Pure audio McGurk effect

Laurent GIRIN

Institut de la Communication Parlée
INPG/Université Stendhal/CNRS
B.P. 25, 38040 Grenoble CEDEX 09, France
girin@icp.inpg.fr

ABSTRACT

In the experiments described in this paper, [aba] and [aga] speech sequences are combined in such a way that an [ada] stimuli is obtained, referring to the well-known McGurk effect. But contrary to the standard experiments, where audio [aba] and visual [aga] stimuli are combined, only audio signals are considered here, resulting in what is called a pure audio McGurk effect. The processing consists in modelling the audio signals by the classical linear prediction model and then linearly combining the [aba] and [aga] sequence LP filters that model the contribution of the vocal tract before resynthesis. The relation of the experiments with the problem of data representation and the nature of the audio-visual integration space is discussed.

1. INTRODUCTION

The probably most demonstrating paradigm of the bimodal (audio and visual) nature of speech can be found in the so-called McGurk effect [1, 2]. The principle is the following: subjects are presented the image of a speaker pronouncing the sequence [aga] while the sound is the sequence [aba], resulting in the perceived sequence [ada]. This effect has been extensively studied by perception experts and psychologists (see further references in [3]). Furthermore, it has become a paradigm for assessing the human behaviour matching of audiovisual speech integration models (eventually dedicated to automatic AV speech recognition). One main reason for the McGurk effect to be so attractive for audiovisual speech experts is that it questions the problem of the nature of the integration space for visual and acoustic data, and the representation of the data in such space. A comfortable explanation of the McGurk effect would be to find a perceptually meaningful space where [d] is placed somewhere between [b] and [g]. Unfortunately, such property does not emerge, neither from perceptual judgments [4] nor from acoustic characterization, through the classical diffuse-rising ([d]), diffuse-falling ([b]) and compact ([g]) gross spectrum shape evidenced by Blumstein and Stevens [5].

For the following of this paper, it is interesting to remind the typical trajectories of the three first formants of [aba], [aga] and [ada] sequences for a male speaker. These trajectories can be grossly characterised by the plot in Fig. 1. Briefly, the three formants drop in vowel to consonants transitions for [aba] while F2 is rising for [aga], resulting in a characteristic “hub-locus” at the F2-F3 convergence. Both F2 and F3 are rising for [ada]. Therefore, from a formant-based distortion criterion, [aga] is between [aba] (differentiated by F2) and [ada] (differentiated by F3). The term “between” is inexact since [g] is rather at the 90° angle of the [bdg] triangle in the F2-F3 space at the consonant release instant, while the common target [a] is on the [bg] hypotenuse (see Fig. 2 and [6] for a more complete description).

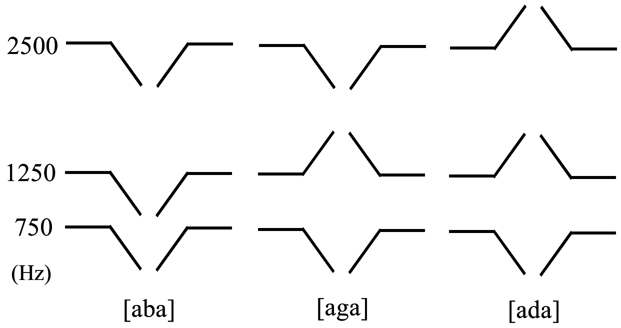

Figure 1: Typical trajectories of F1, F2 and F3 for [aba], [aga] and [ada] sequences.

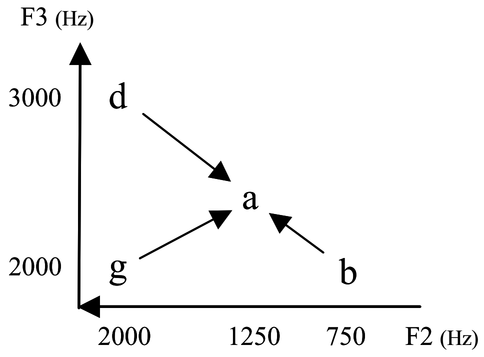

Figure 2: Audio movements of [b], [d], [g] towards [a] in the F2-F3 space (after [6]).

