The Actual Back-End Tendency in Speech Synthesizer

Sergio Suárez Guerra[1]; Ismael Díaz Rangel[1]

[1] Center for Computing Research, National Polytechnic Institute,
Juan de Dios Batiz esq Miguel Othon de Mendizabal s/n, P.O. 07038, Mexico
ssuarez@cic.ipn.mx, idra06@sagitario.cic.ipn.mx

Abstract. At the moment the models of greater use used for the implementation of the synthesis of voice in their stage back-end are: the articulator model, that still is in an immature stage, but is outlined in the future like the model with greater potential; the model by formants, that given its limitation an artificial voice but with great flexibility and intelligibility produces, and the model by concatenation, that enjoys a great popularity nowadays, nevertheless, require of greater computational cost and it practically lacks synthesis flexibility. It is possible to mention that great challenges of the voice synthesis are in their stage front-end, since the determination of how to pronounce the numbers, abbreviations, acronyms, names, etc. they are possible to be turned problems of difficult solution; as well as the correct analysis of prosody. It writes or it sticks the text here to translate.

Keywords: Synthesis, back-end, Front-end, Formants, Glottal Excitation.

1 Introduction

The speech is the main form of communication between the people. A synthesizer of voice is a device able to create of artificial way articulated voice [1]. A type of synthesizer of voice is the call Text Speech System (TTS). The TTS has the capacity to read any text aloud, or introduced by a user, or generated by a system of OCR (Optical Character to Recognizer), or even pertaining when coming out of a system of consultation with data with the results of a request on the part of a user.

The fundamental difference with other systems talking (like it could be a tape reproducer), is that our interest is centered in the capacity to reproduce new phrases or texts automatically, which eliminates of the process the idea that a recording of such mediates. Even so, it can that we need to refine plus our initial definition: systems that, for example, simply concatenate prerecorded words or phrases (typically calls systems of vocal answer), are only applicable when the vocabulary of the application very is limited, of the order of few hundreds of words. In the context of systems TTS,

Computer Engineering.
Research in Computing Science 30, 2007, pp. 119 - 128

it is almost impossible to raise the recording of all the words of the language, so that he is more reasonable to define them as automatic production systems of speech, through a process of transcription of graphemes to phonemes. It is capacity to read must be distinguished to the language and geographic location of the user; since the diction is different not only between languages, but that even in he himself language (it does not sound identical the Spanish in México that the Spanish in Cuba). In this sense, a good margin in México exists to work in voice synthesis.

A system TTS consists basically of two modules (figure 1), the module of text analysis to which we will call "front-end" and the module of speech synthesis to which we will call "back-end". In the analysis module, the following operations are made: normalization of text and abbreviations, syntactic analysis, semantic analysis, syllabication, accentuation and converter grapheme-allophone (the most elementary unit of the sound). Once analyzed the text, it is had a set of textual parameters that are necessary in the synthesis process; for example, Mr. by Mister is translated and he is indicated that the second syllable must go marked, eliminate the letters that are not pronounced and the possible places of prosodic changes are marked.

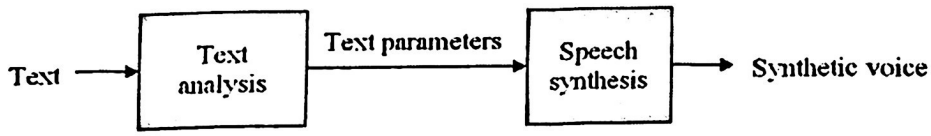

Fig. 1 Simplify diagram of TTS system

In the implementation of a system of voice synthesis first that we must consider it is the model for the stage back-end. At the moment the synthesis is dominated by three systems, that depending on the requirements of the application we must show preference for one or another one. In order to sustain this idea, in the following points we are going to describe briefly to each one of them.

2 Articulator Model

The articulator synthesizers provide synthetic voice of high quality, but its disadvantage is that the parameters are very difficult to obtain and to control them automatically.

The voice wave is the answer to the system of filters of vocal tract for one or more sources of sound. This affirmation, expressed in terminology of acoustics and electrical engineering, implies that the voice waves have unique specifications in source terms and 2 filtrate characteristics [2].

The articulator synthesis determines the characteristics of filter of vocal tract by means of the description of the geometry of vocal tract (like the size of the oral cavity, the trachea and the position of the language, among other variables) and places the sonorous sources within this geometry. These factors are related to each other to produce a voice that is resembled in the greater measurement of the possible thing the human voice. The articulator synthesis applies harmonic signals to the sonorous signal and establishes an analogy between parameters related to the articulator organs, its movements and characteristics.

