GENERATING VOCAL SIGNAL SOURCES

Cătălin - Iulian CHIVU
Transilvania University of Brasov, Romania

Abstract. During last decade speech technology becomes more and more used in fields like: telecommunication and automated dialog systems. The speech technology is developed based on the knowledge of signal processing and comprises two principal areas: automatic speech recognition and speech synthesis. As in any other technology, in speech technology the high performances are obtained based on the quality of the implemented algorithms (method that is used and efficiency). Present paper tries to cover the main methods used to generate a synthetic fundamental vocal signal and to underline the appropriate method for different type of applications.

Keywords: vocal, synthesis, speech, spectrum, spectrogram

1. Introduction

At present, there are in the word, already created, a large variety of vocal synthesisers. For some of these synthesisers the principle of operation is synthesis-by-rule.

Because all existing synthesisers have, as target, one or more languages, it becomes necessary to create a universal synthesiser, which should have the following characteristics: the ability to imitate any kind of voice, to interpret any vocal inflection, to read text in any language and to be able even to play music. Such a vocal synthesiser should have, as fundamental requirement, a high quality of generated vocal signal. A first condition to be able to generate a high quality vocal signal, using synthesis-by-rule, is to obtain a primary vocal signal which is flexible, with clean spectrum and using low computing power.

Based on these aspects, the present paper represents a study case on some methods used to generate vocal signal sources, based on quality of the obtained signal and on necessary computing power.

2. General aspects

The automatic speech synthesis should be defined as the integrated technology simulating the human process that generates speech, ranging from a simple system which comprises a minimum amount of signal processing, to systems which transform symbolic or linguistic representation of utterances in acoustic waveforms [1, 2].

Synthesis-by-rule is the most flexible option that it can get. It allows, on the one hand, obtaining the highest performances, but on the other hand being the most sensitive and hard to attend. Synthesis-by-rule allows adapting the vocal synthesiser to any language or voice or even to some situations when the computer is the vocal singer.

The present paper is focused on the two alternative signal sources that are the input of the time-varying digital filter.

The main methods used to generate de primary vocal signal should be analysed both in terms of generated signal quality and desired computing power.

To be able to understand better those methods it is useful to briefly present the basic structure of a vocal synthesiser based on rule (figure 1 – [1, 2]). The most important functional component of the synthesiser is the group of time-varying digital filters. This group may have one or many filters parallel or serial coupled, the parameters of these filters are changed during time based on the shape parameters of the vocal signal.

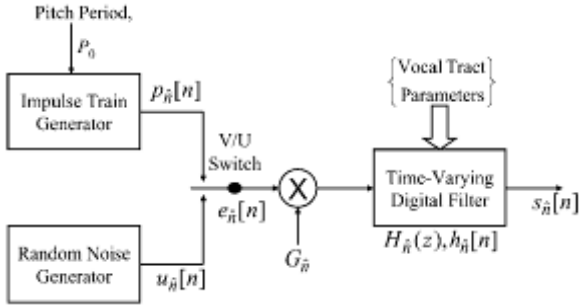

Figure 1. Basic structure of rule based vocal synthesiser

Sounds underlying the spoken language are dived in two main categories: vocal sounds

(vowels) and non-vocal sounds (consonants). These two categories determine the existence of the two alternative signal sources. Vowels are sounds generated by the beating (vibration) of the vocal cords, while consonants do not use the vibration of vocal cords, being the sounds (noises) generated by the pressurised air that flows through small openings.

The switch from figure 1 will choose the proper signal source, depending on the type of sounds needed as output of synthesiser on one moment. Thus, for vowels will have impulse train that simulates the vocal cords beatings and for consonants random noise or, so called, white noise.

In the following there are presented the two type of signal sources and the methods used to generate them.

3. Generating white noise

White noise is a random signal (or process) with a flat power spectral density. In other words, the signal contains equal power within a fixed bandwidth at any centre frequency. White noise draws its name from white light in which the power spectral density of the light is distributed over the visible band in such a way that the eye's three colour receptors (cones) are approximately equally stimulated [5].

Spectrum and spectrogram of a white noise are presented in figure 2 and figure 3, respectively.