All those considerations do not claim in favour of an AV integration space of acoustic nature with a spectral-like representation for the data to be integrated, as it is in the Dominant Recoding model of AV integration, where the video information is recoded into spectral information (be)for(e) integration with the audio information [7, 8]. Now, in this paper, what is called a pure audio McGurk effect is presented, that is the combination of [aba] and [aga] sounds in such a way to obtain an [ada] sound. The signal processing scheme basically consists in modelling [aba] and [aga] audio sequences by the classical linear prediction model [9] and then linearly combining the LP filters that model the contribution of the vocal tract of both sounds before resynthesis. Thus, the [aga] video stimulus of the classical McGurk effect is somewhat replaced with direct information on the vocal tract shape associated with and estimated from the [aga] sound. The process is technically detailed in section 2, but it can be already mentioned here that, after Fig. 1, the effect of the integration of the [aga] filter within the [aba] filter in the presented algorithm should be *a priori* to mean (thus flatten) F2 while leaving the general behaviour of F3 unchanged, leading to a resulting filter globally always closer to [aba] or [aga] than [ada] vocal tract description. These considerations make the results obtained in this study surprising *a priori*. Analysis and possible interpretation of the effect are discussed at the end of the paper, which is organised as follows. In the next section, the speech processing algorithm is described. Then the data are described in section 3 and the results are given in section 4 and discussed in section 5.

2. THE PROCESS

The speech processing algorithm that is used to combine the [aba] and [aga] audio stimuli is basically based on the linear prediction (LP) model [9] and mainly consists in three steps performed on a frame by frame basis (Fig. 3):

1. In the first step, an LP analysis is performed on both [aba] and [aga] speech signals: the coefficients of the LP analysis filters $A_b(z)$ (for the [aba] sequence) and $A_g(z)$ (for the [aga] sequence) are calculated on successive frames of signals by using the autocorrelation method with hamming windowing and the corresponding residual signals $e_b(n)$ and $e_g(n)$ (prediction errors) are extracted by filtering the successive speech signal frames through the corresponding analysis filters.
2. In the second step, an [aba/aga] hybrid analysis filter $A_h(z)$ is calculated for each frame by linearly combining the contribution of both filters (see below).
3. Finally, the "audio McGurk signal" is synthesized by filtering the residual of the [aba] sequence $e_b(n)$ through the LP hybrid synthesis filter $1/A_h(z)$. The $e_b(n)$ (and not $e_g(n)$) signal must be used because it corresponds to the audio modality of the standard McGurk effect, but results involving the use of $e_g(n)$ are also briefly given and discussed.

As the processing is performed on a frame-by-frame basis, some precautions must be taken in order to ensure an accurate analysis and the coherence of the complete synthesis signal. The signals were sampled at 16 kHz. In the experiments that lead to the results presented in this paper, the length of the analysis (hamming) window was fixed to 20 ms with an overlap of 50%, so that the length of the synthesis frame was 10 ms. The order of the LP models was fixed to 20. To ensure smooth and secure transitions between the synthesis filters of adjacent frames, a lattice filtering scheme was implemented [9, 10] involving the use of the reflection coefficients representation for the filters. More precisely, these coefficients, denoted k_i, allowed to (1) ensure smooth transition by linear interpolation of the k_i values between adjacent frames on a transition segment of a few milliseconds, a quite classical routine in LPC-based speech coding application [9, 10] (2) obtain secure (bound to be stable) hybrid $1/A_h(z)$ synthesis filters by linear combination of the [aba] and [aga] k_i coefficients with a linear weight α taking values between 0 and 1 since a synthesis filter is assumed to be stable if the associated k_i have a modulus smaller than 1 [9].