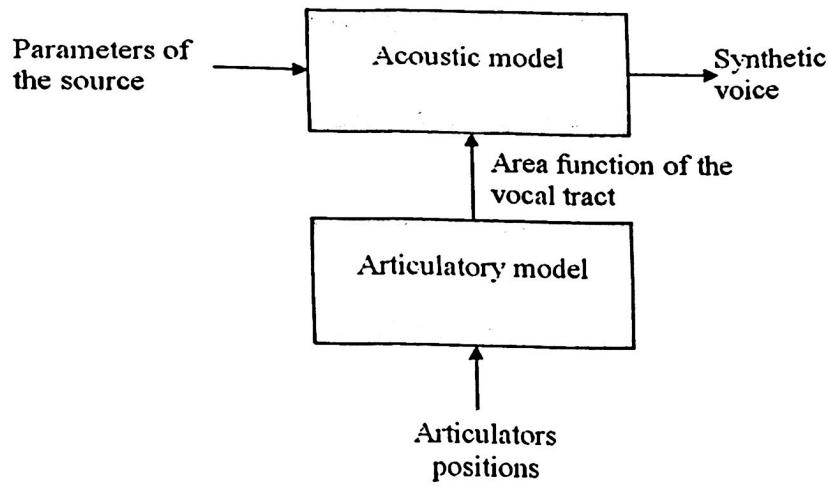

Fig. 2 Basic structure in the articulator synthesis

Depending on the synthesizer, the geometry of vocal tract can be described in one, two or three dimensions. The one-dimensional model represents vocal tract by means of its function of area directly. The area function describes as the area of the representative one of the tubes of the vocal zone varies between the glottis and the opening of the mouth. Assuming a one-dimensional propagation in tract, the function of the area contains all the information to specify the characteristics of the filter. Therefore, with respect to the acoustic simulation, the models bi-dimensional and three-dimensional of the vocal zone also are transformed finally into a one-dimensional function of the area. The advantage of the multidimensional models is that the form and position of the articulator's can be specified of a very direct way.

The artificial articulator's of these models generally are controlled by means of a small set of articulator parameters. The variation of these parameters in the time allows that the function of area of vocal tract changes during a pronunciation. An acoustic model is used to calculate the wave of voice from the sequence of the area functions and its corresponding sources of sound.

In summary, an articulator synthesizer needs the following thing at least:

- A geometric description of vocal tract implemented in a set of articulator's parameters.

- A mechanism of parameters of control during each word.

- A model for the acoustic simulation, including the generation of the sound sources.

The entrance for the acoustic simulation is generally a constant segment of the function of area, corresponding to a zone of vocal tract integrated by several cylindrical sections of a tube according to the illustrated thing in figure 3. The figure shows how the vocal zone is excited by means of a glottal function of the speed of the volume (acoustic source) and radiates a sound wave of the pressure in the nasal orifices and the opening of the mouth.

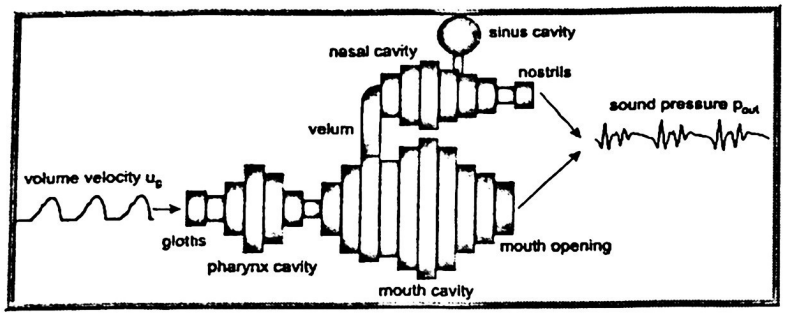

Fig. 3 Three-dimensional model of the vocal tract [4]

The main problem of the articulator models is the enormous amount of internal parameters of control that need and make difficult to the coordination and derivation of the parameters of control available to the entrance of the synthesizer; and on the other hand, the great amount of information that is needed to obtain analyzing (in a three-dimensional space) the position and the movement of the articulator organs of a person that speaks normally, thing very difficult to measure in these conditions.

