CONSTRUCTION OF PERCEPTION STIMULI WITH COPY SYNTHESIS

Yves Laprie and Anne Bonneau

LORIA-CNRS

Yves.Laprie@loria.fr and Anne.Bonneau@loria.fr

ABSTRACT

A number of experiments in perception requires the construction of speech-like stimuli whose acoustic content needs to be manipulated easily. Formant synthesis offers the possibility of editing all the parameters of speech. However, the construction of stimuli by hand is a very laborious task and therefore automatic tools are necessary. This paper describes two main extensions of a copy synthesis algorithm previously proposed. The first concerns formant tracking which relies on a concurrent curve strategy. The second is a pitch synchronous amplitude adjustment algorithm that enables the capture of fast varying amplitude transitions in consonants. In addition, the automatic determination of the source parameters through the computation of F0 and of the friction to voicing ratio enables the speech signals to be copied automatically. This copy synthesis is evaluated on sentences and V-Stop-V stimuli.

Keywords: formant synthesis, Klatt synthesizer, copy synthesis, stop consonants

1. INTRODUCTION

Artificial stimuli have been widely used to investigate perception of speech because they enable the acoustic content to be controlled completely and consequently perceptive effects to be separated from each other. However, these stimuli should sound fairly natural to prevent perception processes from deviating too far from those involved in the perception of natural speech. Indeed, Van Hessen and Schouten [2] showed that the quality of stimuli is a determining factor to explain the perception results achieved by subjects in a categorical perception experiments.

The importance of speech-like stimuli explains the success of formant synthesizers, and particularly that of Klatt [5] which is easily available. Even if the quality of stimuli produced by the Klatt synthesizer is not as good as that reached by nowadays non uniform unit synthesizers, it enables the direct manipulation of all the acoustic cues of speech. In addition, there are several bases of rules developed within the context of text to speech [1], which provide the set of parameters required to use the Klatt synthesizer. Although these rules were derived from the analysis of several speakers, they are sometimes not sufficient to capture all acoustic cues, and more importantly, there are often not easily available, and very difficult to modify or to extend. Indeed, the construction of new sounds requires the adjustment of 39 parameters for the initial version of the Klatt synthesizer. These parameters can be set by trial-and-error comparisons of synthetic and natural speech spectrograms but it is a very laborious work. Obtaining these parameters through an automatic algorithm is thus an important objective.

Works carried out by J. Holmes [3] about formant synthesis and W. Holmes [4] about copy synthesis showed that the parallel branch of the synthesizer (resonators are put in parallel) is the most appropriate architecture for copy synthesis. Indeed, parameters of one resonator (amplitude, bandwidth) can be adjusted independently from those of other resonators, at least to some extent. However, there are very few automatic copy synthesis algorithms [12, 9] and very often a postprocessing step [12] or hand editing of parameters is required [13]. In this paper we present a copy synthesis strategy which extends that proposed in [9] about two crucial aspects of copy synthesis: automatic formant tracking and the adjustment of formant amplitudes. Indeed, one key point to generate good quality stimuli is the possibility to capture fast amplitude variations. In addition, we evaluate copy synthesis and propose a small VCV database that can be used for further phonetic investigations.

2. Overview of the copy synthesis strategy

Once formant trajectories have been determined the copy synthesis operates in two steps. The first step consists of calculating F0 (through an algorithm inspired by that of Martin [11]) and determining source parameters. The source of the Klatt synthesizer is defined by the levels of voicing and friction and the same source is used for all formants. The ratio of friction and voicing is set to render the nature of sounds, and particularly voiced fricatives. When speech is unvoiced the source is a pure friction noise. When speech is voiced, the ratio between the voiced source and the noise source is calculated from the autocorrelation factor of the frequency band [2000Hz-4000Hz].

The second step consists of adjusting formant amplitudes.