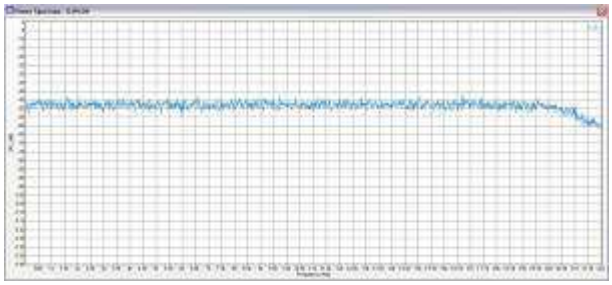

Figure 2. Spectrum of white noise

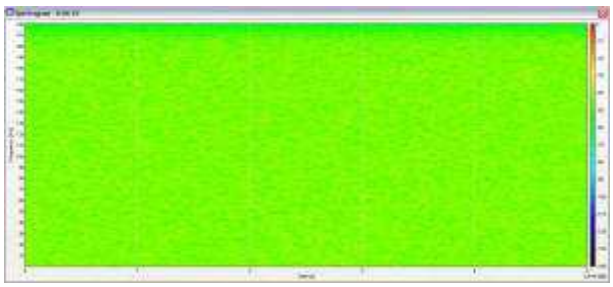

Figure 3. Spectrogram of white noise

The white noise generator can be implemented into a computing system by using a random number generator with uniform distribution. This random number generator is implemented based on an already developed algorithm that is highly efficient.

4. Generating impulse train

To produce vocal sounds (vowels) it should be generated an impulses train. This signal is then passed through a signal filter sets so that the output signal has the same spectrum as the desired signal spectrum. Frequency, amplitude and the shape of the impulses have direct influence upon the output signal spectrum.

As is known, an impulse periodic signal, with F_0 fundamental frequency, can be decomposed as a sum of sinusoids (Fourier series decomposition), F_1, F_2, ..., F_n, where n tends to infinity. Besides frequency parameter of each sinusoid that compose this series (which is simply a multiple of fundamental frequency F_0), each sinusoids has an amplitude coefficient and a phase parameter. When the impulse train passes through the filter the amplitude coefficient and the phase parameter will change for each frequency, F_1, F_2, ..., F_n that compose the series.

Since the human ear cannot distinguish the phase parameters of a signal composed by summing sinusoids series, these parameters will be sent in the background, retaining their importance on the implementation of time-varying digital filters.

Based on this supposition there can be deduced the two distinct methods of generating an impulse train, signal that will substitute the human vocal cords beatings:

1) first method (most intuitive) is to generate a series of samples of zero, through which will occur from place to place, with a frequency equal to the F_0 fundamental frequency, samples that will have as amplitude the maximum amount allowed (highest value that the system allows).

2) second method involves summing a finite number, n, of sinusoids F_0, F_1, F_2, ..., F_n, all having the same amplitude coefficient, and the frequencies $F_1 = 2F_0$, $F_2 = 3F_0$, $F_3 = 4F_0$ etc., where F_0 is the fundamental frequency.

Both methods have advantages and disadvantages:

1.a) The most important advantage of the first method, that which generates an impulse train with F_0 frequency, is the low computing power needed. On the other hand, this method has a major drawback that for a computing system (present situation) is working with samples (discrete time mode), which reduces the number of possible set of

values for F_0 to a finite set of values instead of infinite set of values. Because $F_0 = 1/T$, where T is the fundamental period of generated signal, T may have value only as multiple of the sampling period, T_{es}, (time between two successive samples). Even if generates a clear spectrum signal, as in figure 4, this method cannot be used, in this form, to generate high quality vocal signal, because this type of signal requires a change in F_0 fundamental frequency as a continuous characteristic. However, for an unassuming high quality synthetic voice, this method, in its raw form, is a good compromise between quality and computing power.