The idea that bases to this synthesizer makes us think that it is adapted to totally reach the objectives of the back-end; nevertheless, not yet it has been managed to model correctly to each one of the formants, thus we thought that although at the moment it is not the system that gives better results, the constant improvement of his characteristics will take it to be the dominant system.

3 Concatenation Model

The principle of the concatenation is to produce voice by means of the connection of pre-recorded sounds, with which a sound of intelligible and natural voice is obtained. Nevertheless, the synthesizers by concatenation are limited by only talking one, which must record all the units of speech that the system is going to use. The size of these units is based on the naturalness that is tried to reach, this is, to obtain a greater naturalness is preferable to use great units. The type of unit to concatenate is a critical parameter to obtain a good quality of the synthesized voice: it is necessary to arrive at a balance between the inter-segmental quality possible (to greater length of the segments, less points of concatenation and therefore greater quality) and the amount from memory necessary to store the prerecorded units. The recorded pieces do not have to be words by two fundamental reasons. In the first place, the pronunciation of a phrase is very different from the one from a sequence of recited words separately, since in a phrase the words last one more shorter than when they are isolated and the rate, intonation and accentuation, that depend on semantic and syntactic factors, they are totally unnatural when recorded words are concatenated separately. A second problem is the innumerable existing words in a language, if we consider for example the own names, as well as the formation of words by means of suffixes, area codes and conjugations. The syllable is an interesting unit very linguistically, but there are a great number of them. Another proven unit is the phoneme, whose number is smaller of 30, but the turn out to concatenate phonemes is not satisfactory due to co-articulators effects between adjacent phonemes that produce changes of the acoustic manifestations of a phoneme depending on the context. The co-articulators effects tend to diminish themselves in center acoustic of a phoneme, which took to propose difonema, the voice piece that in the middle of goes from half of a phoneme the following phoneme, like the most satisfactory unit for the concatenation. In Spanish 900 can be considered about. In addition it can be necessary to introduce allophones different to make the distinction between the marked and atonics vowels or the inclusion of triphone, that are an extension to groupings of three phonemes when the co joint effects are so great that the segmentation in difonemas is not possible.

Although this system is the one that more computational resources it requires has become more and more popular, partly because its implementation is not so complicated (but very laborious), and to that the systems of computer every time are quicker and accessible, idem the memories. Nevertheless the parameters of the voice as the fundamental tone cannot be modified; for that reason, if we took like reference to [1] to define voice synthesis, we noticed that this system does not make true synthesis of voice.

Let us think that this system is a good option for a system that require good intelligibility, naturalness, a limited vocabulary and that is sufficient one or two speakers.

4 Formants Model

These synthesizers are based on the acoustic theory of voice production, which in its simpler form, says that it is possible to see the voice as the result of the excitation of a linear filter with one or more sonorous sources.

A simplified approach in the mechanism of the speech production in the acoustic dominion was proposed at the end of the 50 decade and was called "to source-filter model" [3]. In this model, the production system of voice is divided in two:

- Source of excitation.
- Resonance tract.

These two parts assume one "interaction" and one connection not to linear. The formants are the resonances of vocal tract. A synthesizer by formants reproduces the structure of formants of vocal tract.

In the 60 decade, appeared the first done discreet synthesizers with formants. The resonant ones were implemented in configurations of series (in cascade) or parallel. Flanagan (1957) concluded that the form series is better to reproduce sonorous sounds, and the no nasal ones; whereas the structure in parallel is better for the nasal sounds, and the no sonorous.

In 1980 appears the combined parallel/series system of Klatt. This configuration, made to improve the capacity to reproduce nasal and no sonorous sounds of the system. This type of synthesizers has an ample diffusion but the quality of the synthesized voice is smaller [Montero, 2002]. Although with the model parallel/series, using an appropriate specification of the synthesis variables and a correct configuration, it is possible to obtain a synthesized voice of high quality.

