VOICE EVOLUTION

ROSANA G. BARBUZZA* JORGE H. DOORN

Universidad Nacional del Centro, Facultad de Cs. Exactas
ISISTAN Instituto de Sistemas de Tandil
Campus Universitario, Paraje Arroyo Seco, Tandil (7000), Argentina
E-mail: {rbarbu, jdoorn}@exa.unicen.edu.ar

Abstract

One of the less studied aspects in human voice processing is how the transition articulations between phonemes occur in fluent speech. Often, lack of information pushes us to deal with them using black-box approach. This is especially true in the field of voice synthesis. In this article the transitions from nasal consonant to every possible vocalic allophone of Spanish language is analyzed using a cycle based nonlinear transition model. This method has been successfully used in every transition between voiced allophones. The data showed in this paper corresponds to the Riverplatean Spanish.

1. Introduction

The vocal tract apparatus produces a large variety of phones, but only few of them are speech sound. Speech is the acoustic consequences of air flowing through the vocal tract. The basic theoretical unit to describe how speech carries linguistic meaning through sound is called phoneme. A phoneme can be seen as an ideal sound unit produced in a given 'static' position of the vocal tract [1][2].

The latest developments in speech synthesis and recognition researches are based on digital signal processing implemented in computer [3][4][5][6]. The production of natural and continuous speech is a challenging task, since most of the known approaches only produce sound-like "artificial" voice. Speech synthesis has been a field of research along several centuries, with important developments in the last years. Early attempts resulted in mechanical and electrical devices, however the main progress has been accomplished by means of the digital processing of the acoustic signal, providing a new technical profile to the phonetic area [7][8].

The ability to produce speech sounds does not imply ability to speak. Simply concatenating speech sounds would result in an unpleasant and poorly intelligible acoustic signal. In actual speech, articulators move continuously and the resulting sound is continuous rather than a discrete combination of individual sounds. Since the purpose of speech is to convey meaningful messages, humans also help each other to understand their messages by varying pitch, loudness and duration of the basic sound, according to the linguistic structure of the intended message [9].

The production of speech requires movement of the articulators in the vocal tract. Unlike distinct characters found in printed text, phoneme articulations typically overlap each others in time, thereby causing sound patterns to be in transition most of the time. Instead of a quick rigid articulator movement between uniform islands of stationary phoneme production, speech is normally produced with smooth movement and timing of the articulators to form the proper evolution of the vocal-tract shapes to produce the desired phoneme sequence.

The sounds corresponding to a phonetic string are strongly influenced by their position in the sequence, since the incoming and the outgoing phoneme affect each phoneme [10].

In this article the use of a time-domain approach [11] to model the transitions between phonemes, is explored in detail. The set of transitions chosen are for /m/ /n/ /ñ/ phonemes production in the onset of the syllable toward vocalic phonemes in the nucleus of a syllable.

* Also Comisión de Investigaciones Científicas de la Provincia de Buenos Aires

2. Acoustic waveform of a phonetic stream

The purpose of phonetics is to study the different sounds of speech, especially in relation with description, taxonomy and transcription of sounds. In the phonetic alphabet each basic sound is represented by the same symbol, and each symbol must represent the same sound [12][13][14][15].

Phonetic symbols correspond to the allophone units, which do not coincide with acoustic waveform frames, since speech is a continuous signal where limits are difficult to establish. The segmentation of sounds in an acoustic sequence is not a simple task, a voice frame can hold a unique vocal or consonant sound, as well as several simultaneous sounds. When a voice sample is recorded, it is possible to select a segment and to play it in an isolated form. It is thinkable that if this segment corresponds to a unique sound, it would be immediately recognized when heard, like when listening to long streams of syllables or words [10]. However, this is not true since short frames are difficult to identify. This phenomenon is more evident with consonant sounds. Their identification strongly relies on the transitions with the contiguous vowels. Without the perception of the nearby vowel, it is sometimes difficult to know if it is a human sound or not. Moreover, the transition characteristic of consonants depends on the attached vowel, and it is impossible to be sure that if it is part of the consonant segment or not. This phenomenon is also true for vowels. It is frequent that the perception of a vowel extracted from as sequence of phonemes, becomes an unintelligible sound. Normally, if the adjacent phonemes are not appended, the sound is difficult to comprehend.

