A copy synthesis method to pilot the Klatt synthesiser

Yves Laprie, Anne Bonneau

A COPY SYNTHESIS METHOD TO PILOT THE KLATT SYNTHESISER

Yves Laprie and Anne Bonneau

LORIA
BP 239
54506 Vandœuvre-lès-Nancy, FRANCE
laprie,bonneau@loria.fr

ABSTRACT

This paper presents a copy synthesis method to controlling the Klatt synthesizer. Our method allows speech stimuli to be constructed very easily. We accepted the parallel branch of the Klatt synthesizer. After formants have been tracked, the amplitudes of the resonators are measured on a spectrum obtained by an algorithm derived from cepstral smoothing called "true envelope". This algorithm has the advantage of approximating harmonics very accurately. The analysis strategy of a speech signal is straightforward: the fundamental frequency is calculated so that voiced regions are known and the frication energy is set to the value of the spectral energy above 4000 Hz. Stimuli which have been created by means of this method have a timbre close to that of natural speech. This copy synthesis method is incorporated in our software for speech research called "Snorri". Therefore, the user has at his disposal a versatile tool for creating stimuli in the context of the Klatt synthesizer.

1. INTRODUCTION

An acoustic synthesis system enabling the direct control of acoustic cues is interesting for numerous perception and phonetic studies. Although several formant based synthesizers have been developed, that of Klatt [1] is undoubtedly the most widely used because it is easily available and turned out to allow speech, very close to natural speech, to be produced.

In spite of its interest, it remains difficult to use this synthesizer alone because it requires 39 parameters (for the 1980 version) to be specified for each frame. In the framework of a text-to-speech synthesis system these parameters are obtained by means of a set of complex rules [2], after a set of parameters has been determined for the basis sounds. As mentioned by Klatt this determination is often the result of a try-and-error method and, therefore, is a tedious task. This preliminary work becomes a more serious obstacle for a phonetician who wants to have at his disposal sounds close to natural speech without giving to much time to the adjustment of the synthesis parameters.

It seems therefore attractive to derive these parameters from natural speech so that synthesized signals sound sufficiently natural. This idea gave rise to the development of copy synthesis systems, for instance, that of W. Holmes [3] which relies on a synthesizer simpler than that of Klatt. But, to the best of our knowledge, there does not exist any copy synthesis system for the Klatt synthesizer. Compared to graphical software for editing Klatt parameters, our system is integrated into a speech analysis system that provides users with copy synthesis tools. This papers describes these tools and strategies for copy synthesis.

2. CHOICE OF THE SYNTHESIZER BRANCH

The Klatt synthesizer is made up of two branches:

- one is a cascade of resonators, the frequency and bandwidth of which have to be specified,
- in the second branch, resonators are connected in parallel. Each resonator is controlled by its frequency, bandwidth and amplitude.

Theoretically, the cascade branch is sufficient to produce all the oral vowels but it has to be supplemented with the parallel branch to produce nasal vowels and consonants. This means that the copy synthesis should be developed for the two branches. Actually, preliminary experiments, which confirm other works in the area of formant synthesizers, have shown that only the parallel branch should be used for copy synthesis. It is indeed possible to generate vowels, the spectrum of which is the same as that would be produced by the cascade configuration. Furthermore, only the parallel branch enables the copy of consonants.

The parallel branch presents a second advantage. Resonators of the cascade configuration are defined by their frequency and bandwidth. However, the determination of formant bandwidth is always a difficult task. The bandwidth has to be measured during the closed phase of fundamental periods, i.e. when the vocal tract is not coupled with sub-glottal cavities. This requires an accurate detection of the glottal closure instants, which is a not an easy task (see [4], for instance) and can give rise to substantial discrepancies between the original and copied spectra because bandwidth influences the amplitude of formants. On the other hand, amplitude and bandwidth are two independent parameters in the parallel configuration. In some formant synthesizers using the parallel configuration, the amplitude is the only variable parameter provided that the bandwidth has been set to a value consistent for each of the formants involved in the synthesis.