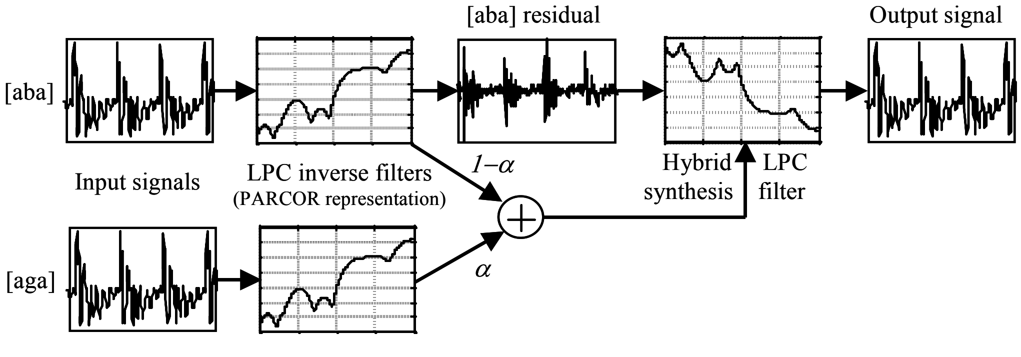

Figure 3: Schema of the process

Thus, the combination of the [aba] and [aga] filters of step 2 is given by:

$$k_i^h = \alpha k_i^g + (1-\alpha) k_i^b \quad \text{for } i = 1 \text{ to } 20, \text{ with } 0 \leq \alpha \leq 1 \qquad (1)$$

where k_i^b, k_i^g and k_i^h denote the coefficients associated respectively with the [aba], [aga] and hybrid [aba/aga] filters. With such convention, α is proportional to the [g] contribution in the audio McGurk effect ($\alpha = 0$ assumes that the original [aba] sequence is unchanged).

3. STIMULI

Two sets of [aba]/[aga] signals were used in this study, corresponding to two French male speakers (let call them PE and JL). The set with speaker PE was used in [3] fo experiments on the standard McGurk effect. The set wit speaker JL was used in [11] for a study on audio-visua enhancement. This last set is in fact composed of [ababa and [agaga] sequence (Fig. 4). When differentiating i unnecessary, the sounds are still denoted by [aba] an [aga] in the following for simplification. Compared to th PE set which was pronounced at a quite low rate (classical condition for the McGurk effect), the JL set wa pronounced at a quite high rate (see Fig 6 and 7).

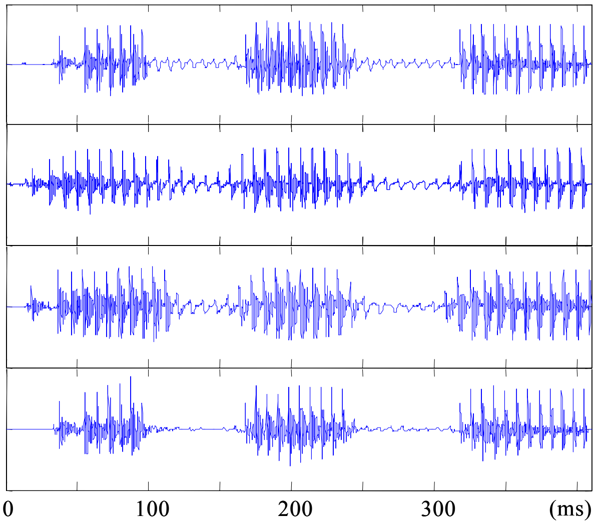

Figure 4: [aCaCa] sequences for speaker JL (samples vs. time in ms); from top to bottom: [ababa], [agaga], [adada] sequences and the hybrid output signal resulting from the process (speaker JL, $\alpha = 0.55$).

No particular pre-processing was applied to these signals before applying the combining process. As a special case, no time warping was processed to perfectly synchronise the sequences phoneme by phoneme. The signals were naturally structurally close enough so that they were simply synchronized at best, in such a way that the consonantal bursts of the two [b/g] pairs were approximately corresponding (the consonantal portion of the stimuli are obviously more sensitive to the combining process than the vocalic portion).

Note finally that [ada] ([adada] for JL) sequences were also recorded and are available for comparison with the results of the processing in section 5.

4. RESULTS

The [aba][aga] signals were processed for different values of α. It was verified that equation (1) provided correct LPC filter/spectrum averaging, as illustrated in Fig. 5.