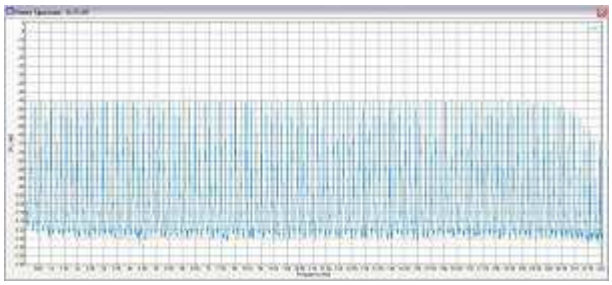

Figure 4. Spectrogram of an impulse train with F_0 = 196 Hz = 44100 Hz/225 frequency

1.b) However, there is a way to improve the quality of this method by computing the phase error and using this value to correct the signal fundamental period T. Phase error is the difference between needed fundamental frequency, F_0, and obtained quantized fundamental frequency, F_{0c}. In this way it will be obtained an impulses train signal with average fundamental frequency F_0 that can change as a continuous characteristic. However, as a tribute to this gain, there is a serious deterioration of the signal spectrum. Maintaining an average fundamental frequency, F_0, different of allowed quantized values, leads to appearance, over the useful signal, of a very evident spectral harmonic series that significantly alters the vocal signal. The improved version determines an insignificant increase of computing power but eliminates the disadvantage given by the discontinuous nature of fundamental frequency obtained using the previous variant. This is obtained with the price of signal spectrum degradation (figures 5a and 5b). This method is still not sufficiently developed to be used in generating high quality vocal signal.

1.c) A third method is based on computing a sampling series that represents the values of sinc(x), or sin(x)/x function. Summing this sampling series with an average fundamental frequency, F_0, and after applying the phase correction, it will be generated a vocal signal with clear spectrum and the possibility of continuous variation of fundamental frequency (figure 6). In conclusion, to obtain a high quality, flexible and clean vocal signal it has to be increase the computing power.

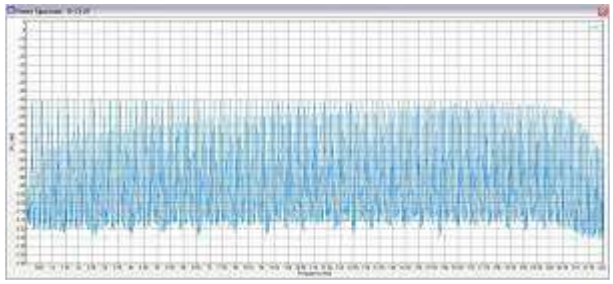

Figure 5a. Spectrogram of an impulse train with average frequency F_0 = 400 Hz = 44100 Hz/220.5

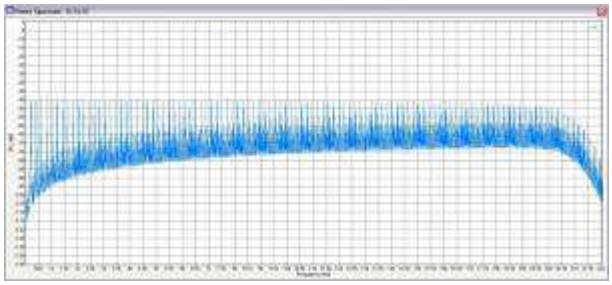

Figure 5b. Spectrogram of an impulse train with average frequency F_0 = 203 Hz = 44100 Hz/217.241379310....

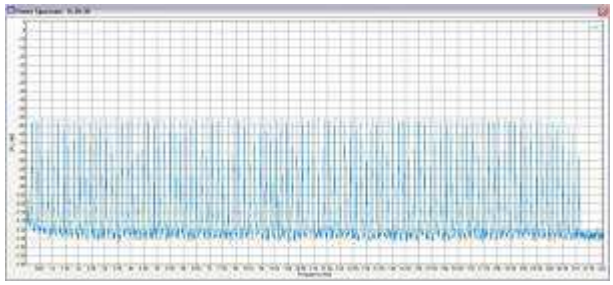

Figure 6. Spectrogram of a "sinc" impulse train with average frequency F_0 = 203 Hz = 44100 Hz/217.241379310....