In the following section, the division in units of an acoustic signal corresponding to a phonetic stream is presented to study the transitions between phonemes.

3. Representation of the acoustic waveform

Grapheme to allophone transcription is an essential phase in text-to-speech system. It can be described as a function mapping the spelling form of words to a string of phonetic symbols, along with possible diacritic information, such as stress placement, representing the pronunciation of the word. This approach is naturally best suited to languages like Spanish where there is a relatively simple relation between graphemes and allophones. For other languages like English, however, it has generally been recognized that a highly accurate word pronunciation module must contain a pronouncing dictionary that at very least records words whose pronunciation could not be predicted on the basis of general rules. A finite-state transducer was developed to produce the grapheme to allophone translations for Riverplatean Spanish language.

There is a continuous transition from one phoneme toward the next phoneme in natural speech. For example the acoustic signal corresponding to the word *mamá* can be represented by a sequence of quasi-stationary waveform and transition segments between consecutive phonemes, as it is shown in figure 1.

Graphemes		m	a	m	á	
Phonemes	/sil/	/m/	/a/	/m/	/a/	/sil/
Allophones	[ ]	[m1.3]	[a..3]	[m2.1]	[a..1]	[ ]
Acoustic Sig						

Figure 1. Phonetic transcription and acoustic signal of the word mamá

Figure 1 shows the phonetic representation and allophone codification of a word. Shadow segment stands for the stationary waveform or central part of every phoneme and white segment stands for the transition waveform between consecutive phonemes. The allophone codification[1], indicates [m1.3] for the /m/ located in initial position -1-- affected by weak stress ---3, and [m2.1] for the second /m/ located in the onset on the syllable -2-- having main stress ---1. Similarly, [a..3] and [a..1] correspond to /a/ phoneme, both placed in the medium of a word, and with different degree of stress: weak and main, respectively [16].

This article reports results obtained using concatenative synthesis. In this method, to convert a string of allophones into synthetic voice, segments of speech are connected to form the desired speech signal.

The basis of concatenation method is the simply putting two sounds together. The problem is how to combine them, since fluent speech requires smooth transitions between the phonemes. This dilemma is usually overcome using diphones as a basic unit for the

[1] For allophone code, 4 characters are used to represent variant of sounds of the Riverplatean Spanish

concatenation. Diphone is the sound contained in a time window beginning in the stationary portion of one phoneme and ending in the stationary portion of the next one. In this way, this time window will hold all the articulatory evolution of the vocal tract that takes place from the first phoneme to the second one [17].

In Riverplatean Spanish the complete inventory of transitions is 7885, resulting from the permissible combination between phonemes: Silence-to-Consonant, Silence-to-Vowel, Vowel-to-consonant, Consonant-to-Vowel, Consonant-to-Consonant, Vowel-to-Vowel, Consonant-to-Vowel and Vowel-to-Silence. Using this amount of samples produces a speech with good quality, but the number of samples required is high. Therefore, the present approach consists of reconstructing the transition segment, based on the quasi-stationary part of the adjacent phonemes [18][11]. A detailed example of Consonant-to-Vowel group is analyzed: the nasal consonant /m/, /n/ or /ñ/ to vocalic allophones. This study corroborates that the evolution from a sound to the following one, is not linear.