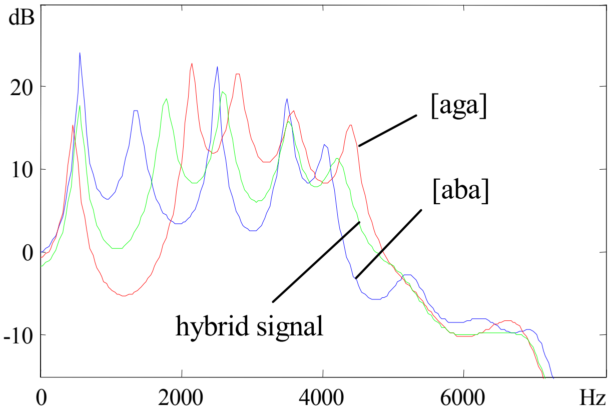

Figure 5: normalized (centered log) LPC spectrum of corresponding frames of the [aba], [aga] and hybrid output signals (speaker JL, $\alpha = 0.55$).

Since the sound generated with α close to 1 was clearly identified as [aga] and $\alpha = 0$ provided no modification to the original [aba] sequence, preliminary experiments consisted in progressively adjusting the value of α with a dichotomy strategy and abundant informal listening tests. The output signal was found to balance from [aba] to [aga] around a quite small interval of α values with progressive modification of the signal in this interval. More precise adjustment of α lead to the generation of what could be identified as [ada] for JL and "between [aða] and [ada]" for PE, thus resulting in the so-called pure audio McGurk effect. "Optimal arbitrary values" of α were quite different for the two speakers (0.55 for JL and 0.15 for PE). This point is discussed in the next section.

Once this subjective optimal α value was fixed for each speaker, more formal listening tests were conducted: a total of seven French subjects were asked to identify the output sounds. They could listen to the stimuli as many times as desired, in a quiet room and with headphones, and without any a priori information of any kind on the stimuli and the purpose of the study. For speaker PE, four naïve subjects identified the stimuli as [ada] and three subjects with strong skills in phonetics identified it as [aða]. For speaker JL, four subjects identified the two consonants as [d] (remind that the stimuli from speaker JL are of the form [aCaCa]) and three subjects (two among them having strong phonetic skills) identified it as [agada].

Complementary informal experiments consisted in substituting the [aba] residual signal $e_b(n)$ with the [aga] residual signal $e_g(n)$ to excite the hybrid synthesis filter. In this case, the effect was more difficult to obtain for speaker JL: for the "optimal" value of α, [aga] was now perceived. So, α had to be decreased to increase the contribution of the [aba] spectral characteristics and the resulting hybrid sound was more difficult to identify (e.g. between [aða] and [ava] depending on α values). However, in the case of speaker PE, the permutation of the $e_b(n)$ and $e_g(n)$ excitation signals had less important effect. So, it may be slightly exaggerated to conclude that the excitation contribution makes the present study an additional step closer to the classical McGurk effect, since both are obtained if the [b] source signal is involved in the process.

The reader is given the possibility to conduct its own experiments and make its own judgment on the output signals with the material at disposal (see section 6).

5. DISCUSSION

Acoustic analysis: Since these experiments concern pure audio signals in contrast with the classical audiovisual McGurk effect, discussion of the results begins with an acoustic analysis of the signals. These results can be explained in the spectral domain by considering the spectral modification of the synthesis filter compared to the "original" [aba] synthesis filter, which would allow perfect reconstruction of the signal (as is the case with $\alpha = 0$). Fig. 6 gathers the spectrograms of the different transitions involved in the process for speaker JL. We can see that F1 and F2 globally follow the behaviour given in Fig. 1 but it can be noted that the variation of F2 is significantly larger, with higher values reached, for [aga] than for [ada]. Thus, the combination of the [aga] filter with the [aba] filter (plotted in Fig. 6 bottom-right) attenuates the F2 transition of the former and makes it finally closer to the [ada] F2 transition of this speaker.