We therefore chose to use only the parallel branch of the synthesizer, and in a first time, we set bandwidths to values proposed by Mrayati and Guérin [5].

3. COPY SYNTHESIS

Copy synthesis consists of determining at each frame frequency and amplitude of formants which enable the reproduction of a signal as close as possible to the original speech signal. In the soft-

ware we developed the user has the choice between automatic extraction of formants trajectories by means of our formant tracking algorithm [6] and editing formants trajectories by hand from the spectrogram (in this case the user draws formants directly onto the spectrogram displayed). Moreover, in the later case the user can use two tools:

- an algorithm to register trajectories entered by user onto the spectral peaks of the spectrogram calculated by linear prediction or linear cepstral smoothing,
- a B-spine smoothing algorithm with one control point each 80 ms which is the average duration of a vowel.

At this point, source parameters as well as formant amplitudes have to be determined. We developed a fundamental frequency detector which derives from the algorithm proposed by Martin [7].

The transfer function of speech produced by the synthesizer is as follows: $P(z) = S(z).H(z)$ where S represents the transfer function of source and H the contributions of vocal tract and lip radiation: $H(z) = -H_1(z)+H_2(z)-H_3(z)+H_4(z)-H_5(z)+H_6(z)$ where $H_i(z)$ represents the transfer function associated to formant Fi.

Amplitude parameters of functions $H_i(z)$ are obtained by measures from the original spectrum and source spectrum. It would be possible to calculate $S(z)$ from the analytical form of the temporal excitation signal $s(n)$. Actually, Klatt added several parameters to produce an excitation signal close to real excitations, which gives $s(n)$ a somewhat complex analytical form. Moreover, as it is possible to use an excitation signal stemming from a natural speech signal we preferred to calculate $S(z)$ directly from the file of the excitation signal obtained by synthesis or inverse filtering.

The influence of a formant is almost reduced to the vicinity of this formant, and it is therefore possible to obtain the amplitude parameter simply by measuring it on the speech spectrum. Let Fi be the transfer function of formant Fi: $H_i(z) = \frac{A_i(1-B_i-C_i)}{1-B_iz^{-1}-C_iz^{-2}}$ where B_i and C_i are two real constants defined from the frequency and bandwidth of formant Fi, A_i is given by: $A_i = 20log_{10}|P(z_i)| - 20log_{10}|S(z_i)| - 20log_{10}|H(z_i)|$ where z_i is the complex frequency corresponding to formant F_i.

The source signal saved after synthesis is pre-emphasized to weaken the influence of formants $F2$ to $F6$ in low frequency. On the other hand, the excitation signal of $F1$ has not to be pre-emphasized and H_1 is defined by equation:
$H_1(z) = \frac{A_1(1-B_1-C_1)}{(1-B_1z^{-1}-C_1z^{-2})(1-z^{-1})}$

4. DETERMINATION OF AMPLITUDES

The determination of amplitudes is simple provided that the spectral amplitude of the natural speech and that of the excitation have been evaluated correctly. Actually, this represents one of the major difficulties of the copy synthesis. Formants are often extracted from a linear prediction spectrum. However, weaknesses of this kind of spectral analysis are well known: formants are "attracted" towards intense harmonics and non-vowel sounds are not well modeled. Moreover, there is a substantial lack of precision on the amplitude measures that can often reach 10 dB, and errors done on the original speech spectrum and on the excitation signal spectrum can add together.

Linear cepstral smoothing presents less weaknesses but it is rather sensitive to the position of the analyzing window and the smoothed spectrum may be fairly lower than harmonics. We therefore accepted a method proposed by Imai and Abe [8, 9] called "true envelope". Its principle is to start from a cepstrally smoothed spectrum and to correct it in an iterative manner to cancel the contribution of original spectral values below the current smoothed spectrum. This method requires more computation than traditional cepstral smoothing since two Fourier transforms are used at each iteration. Nevertheless, it gives very good results for the determination of the original speech spectrum as well as the excitation spectrum (see Fig. 1). This method provides spectra close to those obtained by the discrete cepstrum algorithm [10] but it does not requires the prior determination of points to be taken into account.