3. Automatic formant tracking

Although automatic formant tracking received a sustained attention for several decades it is still a problem not satisfactorily resolved. In [7] it was shown how active curves could be used to track formants. The underlying idea is to deform initial rough estimates of formants under the influence of the spectrogram to get regular tracks close to lines of spectral maxima which are potential formants. A formant track is thus represented by a curve $t : [t_i, t_f] \rightarrow \mathbb{R}^2$, $t \rightarrow (t, F(t))$ in the time frequency domain, t_i and t_f are times of the beginning and end of the formant track, and $F(t)$ is the frequency of the formant at time t. The compromise between proximity to spectral peaks and regularity is given by the following functional E

$$E(F) = -\int_{t_i}^{t_f} E_{Spectro}(t, F(t))dt + \lambda \int_{t_i}^{t_f} \alpha|F'(t)|^2 + \beta|F''(t)|^2 dt \quad (1)$$

where the overall energy $E(F)$ has to be minimized. The first term represents the spectrogram energy $E_{Spectro}$ along the formant track. It is thus all the bigger since the curve is close to a line of spectral peaks. The second term represents the length and the curvature and is thus all the smaller since the curve is regular. α influences the curve length, β its curvature and λ the compromise between the spectrogram energy explained by formant tracks and the smoothness of the curve.

Each formant curve becomes deformed under the influence of the spectrogram independently of the other formant curves, what requires a complex control strategy [7] to manage interactions between formants. The main difficulty is when two formants are competing with each other to catch the energy of one spectral peak. This problem occurs when one spectral peak is too weak compared to the other and leads the two formant tracks to get closer to the prominent spectral peak. Another difficulty is the initialization of the tracks that requires the construction and the labelling of elementary tracks (i.e. small pieces of formant tracks obtained by applying a simple continuity constraint) in terms of formants.

We thus recently proposed [6] to modify the deformation equation to incorporate interdependency between formants. This way, formants are deformed by taking into account the deformations of their neighbouring formants with two advantages: a better coverage of the spectrogram energy, a simpler and more robust control strategy. Moreover, the initialization stage can be substantially simplified because the interdependency of formant tracks enables a more dynamic exploration of solutions than that possible with the labelling of elementary tracks based on a static strategy.

This new strategy turns out to be more efficient than that reported in [7] with only a very simple control strategy. The key point is thus the construction of initial rough estimates of formant trajectories. The previous algorithm used a moving average applied onto LPC roots. The window is sufficiently long (200 ms) to remove fast varying variations due to the detection of spurious roots. The counterpart of this long duration is that the moving average prevents formants fairly far from the average frequency to be kept. This is particularly sensitive in the case of F2 which presents low frequency values for back vowels. A simple algorithm to detect back vowels from the overall spectral shape and particularly energy levels has been added in order to keep extreme values of F2 which are relevant. Four curves are used although only the first three ones, i.e. F1 to F3, are relevant. Indeed, the role of the last one (that for F4) is to exert a repulsive force on F3 to prevent the F3 curve to capture spectrogram energy in higher frequency. Although curves are defined along the entire speech segment only points where the spectrogram energy is sufficiently high are relevant. Other points bridge the gaps where the spectrogram energy is weak, for instance during stop closures. For the same reason portions of curves in low energy regions present slopes which are not relevant.

4. Pitch synchronous formant amplitude adjustment

Each resonator of the parallel branch of the Klatt synthesizer is defined by its frequency, bandwidth and amplitude. Once the frequency has been determined through automatic formant tracking, the bandwidth and amplitude have to be estimated. Previous works about the parallel branch of a formant synthesizer and copy synthesis [3, 4, 9] showed that the adjustment of amplitude is sufficient to approximate speech correctly provided that the bandwidth is set to a relevant default value given by literature about acoustics of vowels.

The adjustment of formant parameters thus amounts to finding their amplitudes. However, the adjustment procedure should capture fast events so that consonants can be rendered properly. This requires very short time windows to compute spectra, and consequently the synchronization of these windows on relevant time events. One solution could be the determination of glottal closure instants [14] in order to compute linear prediction spectra at these points. However, the determination of glottal closure instants is not sufficiently robust to be used without further control. We thus use a pitch marking algorithm previously developed to decompose the speech signal into pitch periods [10]. For each pitch period a relevant time instant is localized by searching the time where the amplitude of F1 is maximal. This is achieved by moving a 4ms window (or less if F0 is very high) between two consecutive pitch marks (see Figure 1) and measuring the amplitude of F1 from a wide band spectrum calculated on this window. All formant amplitudes are then measured from the narrow band spectrum where F1 is maximal.