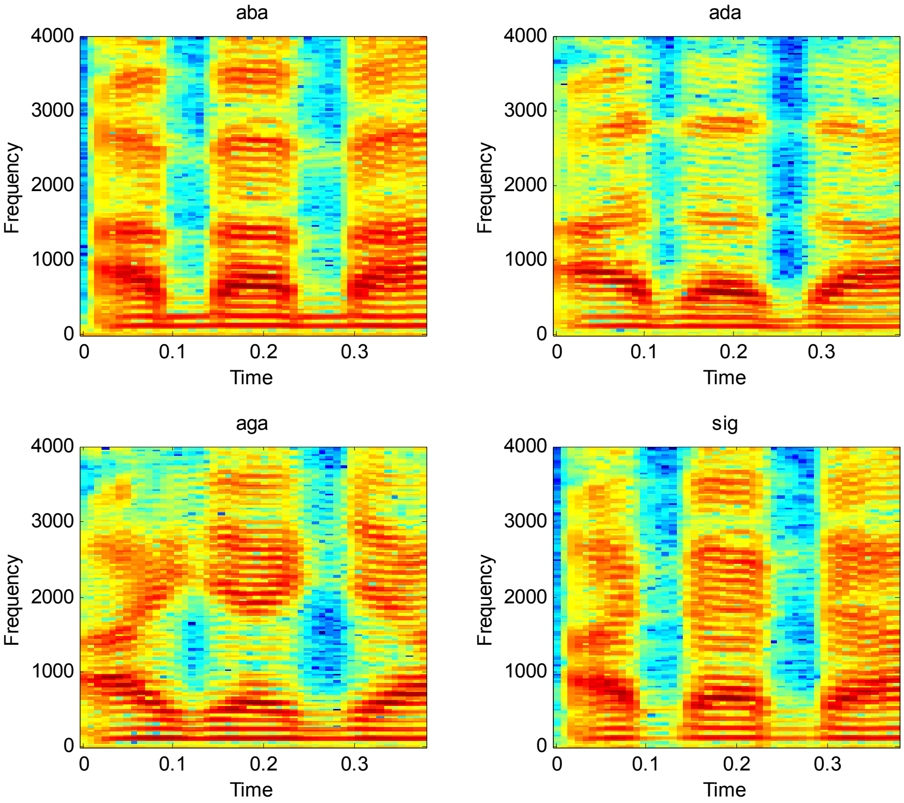

Figure 6: FFT-based spectrograms of the [aCaCa] sequences for speaker JL; from left to right and up to down: [ababa], [adada], [agaga], and the hybrid output signal ($\alpha = 0.55$).

Since the trajectories of formant F3 seems generally more flat on all these spectrograms than on the schema of Fig. 1, they may be perceptually poorly discriminating and therefore may not prevent the observed effect. This is particularly true for the last CV transition with an F3 behaviour already close to [da] in the [agaga] stimulus! This last point may explain the choice of the three subjects that perceived [agada]. Altogether, considering those data, the perceptually important spectral modification generated by the process (integration of [aba] and [aga] characteristics) may be the opening of the [aga] characteristic hub, leading to [ada] perception.

For speaker PE, the results are more difficult to explain from the spectrograms (Fig. 7) because F3 was particularly difficult to track (even with the use of LPC spectrograms that were less confused than FFT-based spectrograms, as the ones plotted in Fig. 6). It can only be noted that once again, the effect of the [aba]/[aga] integration process is to lower the amplitude of F2 trajectories, making them closer to the [ada] trajectories. This is done more drastically than with speaker JL because the "optimal" value of α is here 0.15, making the [aba] contribution more important than with speaker JL (where the optimal α was 0.55).

A new insight into the audio-visual integration modelisation problem? Taking the opposite view of section 1, such convenient acoustic analysis of the results may now be considered as a strong support for the Dominant (audio) Recoding model of audio-visual integration [7, 8], provided that the video [aga] of the classical McGurk effect is recoded into a spectral LPC-like time-varying envelope. Of course, this study does not provide any direct proof of this last point (the video to spectrum recoding process). But it provides a basis for considering the possibility of simple (linear) combination of [b]/[g] patterns in a spectral space resulting in a pattern acoustically and perceptually close to [d]. However, great care must be taken concerning the generality of the results, given that this assumption supposes to partly "relax the [b][d][g] geometry" discussed in section 1.