Nasals to vowel combinations are always present at the beginning of the syllable. Nasal phonemes are bilabial /m/, alveolar /n/ and palatal /ñ/. The evidence for opening and closing degree of constriction in vocal allophones is smaller in Spanish than in other languages such as French or Italian [19]. In relation to these languages, a Riverplatean Spanish open sound is really less open and the closed one is partially closed. Current transcriptions for the regional language do not consider these variants of vocal sounds, so in this article they are called "medium" variant of the vowel [7][13][16]. The code used for these variants are strong [a...][e...][o...] and weak [i2..][u2..]. The third character of the code is used only for certain consonants and the fourth place stands for the stress of the syllable. Moreover, the vocal allophone for i or u vowels when the vowel is present forming part of a diphthong is different from the medium /i/ or /u/. These variants of sound are called semiconsonants (i or u being the first vowel in a diphthong) or semivowels (i or u being last vowel in a diphthong) [2]. In the semiconsonant allophones, the articulation movement from one vowel toward another is done in such a way that amplitude follows an increasing pattern, conversely in semivowel allophones the articulation movement produces a decreasing amplitude scheme. This pattern also appears in the diphthongs -iu- and -ui-. The code used for semiconsonant allophones are [i1..] and [u1..] and for semivowels [i3..][u3..]. Semivowel and semiconsonant allophones differ from the medium allophone corresponding to the same vowel. The first group has a non-uniform waveform since changes in the vocal-tract configuration are required during the production of them, but the last group is stationary or uniform. Since this example considers the combination from nasal to vocalic sound, the semivowel variant is never used.

4. Non stationary and quasi stationary segments

The parameters calculated in this model are obtained based on a time-domain approach. In voice signal, the characteristics of 10-30 ms segments of speech are presumably invariant. The parametrization extracts the specific acoustic parameters of units, relying on the redundancy of information that is present in neighbour voice segments, and using them to reconstruct the voice signal. In the voiced fragment of the acoustic signal, one cycle is very similar to the previous and the following one, so it is possible to consider the reconstruction of the whole frame based on a few cycles.

Considering a transition segment that holds n cycles, and calling $f_i(t)$ to the i^{th} cycle, it is thinkable to rebuild the cycles from $f_2(t)$ to $f_{n-1}(t)$ with the only knowledge of $f_1(t)$ and $f_n(t)$. To perform such continuous transformation from $f_1(t)$ to $f_n(t)$, the function $\Gamma_i(t)$ is defined as:

$$\Gamma_i(t) = \alpha_i \cdot f_1(t) + \beta_i \cdot f_n(t), \qquad (1)$$

where $\alpha_1 = 1$, $\beta_1 = 0$, $\alpha_n = 0$ and $\beta_n = 1$. The values of α_i and β_i for i from 2 to n-1 will be chosen to let be $\Gamma_i(t)$ as close as possible to $f_i(t)$. Then, (α_i, β_i) pairs are determined by solving the equation 1, so as to minimize the estimation error between the natural signal $f_i(t)$ and the reconstructed signal $\Gamma_i(t)$.

The assumption that the evolution of α_i and β_i are linear and continuous, for $1 \leq i \leq n$, results in two crossed lines where α_i is continuously decreasing from 1 to 0 and β_i is continuously increasing from 0 to 1. However, this supposition is far from being true for the natural evolution of the voice signal. In next section the

/ma/ transition and the parameters obtained are presented.

5. The /ma/ transition

If the transition /ma/ is observed, an intuitive approach to model is thinkable, that is to say a /m/ phoneme changes upwards to acquire the /a/ phoneme waveform. In figure 2 the acoustic waveform of natural /ma/ transition is showed, extracted from the word *mamá,* containing 12 cycles. Figure 3 (up) shows 5^{th} cycle and 8^{th} cycle, predicted by linear criterion, superposed with the same cycles of the natural signal. The difference of the waveform signals is evident and obviously the sound corresponding to them is perceptible by a listener.

If it is now assumed that the transition evolution is not linear, a recreation of the transition /ma/ can be obtained choosing a better set of values for α_i and β_i for i from 2 to n-1 that makes minimum the error function between the natural and the reconstructed signal. Then, it is possible to reach a good description of the waveform as it is shown in figure 3 (down). Further, this model is notoriously better. In this case, during a large portion of the transition the waveform of the phoneme /m/ is preserved and is abruptly transformed into /a/ phoneme. Probably, this happens when the oral cavity is opened and simultaneously the complete stop of the airflow from the nasal cavity takes place. In linear approach the pairs (α_5 = 0.64, β_5 = 0.36) and (α_8=0.36; β_8 = 0.64) are used for equation 1. However, in non-linear technique the pairs are (α_5 = 1.07; β_5 = 0.0) and (α_8 = 0.02; β_8 = 1.14). This difference qualifies the naturalness of the signal when the natural and synthesized signals are compared.