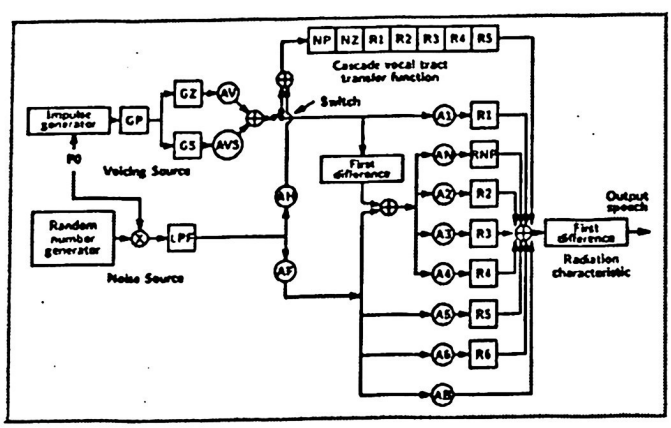

Fig. 4 Parallel/series model formants synthesizer [5]

The factor most important to obtain a synthesis of voice of high quality is the extraction of the parameters of synthesis, applying adapted procedures of analysis to a voice signal. Most of these procedures they use an acoustic signal of voice like source to determine the formants [Alku, 1992; Childers and Lee, 1991; Klatt and Klatt, 1990; Markel and Gray, 1974; McCandless, 1974; Trim off lower branches of, 1971]. Another factor important to obtain a synthesis of high quality, is the design of the excitation source [Childers, 1995; Childers and Ahn, 1995; Childers and Hu, 1994; Childers and Lee, 1991; Childers and Wu, 1990].

Synthesis parameters extraction

The parameters of our interest are bandwidth and the frequency of the formants, which are the picks in the surrounding one of the spectral of the voice signal which they represent the frequencies of resonance of vocal tract [6]. The formants frequencies can vary from a person to another one because all we have a constitution of vocal tract only; but in general form they are within a well-known rank.

Several methods for the calculation of formants of a voice signal exist; the methods that were used to extracts the formants from the spectrum of the lineal predictions coefficients (LPCs) of a voice sample. The steps to obtain the LPCs are:

1. Segment the useful signal, without silences at the beginning and end, in units of 25 or 20 ms (I segments).

2. To apply a window of Hamming to each segment of signal.

3. To calculate the amount of **p** values of the coefficients of autocorrelation for each segment.

4. To calculate the $\mathbf{a_p}$ coefficients LPC for each segment.

5. To obtain the LPCs average.

Once obtained the LPCs, graphical its spectrum in frequency, and took the frequency from the picks of the surrounding one like the formants, also of this graph we can calculate the bandwidth of each formant. In figure 5 it is showed to the frequency response of the LPCs of a recording without controlled conditions of the vowel 'a', monaural and with a frequency of sampling of 16000 Hertz. In figure 6 we showed an approach of the first pick of the spectral shown in figure 5, and of which we can accurately determine the frequency of the first formant.

In figures 5 and 6 the vertical axis corresponds to the amplitude, and horizontal axis corresponds to the frequency in Hertz.

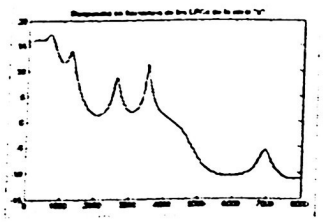

Fig. 5 Frequency spectrum response of LPC Vowel 'a'

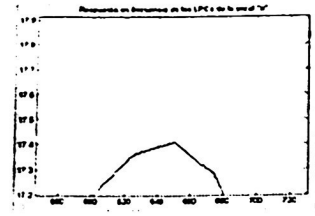

Fig. 6 First format pick of the frequency spectrum frequency spectrum of vowel 'a'

This way it is been able to determine bandwidth formants and for each one of the units of voice of sonorous type that our system requires. The following table presents the data collected for the five vowels of the Spanish; in her we both showed to first respective formants and their bandwidths:

Table 1. Vowels formats

	F1 / W1 [Hz]	F2 / W2 [Hz]	F3 / W3 [Hz]	F4 / W4 [Hz]	F5 / W5[Hz]
A	650 / 100	1200 / 150	2600 / 250	3500 / 300	4500 / 200
E	400 / 200	2000 / 150	2700 / 350	3650 / 150	3850 / 200
I	250 / 50	2200 / 150	3000 / 300	3500 / 200	4000 / 200
O	400 / 100	800 / 200	2000 / 200	3200 / 200	4000 / 200
U	400 / 100	800 / 200	2500 / 200	3500 / 200	4000 / 200

In the tests it is observed that the quality of the synthesis improves when increasing the amount of formants; but this is only truth with the first formants, after the first three formants the improvement is every smaller time; for that reason one says that five formants are an suitable amount to approach us the Maxima quality that this type of synthesizer can offer.