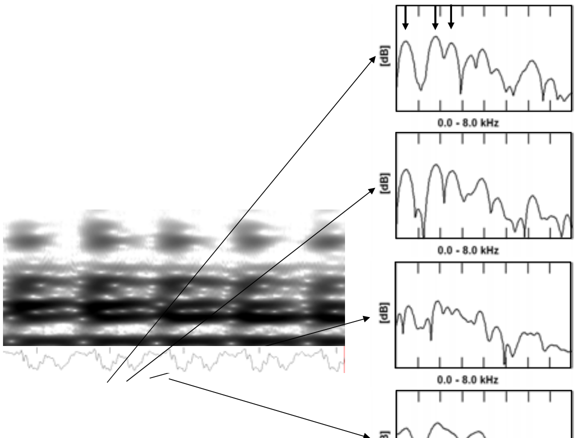

Figure 1: Principle of pitch synchronous amplitude adjustment. Left column: signal with pitch marks (bottom) and wide band spectrogram (top). Right column: four wide band spectra. The first three formants are indicated by arrows in the top spectrum.

5. Evaluations and concluding remarks

This new copy synthesis algorithm has been implemented in the current version of the speech analysis software WinSnoori (http://www.winsnoori.fr). It is fully interfaced with other tools developed to facilitate the edition of Klatt parameter files, and which enable a visual (with respect to the spectrogram) and perceptive control of formant parameters obtained.

We conducted two evaluations, one fully automatic for sentences, and the second one semi-automatic for VCV sequences. The first informal evaluation is about the intelligibility of copied sentences. The time shift between two vectors of Klatt parameters is 4 ms and six formants were tracked. Note that this choice corresponds to a finer analysis than that of formant synthesizers developed in the eighties. As explained above the first formants (F1 to F3) are more relevant than F4, F5 and F6. However, these last three formants add some naturalness to the copied speech signal. The copy synthesis is fully automatic and no correction was done for these sentences. The phonetic information, as well as some speaker specificities are preserved by the copy synthesis algorithm (see example attached to this paper).

The second evaluation focuses on the stop consonants because they involve very fast transitions. We analyzed VCV stimuli with V belonging to /a,i,u/ and C to /p,t,k/ (original and copied stimuli are attached to the paper). Unlike sentences, these stimuli are more difficult to identify. The first reason is that the time shift previously used (4ms) is too large. We thus used a 2 ms time shift between two consecutive vectors of Klatt parameters and edited them by hand to correct most flagrant errors. The correction consisted in adding extra formants when there is no continuity between spectral peaks of the burst and vowel formants, and in correcting amplitudes computed by the copy synthesis algorithm. The objective is to investigate how well stimuli generated by a formant synthesizer and derived from natural speech can be used in perception experiments.

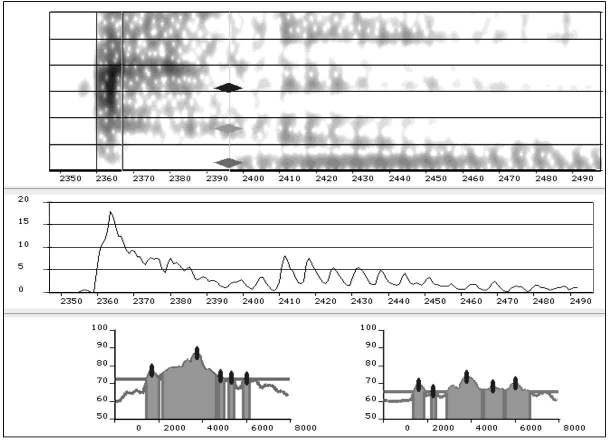

Figure 2: Analysis of the original burst for /atu/. Top: wide band spectrogram with the automatic decomposition of the burst into transient and frication noise, Middle: energy contour calculated on a 4 ms window, Bottom: stylized spectra of the transient and noise