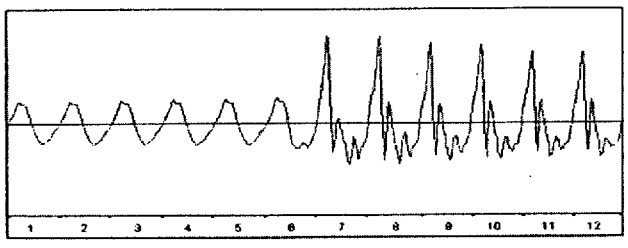
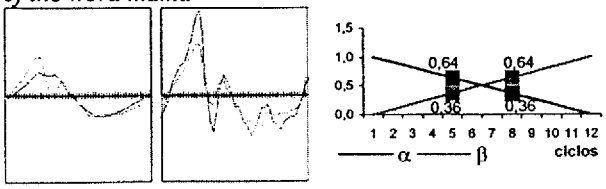


Figure 2. /ma/ transition extracted from the first syllable of the word mamá



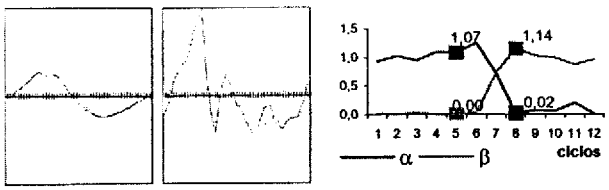


Figure 3. 5^{th} and 8^{th} cycles reconstructed with the linear method (up) and non linear method (down)

Figure 4, shows the reconstructed signal using the non linear method, which is very similar to natural signal, showed in figure 2.

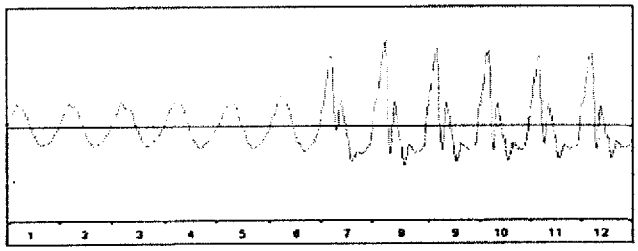


Figure 4. /ma/ reconstructed transition

6. /m/ toward medium vowels

Similarly to the /ma/ transition, the evolution from /m/ to another vowel was analyzed. Word samples were uttered at normal speaking rate to study the transitions between phonemes. In all cases, the evolution of the acoustic signal from /m/ to vowel is a fast transformation where the acoustic waveform of the first cycle changes progressively to the last one. In the special case of /a/ and /o/ vowels an increasing amplitude is observed, followed by a sudden transformation into the vowel waveform. Inversely, in the evolution toward /i/, /e/, /u/ vowels, the /m/ cycle remains invariable up to the middle of the transition, followed by a smoothed evolution with characteristics of both phonemes ending with a completely vowel waveform. Figure 5, shows [m...a...][m...o...] and [m...e...][m...i2..] and [m...u2..] average curves obtained from different samples of captured voice signal. In the codification, [m...] stands for all variants of /m/ phoneme; and /a/, /e/, /o/ belong to the strong medium vowels, while [i2..] and [u2..] represent the medium weak vowels that are not forming diphthong or triphthong. It is clear that the actual evolution is very different than the predicted by the linear model.

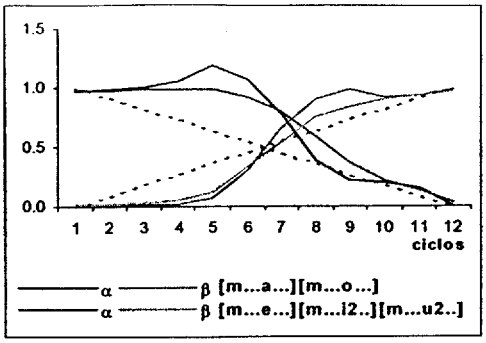


Figure 5. α y β evolutions for /m/ toward medium vowels.