4.1. Amplitude of the first harmonics

To this point of the adjustment of synthesis parameters, the main difference concerns the first harmonics which are generally stronger in natural speech. Synthesized speech sounds not as low as original speech. However, J.N. Holmes [11] assures that the contribution of these harmonics to the overall perception of speech sounds is weak. He proposes to control the very first harmonics through a formant with a fairly high bandwidth of 150 Hz and a frequency of 200 Hz. This formant is also used to generate nasal vowels. Therefore, we used formant FN of the Klatt synthesizer to modify amplitude of the first harmonics. Its amplitude is set in the same manner as $F1$ since the excitation signal has not to be pre-emphasized.

4.2. Analysis strategies of speech

The analysis strategy we accepted in our first attempts gave good results despite its simplicity. Once the fundamental frequency has been calculated, the voicing intensity is set to 60 dB for all the voiced regions and that of frication to 60 dB for all the unvoiced sounds (unvoiced fricatives and stops).

The automatic amplitude determination presented above ensures that formant amplitudes are correct.

4.2.1. Fricatives

In fricative sounds the first formants are considerably weakened and formants $F4$ to $F6$ correspond to the frication noise. We use a very long window (48 ms) in spectral analyses to determine formant frequencies of $F4$ to $F6$ in frication noises. This prevents very unsteady formant trajectories.

The analysis strategy we just described is slightly modified to take into account voiced fricatives for which voiced and unvoiced sources are combined. We set the frication amplitude to the value of the energy above 4000 Hz and this component is incorporated in the computation of the excitation signal. However, the modeling of voiced fricatives is slightly inadequate for two reasons. First, proportions of frication and voicing should be controlled with more precision as proposed by J.N. Holmes in the JSRU synthesizer [12] that enables adjustment of the frication level for each formant. Secondly, it would be useful to exploit a noise detection method to localize the frication contribution in the frequency domain.

4.2.2. Stops

Until now, we did not implement specific analysis for stop consonants because we consider that results are satisfactory. The only point we will correct is the beginning of the transient of stops which requires the use of the bypass parameter. This parameter

enables the bypassing of all the resonators to produce a flat transfer function which is necessary in the case of labial and dental stops. We will use the segmentation developed in [13] to find the transient noise which contains the most salient acoustic cues of stops.

4.3. Bandwidth optimization

So far we set bandwidth to a default value depending on formant. This allows the setting of formant amplitudes by means of simple measures but does not guarantee a very precise approximation of the original spectrum, particularly when two formants are close together. We thus added an optimization step which is initialized with parameters measured on the spectrum and minimizes the deviation between original and synthesized spectra. Optimization can be carried out for all the parameters (frequency, amplitude and bandwidth of resonators). However, it is not possible to optimize them all together because the low frequency region of the synthesized spectrum is not always correctly modeled by the Klatt synthesizer. Whatever the optimization method (gradient based methods as the Davidson-Fletcher-Powell algorithm as well as methods which does not require the gradient as that of Powel) it tends to use free parameters to correct the low frequency region. Therefore, some parameters, especially F1 and FN frequencies, must be kept constant and the spectrum is optimized separately on several independent frequency regions.

5. EXPERIMENTS

We are now constructing a small database of sounds including logatoms and sentences. This database will be used in the framework of our perception studies. Sounds synthesized are created from natural speech uttered by the same speaker.

Before giving more details on how the stimuli were created, we briefly present the graphical user interface developed to assist users. The interface provides the users with the possibility of calculating synthesis parameters automatically (by means of the copy synthesis method presented above) or of editing parameters superimposed onto the spectrogram. Both possibilities are complementary since it is possible to correct parameters calculated automatically by hand, or to optimize parameters drawn by hand automatically. Each parameter is given a default value. The copy synthesis tools have been implemented in our software Snorri [14] and the graphical user interface of the Klatt synthesizer is available in the Windows version called WinSnoori. Besides copy synthesis tools, this software offers numerous tools to compute and display acoustic cues and parameters.