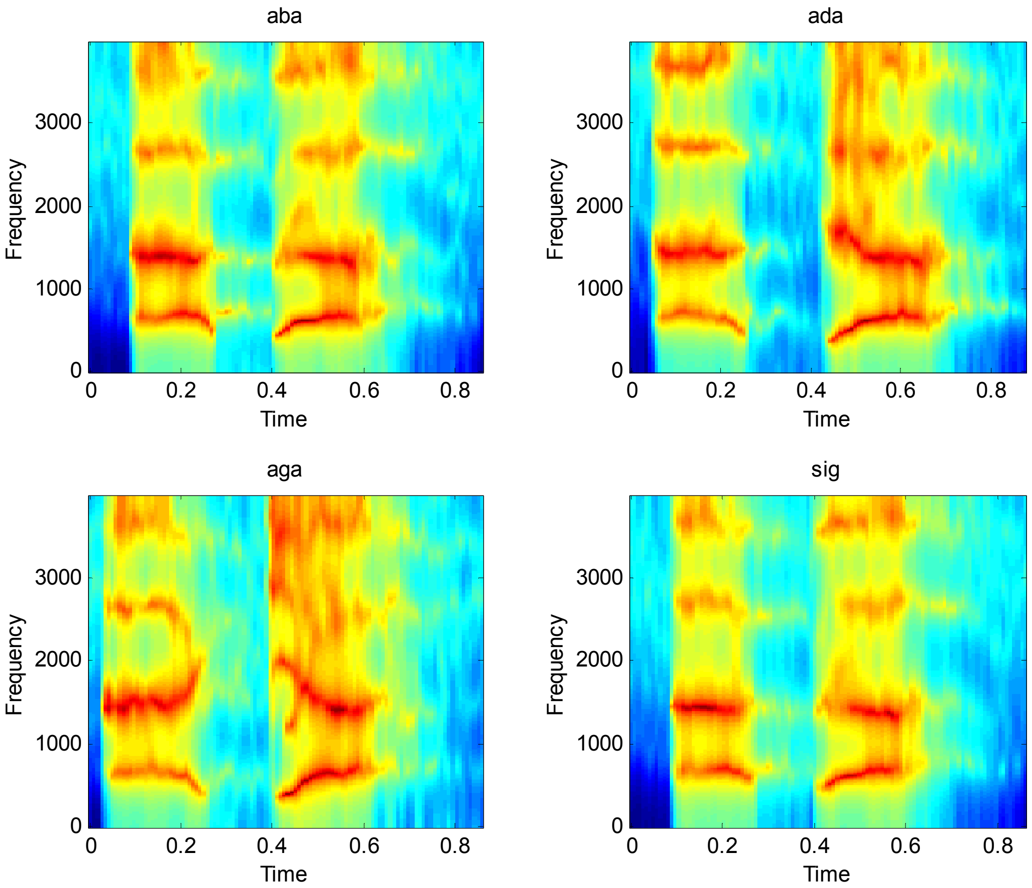

Figure 7: LPC-based spectrograms of the [aCa] sequences for speaker PE; from left to right and up to down: [aba], [ada], [aga], and the hybrid output signal ($\alpha = 0.15$).

In other words, some "[d] between [b] and [g]" unexpected geometry is verified in the spectral space by the data and process used in this study, but that may precisely depend… on the data. Further application of the process on other sets of stimuli (including other phonemic interactions), and the test of other spectral representations for the integration process (e.g. LSP coefficients) may provide partial answer to this question.

The last point that will be mentioned in this discussion section concerns the unsymmetrical effect of the [aba] and [aga] excitation signals, briefly described in the result section. A possible explanation may be found in the signals energy, since they both have flat spectra after LPC filtering. The repartition of the excitation energy between vocalic and consonant parts of the signal is slightly different for [aba] and [aga], with a weaker energy in the consonant part for the former, as illustrated in Fig. 8. This may explain for example why the consonant part of the hybrid output signal (obtained with [aba] excitation and a filter integrating [aga] characteristics) of Fig. 4 is weaker than in the original input signals. On the contrary, filtering a synthesis filter integrating dominant [aba] characteristic with the [aga] excitation produces a high-energy signal in the consonant part, leading to more difficulty to obtain the audio McGurk effect. This point and its relation with the Dominant Recoding hypothesis should be considered in more details in the future.