7. /n/ toward medium vowels

The transitions from /n/ toward medium vowels are very similar one to another. As a consequence, the average pattern is identical for the complete [n...a...][n...o...][n...e...][n...i2..] [n...u2..] set. Moreover, it can be observed that /n/ toward vowels look like [m...e...] [m...i2..] [m...u2..] depicted in figure 5. Again, the difference between natural speech and reconstructed signal is smaller using non-linear coefficients.

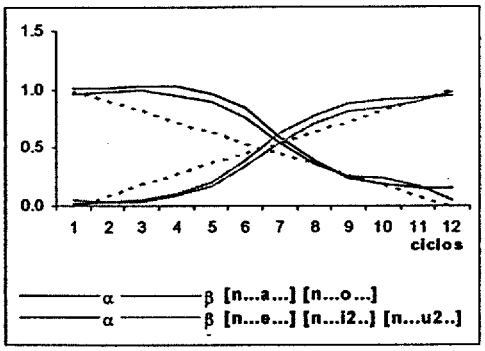


Figure 6. α and β evolution for /n/ toward medium vowels

8. /ñ/ toward medium vowels

/ñ/ toward vowel transitions differ from the previous consonants in such a way that the complete duration of the transition is twice the time required in previously described articulations. For example, for a pitch of 220 Hz. transition of figure 5 contains 12 cycles, while transitions of figure 7 contain 24 cycles. As occurs with /m/ to vowel transitions, for [ñ...a...] and [ñ...o...] the first cycle increases its amplitude to reach the middle of the transition, followed by a mixed waveform with certain overlap between two phonemes, and finally the signal acquire the complete vowel waveform. [ñ...e...][ñ...i2..][ñ...u2..] are smooth transitions, with gradual decreasing α and increasing β parameters from the beginning towards the middle. Then it is a stable waveform arriving at end with complete vowel form. Similarly to other nasal to medium vowel transitions, a perceptible difference exists when synthetized voice using the linear and non linear model are heard.

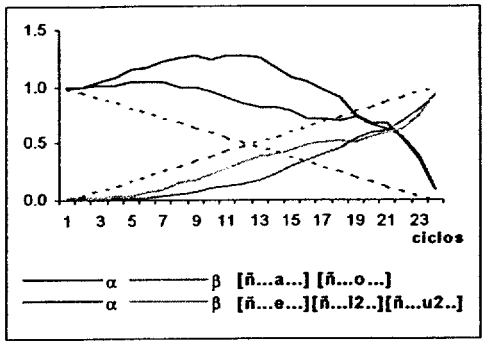


Figure 7. α and β evolution for /ñ/ toward medium vowels

9. Nasal to semiconsonant vowels

The duration of nasal consonants to semiconsonant vowel allophone transitions are longer than the transition of the same nasal consonant to medium vowel allophones. This evolution like /i//e//u/ group of previous transitions is a weak transformation. The natural segmental duration is identical to /ñ/ to medium vowels. Figure 8 shows /m/ and /n/ followed by semiconsonant vowels, overlapped with a /ñi//ñu//ñe/ curve of figure 7. The /ñ/ consonant was defined in [2] as a unique sound articulated in an interval of time as do the /m/ and /n/ consonants. But this result confirms that /ñ/ sound is produced by two consecutive sounds: nasal /n/ followed by palatal /y/. As can be seen in figure 8, α and β evolutions for /ñ/ toward medium vowel group are similar to /n/ toward semiconsonant vowel group. As expected, the results confirmed that /ñ/ is formed by two sounds [8][18]. Similarly, /m/ toward semiconsonant vowels presents similar data and it is included in an average curve.