Excitation source

The other point of relevance is the source of excitation for the system of synthesis by formants, was proven with different models: Delta, Rosenberg and Liljencrants-Fant (LF). Being this last one from that better perceivable result we obtained. Model LF tries to equal the waveform produced in vocal tract (figure 7), and changing some of his parameters it is possible to obtain different pitch characteristics in the synthesized signal.

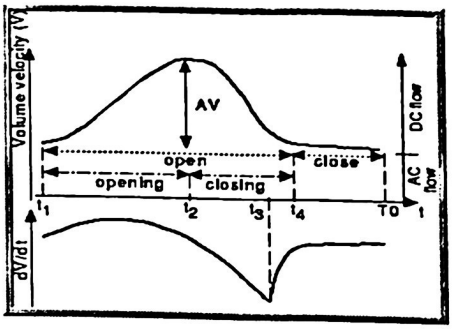

Fig. 7 Glottal Speech volumetric wave and its derivate [7]

In figure 7 we can observe the following parameters:

- t1 initiates the opening of the vocal cords and begins to flow the air.
- t2 moment of maxima glottal opening (AV) and maximum air flow.
- t3 beginning of the closing of the glottal cavity and maximum change of the glottal flow (harsh- introduces components of greater frequency).
- t4 the glottal cavity is completely closed and ideally there is air flow no.
- To duration of a complete period.

The equations that model LF proposes to model the derived one from the waveform is:

$$g(t) = E_0 e^{\alpha t} \sin(\omega_g t) \qquad t_1 \le t \le t_3$$

$$g(t) = \frac{Ee}{\varepsilon t_a}[e^{-\varepsilon(t-t_3)} - e^{(T_0-t_3)}] \qquad t_3 \le t \le t_4 \le T_0$$

Where:

- E_e is the amplitude of the excitation.
- tα constant of the exponential curve that determines form of the curve between T3 and t4.
- ω_g establishes the duration of the phase of opening.
- E_0 is an amplitude factor.
- ε is coefficient that it increases of exponential way to the sinusoid.

In order to obtain the waveform that will be used in the synthesizer only it is necessary to integrate the equation of the LF model.

In figure 8 we can observe the vowel 'a' real and in figure 9 we observed the vowel 'a' synthesized. The graphs show amplitude against sample. The source of used excitation was provided by model LF.

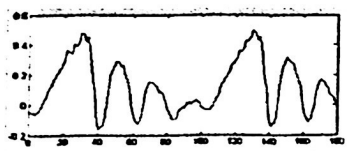

Fig. 8 Vocal 'a' real

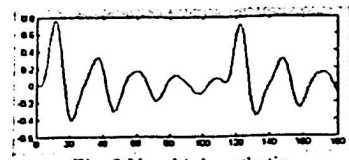

Fig. 9 Vocal 'a' synthetic

5 Conclusions and future works

In intelligibility approach the voice synthesis already reach an acceptable level; nevertheless, the naturalness implies so many dynamic changes that at the moment it is continued investigating the most suitable method to reach an acceptable level; in that sense, the articulator synthesis seems to be the key, but lack work to do much. The synthesis would concatenate, although it is not synthesis in a strict sense, at the moment widely is used, due to the superiority in naturalness that obtains, but its lack of dynamism and discharge demand of resources prevents its universality. On the other hand, the model by formants, although produces an artificial voice, doing a configuration adapted of its modules, a correct calculation of the formants and providing a source to him of excitation next to which generates vocal tract, we can obtain satisfactory results, with the advantage to be able to change parameters in the speech synthesized.

This word is support by National Polytechnic Institute (IPN) of Mexico in de project 20070331.

References

1. Donald G. Childers. "Speech Processing and Synthesis Toolboxes", Jonh Wiley & Sons, Inc. (2000)

2. Donald G. Childers. "Speech Processing and Synthesis Toolboxes", Jonh Wiley & Sons, Inc. (2000)

3. Gunnar Fant. "Acoustic Theory of Speech Production", Mouton, The Hague. (1960)

4. http://wwwicg.informatik.uni-rostock.de/~piet/speak_main.html

5. J. Acoust. "Software for a cascade/parallel formant synthesizer", Soc. Amer., vol. 67, pp. 971–995, 1980.

6. Alfredo Mantilla. Tesis doctoral "Análisis, reconocimiento y síntesis de voz esofágica". Escuela Superior de Ingeniería Mecánica y Eléctrica, Culhuacan DF México. (2007)

7. http://www.ims.uni-stuttgart.de/phonetik/EGG/page13.htm