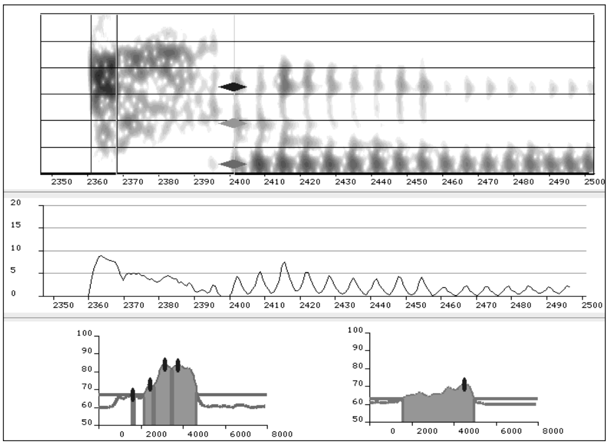

Figure 3: Analysis of the copied burst for /atu/. Top: wide band spectrogram with the automatic decomposition into transient and frication noise, Middle: energy contour calculated on a 4 ms window, Bottom: stylized spectra of the transient and noise

We will first deal with the quality of the stimuli in general, and then will comment upon their perceptual resemblance with the sound they are supposed to produce and their “identifiability”. The stimuli can be easily identified, with relatively few exceptions, and we believe their quality is quite satisfying with respect to stimuli used at present. There still remain some slight problems concerning the stimulus quality, due to the specificities of stop bursts: their sharp attack (the transient part of the burst), the importance of the distribution of energy all over the spectrum, which differentiates diffuse from compact consonants, and their rapid rate of change along the time. These specificities make stop bursts very difficult to replicate. The absence of a very sharp attack makes synthetic stops sound softer than the natural ones, but does not affect the identification of the stimulus (at least in intervocalic position). Our method does not favour the capture of diffuse consonants. This leads to the absence of energy in low and high frequencies, for dentals, and in high frequencies for labials. This lack of diffuseness affects the quality of these stops, dentals being probably especially affected by the lack of energy in very high frequencies. We will discuss below the impact of this absence on the identification of the stimuli.

Figures 2 and 3 illustrate the weakening of the attack. The automatic burst segmentation as well as the spectral stylization were obtained by means of algorithms designed within the context of a study about acoustic cues of stop consonants [8]. It can be seen that the transient durations are quite similar in the original and copy. The most noticeable discrepancies are the smooth energy profile of the transient instead of a sharp one, and the shape of the spectral prominence of the transient which is rendered by two close peaks instead of only one with a larger width. The first problem originates partly in the analysis stage which should be finer and from the time shift (2 ms) used to sample Klatt parameters. Given the sharpness of the transient attack the time shift should be no more than 1 ms and the analyzing window about 2 ms. The second problem is related to the representation of the burst by formants. Although the system of cavities, just after the release of the occlusion, specially for /t/, substantially differs from that of the following vowels (one front cavity excited by a noise source for the transient, and the whole vocal tract excited by a voiced source), both are represented by the same number of formant. In order to better approximate the transient parta very specific frequency and amplitude adjustment is necessary. Additionaly, using six formants does not allow the energy in high frequency (above 5000 Hz) to be copied.

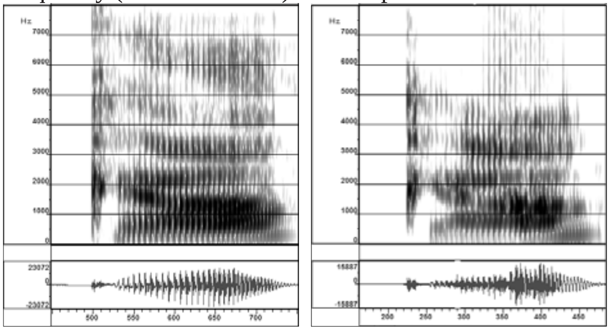

Figure 4: Comparison of the original (left) and copied (right) bursts of /ka/. The absence of energy in high frequency is due to the fact that only six formants are used to analyze speech.