We have already created some very distinguishable logatoms. This first set of stimuli has been created with formant frequencies and amplitudes as the only variable parameters. All the other parameters were given a default value except F0 which has been extracted from the original utterance. Formant frequencies and amplitudes have been determined by the “true envelope” spectral analysis. Fig. 2 shows the spectrograms of some stimuli synthesized. Synthesized syllables are not only very intelligible but they also sound like natural speech. Our next work will consist of improving the quality of stimuli, especially that of stops which play a crucial role in the perception of speech, and of augmenting the size of the database so that it will contain CV logatoms combining all the phonemes of French.

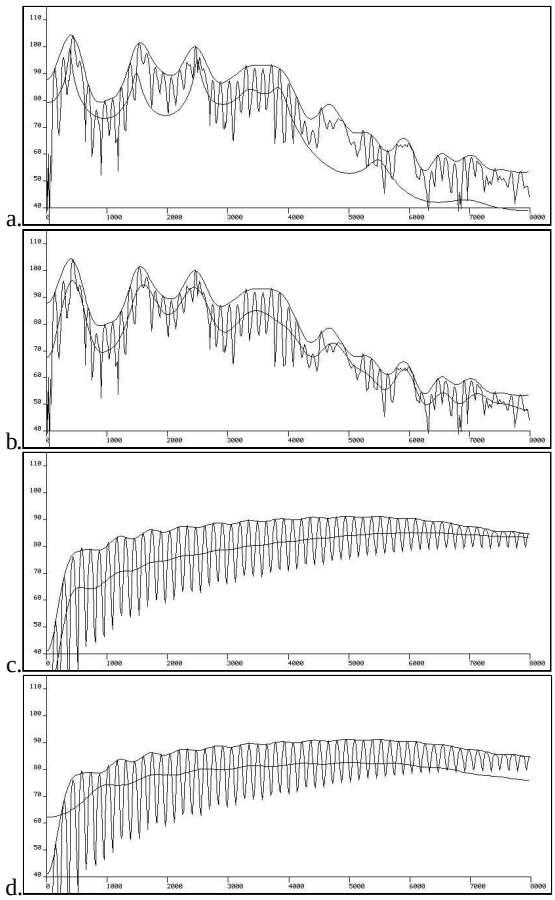

Fig. 1. Comparison of the “true envelope” against linear prediction (a and c) and cepstral smoothing (b and d) for a speech spectrum (a and b) and for an excitation signal (c and d). Figures represent the spectrum obtained by narrow band Fourier transform, “true envelope” and the other spectral analysis under consideration.

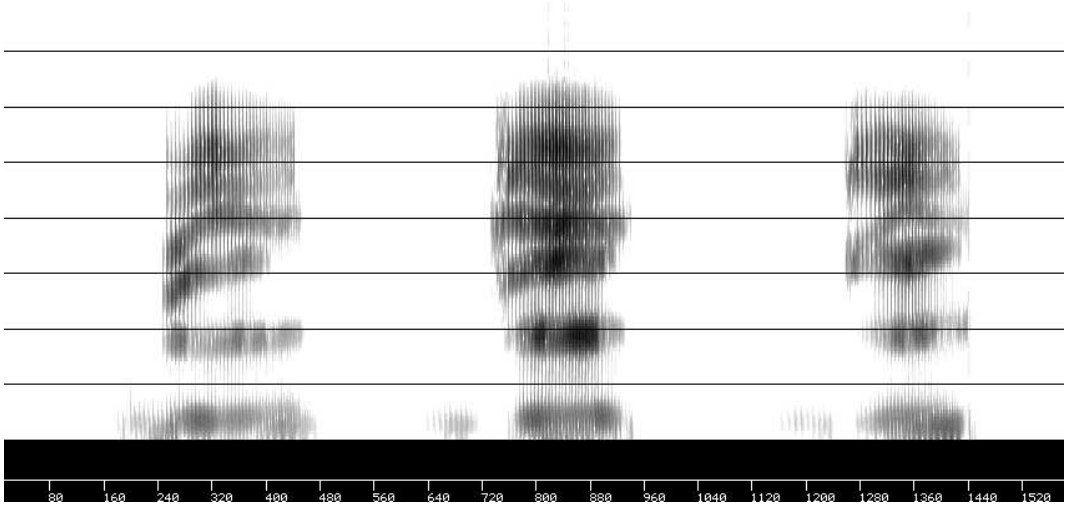

Fig. 2. Spectrograms of signal /bi/ /di/ /gi/ generated by copy synthesis.