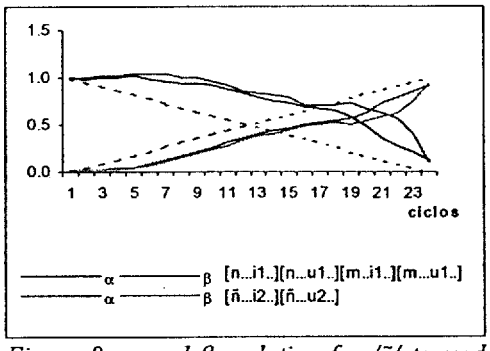


Figure 8. α and β evolution for /ñ/ toward medium /i//u/ and /n/,/m/ transitions toward semiconsonant vowels

Conclusions

This paper enforces the conclusions of [11]

in the sense that:

- The use of deep phonetic knowledge improves the quality of synthesized speech in a remarkable way.
- Time-domain approaches may produce good quality in the synthesized speech.
- The combinatory explosion in the number of transitions can be neatly handled without an important loss of information.

In the specific case of nasal to vocal transition it can be seen that the transition from nasal to /a/ and /o/ vowels have a sudden change in the waveform whereas transition to /e/, /i/ and /u/ vowels exist a certain mix between the two sounds. In the first case the nasal increases in amplitude and becomes vocal in a sort of plosive behavior while in the second one there are gradual "vocalization" of the nasal sound.

Bibliography

[1] Deller J., J. Proakis y J. Smith.(1993). *Discrete time processing of speech signal.* Mac.Pub.,NY.

[2] Navarro Tomás, T.(1957). *Manual de la pronunciación española,* Artes Fénix, Madrid.

[3] Pietra, R., A. Fritsch (1996). *Digital signal processing come of ages.* IEEE Comp., Vol. 33, pp.70-74.

[4] Manaris B., B. Slator (1996). *Interactive natural language processing.* IEEE Comp., Vol. 29, pp. 28-32.

[5] Martin, P., et al.(1996). *SpeechActs: A spoken-Language Framework.* IEEE Comp., Vol. 29, pp. 33-40

[6] Wactlar H. et al.(1996). *Intelligent Access to digital video.* IEEE Comp., Vol. 29, pp 28-32.

[7] Massone M., M. Borzone(1985). *Principios de transcripción fonética,* Ed. Macchi

[8] Luk R., R. Damper (1998), *Computational complexity of a fast Viterbi decoding algorithm for stochastic letter-phoneme transduction,* IEEE trans. Speech and audio proc. Vol 6, 3, pp. 217-225.

[9] Levinson, S., et al.(1993). *Speech synthesis in telecomm.* IEEE Com., Vol. 31, pp.46-53.

[10] D'Introno, F, et al (1995). *Fonética y fonología actual del español,* Ed. Cátedra, Madrid.

[11] Barbuzza, R.; J. Doorn (1997) *A time domain approach to model transition between phonemes,* SIP'97 International Conference on Signal and Image Processing", San Francisco, EEUU.

[12] Navarro Tomás, T. (1946). *Estudios de la fonología española.* Syracuse Univ. Press, NY.

[13] Gili Gaya, S. (1988). *Elementos de fonética general.* Cátedra Ed. , Madrid.

[14] Vidal, B. (1964). *El español de la Argentina.* Cons. Nac. Ed., Buenos Aires.

[15] Quilis A. (1981). *Fonética acústica de la lengua española.* Gredos Ed., Madrid.

[16] Barbuzza, R.; J. Doorn (1996). *The phoneme taxonomy for Riverplatean Spanish.* Int. Rep., Tandil.

[17] Portele, T.; F. Höfer and W. Hess (1997) *A mixed inventory Structure for German Concatenative Synthesis, pp.* 263-277, Progress in Speech Synthesis, *Springer-Verlag, New York.*

[18] Barbuzza, R.; J. Doorn (1997) *Estudio de las transiciones entre fonemas rioplatenses para la síntesis de voz utilizando algoritmos temporales.* VII RPIC, San Juan, Argentina

[19] Ferri, G.; P. Pierrucci and D. Sanzone (1997) *A complete linguistic analysis for an Italian Text-to-Speech System,* pp. 123-138, Progress in Speech Synthesis, *Springer-Verlag, New York.*