We obtain the best stimuli for velars, the less satisfying for labials. Velar sounds are well reproduced by our method (see Figure 4) because of the continuity between the main velar peak and the following vowel F2 (F3 for front vowels). In particular, their natural resonances were correctly captured by our system. Another characteristic of velars which makes them easier to reproduce is their relatively long burst (around 30 ms and more before high vowels). In French, unvoiced stops are not aspirated, so the duration of french unvoiced stop bursts is very short, in particular for labials (less than 20 ms in most of the cases), and to a lesser extent for unvoiced dentals (around 20 ms, sometimes more). We are quite satisfied with the copy-synthesis of dentals, with the following reserves. Before the high front vowel /i/, the fricative noise is very long and very high in frequency. If the attack is not sharp enough, the stimulus, although identifiable, may sound like an /s/. Another problem comes from the absence of (substantial) energy in high frequencies, for most of the dental stimuli. In one case, in a /ita/ sequence the dental consonant might be confounded with a velar. Labials are the most difficult consonants with regard to the copy-synthesis. This essentially comes from their short duration.

This evaluation shows that simple corrections allow good quality stimuli to be copied from original speech signals. The main future improvement to fulfil the requirement of quality expressed by van Hessen and Schouten [2] to evaluate categorical perception will consist in increasing the time precision of the analysis that pilots copy synthesis. This is quite normal since formant synthesizers were designed in the eighties when the constraint of disk space led to adopt fairly rough time and frequency precisions. This modification is particularly important to increase the quality of bursts which play a major role in the perception and identification of consonants. Furthermore, as mentioned above, we will focus future work on the development of a specific automatic burst copy synthesis procedure. However, the advantage of formant synthesis compared to sinewave representation or another speech coding approach is to enable the direct acoustic and perceptive impact of all the acoustic cues to be evaluated, and the quality of stimuli to be adjusted by manipulating parameters of a higher level than those of sinwave synthesis for instance.

6. REFERENCES

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

[2] van Hessen, A., Schouten, M. 1999. Categorical perception as a function of stimulus quality. *Phonetica* 56, 56–72.

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

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

[5] Klatt, D. March 1980. Software for a cascade/parallel formant synthesizer. *J. Acoust. Soc. Amer.* 67(3), 971–995.

[6] Laprie, Y. Oct. 2004. A concurrent curve strategy for formant tracking. *Proc. ICSLP* Jegu, Korea.

[7] Laprie, Y., Berger, M.-O. October 1996. Cooperation of regularization and speech heuristics to control automatic formant tracking. *Speech Communication* 19(4), 255–270.

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

[9] Laprie, Y., Bonneau, A. Sept. 2002. A copy synthesis method to pilot the Klatt synthesiser. *International Conference on Speech and Language Processing, Denver, USA*.

[10] Laprie, Y., Colotte, V. September 1998. Automatic pitch marking for speech transformations via td-psola. *Proceeding of the European Signal Processing Conference* Rhodes, Greece. 1133–1136.

[11] Martin, P. 1982. Comparison of pitch detection by cepstrum and spectral comb analysis. *Proc. of Int. Conf. Acoust., Speech, Signal Processing 1982* 180–183.

[12] Scheffers, M., Simpson, A. August 1995. Lacs: Label assisted copy synthesis. *Proceedings ICPhS* volume 2 Stockholm. 346–349.

[13] Simpson, A. August 1995. A tool for the complete production of copy syntheses from natural tokens. *Proceedings ICPhS* volume 2 Stockholm. 350–353.

[14] Smits, R., Yegnanarayana, B. September 1999. Determination of instants of significant excitation in speech using group delay function. *IEEE Trans. on Speech, and Audio Processing* 3(5), 325–333.