Copy synthesis allows synthesized sounds to be given a quality and an intelligibility close to those of natural speech. Indeed, slight variations of the fundamental frequency and of formant frequencies, and more generally all the "imperfections" characterizing human speech, are copied by this technique. Furthermore, this technique reproduces complex links between the speech parameters.

6. CONCLUSION AND PERSPECTIVES

The advantage of the copy synthesis presented above lies in its high flexibility and its integration into the our speech analysis software.

One of our objectives was to not modify the Klatt synthesizer so that stimuli obtained by copy synthesis can be used independently of our speech analysis software. However, we will probably question this choice in the future because it would be interesting to control voicing parameters and frication of the excitation signal as well, according to the formant under consideration.

7. REFERENCES

[1] D.H. Klatt, "Software for a cascade/parallel formant synthesizer," *J. Acoust. Soc. Amer.*, vol. 67, no. 3, pp. 971–995, March 1980.

[2] J. Allen, M. S. Hunnicutt, and D. Klatt, *From text to speech, The MITalk system*, Cambridge University Press, Cambridge, 1987.

[3] W. J. Holmes, "Copy synthesis of female speech using the JSRU parallel formant synthetiser," in *Proceedings of European Conference on Speech Technology*, Paris, France, September, 1989, pp. 513–516.

[4] Y. M. Cheng and D. O'Shaughnessy, "Automatic and reliable estimation of glottal closure instant and period," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-37, no. 12, pp. 1805–1815, December 1989.

[5] M. Mrayati and B. Guérin, "Étude des caractéristiques acoustiques des voyelles orales françaises par simulation du conduit vocal avec perte," *Revue d'Acoustique*, , no. 36, 1976.

[6] Y. Laprie and M.-O. Berger, "Cooperation of regularization and speech heuristics to control automatic formant tracking," *Speech Communication*, vol. 19, no. 4, pp. 255–270, October 1996.

[7] Ph. Martin, "Extraction de la fréquence fondamentale par intercorrélation avec une fonction peigne," in *Actes des* $12^{èmes}$ *Journées d'Etudes sur la Parole*, 1981, pp. 223–232.

[8] S. Imai and Y. Abe, "Spectral envelope extraction by improved cepstral method," *Trans. IECE*, vol. J62-A, no. 4, pp. 217–223, 1979 (en japonais).

[9] P. Halle, "Techniques cepstrales améliorées pour l'extraction d'enveloppe spectrale et la détection du pitch," in *Actes du séminaire "Traitement du signal de parole"*, Paris, 1983, pp. 83–93.

[10] T. Gallas and X. Rodet, "Generalized functionnal approximation for source-filter system modelling," in *Proceedings of European Conference on Speech Technology*, Genova, Italy, September, 1991.

[11] J. N. Holmes, "Formant synthetisers: cascade or parallel ?," *Speech Communication*, vol. 2, pp. 251–273, 1983.

[12] J. N. Holmes, "A parallel formant synthetizer for machine voice output," in *Computer Speech Processing*, F. Fallside and W. A. Woods, Eds. Prentice Hall International, 1985.

[13] Y. Laprie and A. Bonneau, "Burst segmentation and evaluation of acoustic cues," in *Eurospeech, Aalborg, Danemark*, Sept. 2001.

[14] Y. Laprie, "Snorri, a software for speech sciences," in *Proceedings of Matisse 99 (Methods and tools inovations for speech science education)*, London, April, 1999, pp. 89–92.