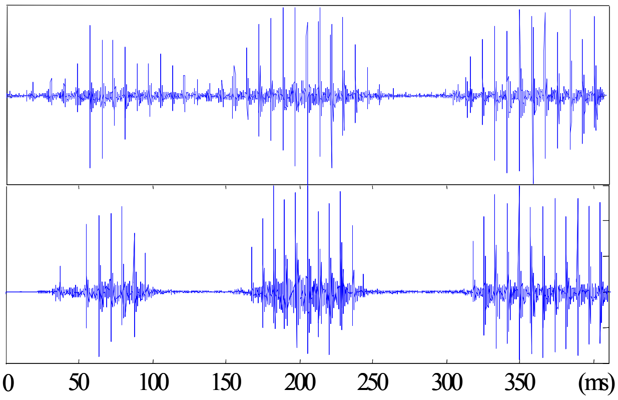

Figure 8: Excitation (LPC residual) signals for [agaga] (top) and [ababa] (bottom) sequences of speaker JL.

6. MATERIAL

The presented algorithm was implemented within the MATLAB environment. The (commented) program and data files (".mat" format) can be retrieved at http://www.icp.inpg.fr/~girin/. All the analysis / synthesis parameters of the program can be easily modified and interested users are invited to give a feedback to girin@icp.inpg.fr.

ACKNOWLEDGEMENTS

The author would like to thank J.L. Schwartz and C. Abry for their helpful comments on this paper.

REFERENCES

1. H. McGurk & J. MacDonald (1976), Hearing lips and seeing voices, *Nature*, vol. 264, 1976, pp. 746-748.
2. J. MacDonald & H. McGurk (1978), Visual influences on speech perception processes, *Perception and Psychophysics*, vol. 24, pp. 253-257.
3. M.A. Cathiard, J-L. Schwartz & C. Abry, Asking a naïve question to the McGurk effect: why does audio [b] give more [d] percepts with visual [g] than with visual [d]?, *Proc. AVSP'2001*, Aalborg, Denmark, 2001,
4. B.E. Walden, A.A. Montgomery, R.A. Prosek & D.M. Schwartz, Consonant similarity judgments by normal and hearing-impaired listeners, *J. Speech Hearing Res.*, vol. 23, March 1980, pp. 162184.
5. S.H. Blumstein & K.N. Stevens, Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants, *J. Acoust. Soc. Am.*, vol. 66 (4), Oct. 1979, pp. 1001-1017.
6. C. Abry, [b]-[d]-[g] as a universal triangle as acoustically optimal as [i]-[a]-[u], *Proc. Int. Cong. Phonetic Sciences (ICPhS'2003)*, August 2003, Barcelona.
7. Q. Summerfield, Some preliminaries to a comprehensive account of audio-visual speech perception, in *Hearing by Eye : The Psychology of Lipreading*, B. Dodd and R. Campbell (Eds.), Lawrence Erlbaum Associates, London, NJ, pp. 3-51, 1987.
8. J-L. Schwartz, J. Robert-Ribes, and P. Escudier, Ten years after Summerfield... a taxonomy of models for AV fusion in speech perception, in *Hearing by Eye, II. Perspectives and Directions in Research on Audio-visual Aspects of Language Processing*, R. Campbell, B. Dodd and D. Burnham (Eds), Erlbaum /Psycho. Press, Hillsdale, NJ, pp. 85-108, 1998.
9. J.D. Markel & A.H. Gray, *Linear Prediction of Speech*, Springer-Verlag, New-York, 1976.
10. L.R. Rabiner & R.W. Schafer, *Digital processing of speech signals*, Englewood Cliffs, NJ: Prentice Hall, 1978.
11. L. Girin, J-L. Schwartz and G. Feng, Audio-visual enhancement of speech in noise, *J. Acoust. Soc. Am.*, vol. 109, June 2001, pp. 3007-3020.