2) Even if it is obvious that the second method will determine a high quality vocal signal, with clean spectrum (figure 7), will have the main disadvantage of needing a high computing power. The method allows computing an impulse vocal signal that can have any value of fundamental frequency, F_0. This method involves summing n, finite number of sinusoids, F_0, F_1, F_2, ..., F_n, all having the same amplitude coefficient and the frequencies $F_1 = 2F_0$, $F_2 = 3F_0$, $F_3 = 4F_0$, and so on, where F_0 is the fundamental frequency. The significant increase of needed computing power is justified by the fact that the sum of all sinusoids must be done for each sample separately.

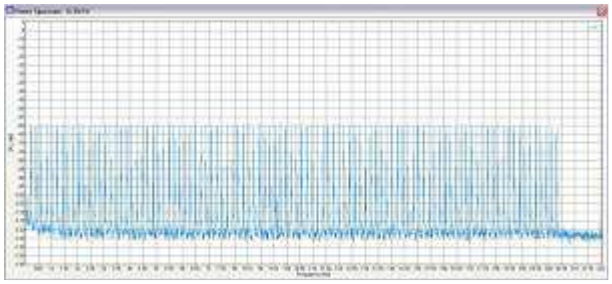

Figure 7. Spectrogram of an impulse train obtained by summing sinusoids with frequencies multiple of F_0 = 203 Hz = 44100 Hz/217.241379310....

In fact, not computing the sum will determine the increase of computing power, but the computing of the n sinusoids. Thus, the number of sinusoids from Fourier series, which can be computed in real time, depends directly proportional on the available computing power.

It is known that the number of sinusoids in the Fourier series directly influence the accuracy of recalculation of the original signal. Thus it becomes necessary to accept a compromise between obtained signal quality and the needed computing power. However, the fact that the signal filtering can be obtained before summing the n sinusoids, by changing the coefficient of each Fourier series sinusoids, is the most important advantage of the second method. This means, on one hand eliminating the time-varying digital filter and, on the other hand, a significant gain in terms of needed computer power.

In terms of computing power, choosing a more expensive model to obtain the row vocal signal, determines, in the end, a saving of global computing power needed. However, besides all the advantages this method generates a rigid vocal signal. This rigidity of the signal derived from the difficulty to change the fundamental frequency of the vocal signal as a continuous function, function with a non-continuous first order derivative.

Even if the value of F_0 fundamental frequency is not constrained to a limited variety of quantised values, trying to change it can determine uncontrolled variations of it. This phenomenon occurs when the first order derivative of the parameter F_0 variation function, passes through a point of discontinuity.

This generates constrain of using, to control the fundamental frequency F_0, a continuous differentiable function with continuous first order derivative.

5. Conclusion

The above analysis allows the user to make the appropriate choice for a particular case when a synthesis-by-rule synthesiser is used. It can be chosen the most convenient method of vocal primary signal synthesis based on the available computing power, on the desired quality performances and on the flexibility of vocal synthesiser that is desired to be implemented, considering all the advantages and disadvantages of the methods. Based on these criteria, method 1.c) presented in this paper is the most suitable compromise to generate the primary vocal signal to be used to implement a synthesis-by-rule synthesiser designed to run on a personal computer.

References

1. Rabiner, L.R., Schafer, R.W. (1978) *Digital Processing of Speech Signals*. Prentice Hall, ISBN 0-132-136-031, London, United Kingdom
2. Rabiner, L.R., Schafer, R.W. (2007) *Foundation and Trends in Signal Processing, vol. 1 Introduction to Digital Speech Processing*, Available from: www.nowpublishers.com/ product.aspx?product=SIG&doi=2000000001, Accessed: 09/10/2011
3. Rossing, T.D. (2007) *Handbook of Acoustics*. Springer Verlag, ISBN 978-0-387-30446-5, New York, USA
4. Roach, P. *Speech Technology: a Look into the Future*. Available from: www.racai.ro/books/awde/roach.html, Accessed: 09/10/2011
5. *** *White noise*. Available from: www.en.wikipedia.org/ wiki/White_noise, Accessed: 09/10/2011
6. *** *Random number*. Available from: www.mathworld. wolfram.com/RandomNumber.html, Accessed: 11/10/2010