A PRELIMINARY STUDY OF AN AUDIO-VISUAL SPEECH CODER : USING VIDEO PARAMETERS TO REDUCE AN LPC VOCODER BIT RATE

Elodie FOUCHER, Gang FENG & Laurent GIRIN

Institut de la Communication Parlée
INPG/ENSERG/Université Stendhal
B.P. 25, 38040 Grenoble CEDEX 09, France

foucher@icp.inpg.fr

For presentation at EUSIPCO'98, Rhodes, Greece

ABSTRACT

Today there exists numerous speech coding techniques which allow to transmit and stock efficiently the acoustic signals. But speech is both auditory and visual : audio and video information are complementary and the lip shape of the speaker can help the listeners to better understand what is said. This paper aims to show the interest of video information in the speech coding domain in general and in terms of transmission rate in particular.

1 INTRODUCTION

Speech is multimodal : it is both auditory and visual. Thus, seeing the speaker's face, in particular his lip shape, can help the listeners to better understand the message, especially when the environment is noisy. The lip movements are strongly correlated with the acoustical signal, and audio and video information are complementary. The visual information, in particular the lip movements, has already been exploited in different fields, as speech recognition and noisy speech enhancement, in which significant improvements have been obtained [2] [6] [7]. Although there exists today a great number of speech coding techniques which allow to significantly reduce the bit rate for transmission and storage of speech signals, none of them exploits the speaker's face information. So, it is interesting to study a speech coding system which exploits the complementarity between the acoustic signal and the information about the speaker's lip movements. This paper describes a first study which aims to show the contribution of video information for the speech coding purpose.

The contribution of visual information to speech perception is uncomplete : speech signals cannot be well coded only from video information. But, one know that seeing the lip movements of the speaker helps to better understand what is said so the visual information improves the intelligibility of the message. But, none information about the speech signal quality is brought by the vision of the speaker. This suggests that visual information is only linked to global perceptive shape of the signal synonymous with intelligibility (spectral envelop), and not to its detailed structure, only linked with the excitation source that determines the quality of the speech signal. In order to illustrate the contribution of visual information, we have chosen to work on a very low bit rate coder which privileges the aim of intelligibility of the transmitted speech signal instead of its quality. So, we aim to build an audio-visual speech coding system which improves the performances of a very low bit rate coder. For this study, we have chosen to work on a vocoder based on linear prediction coding (LPC) with a bit rate 2,4 kbit/s.

In this paper, we present the original study of an audio-visual vocoder which exploits visual information to reduce the transmission rate of a classic vocoder. In this condition, we focus on the improvement of intelligibility instead of the quality of speech signals. Compared with a classic vocoder, the principle of an audio-visual vocoder is to estimate the filter representing the vocal tract from acoustic and visual parameters during the synthesis step. In section 2, we describe in details the audio-visual vocoder operating. In section 3, we expose the method to estimate the LPC filter. In sections 4 and 5, we present the protocol of experimentation and the results.

2 STRUCTURE OF THE AUDIO-VISUAL VOCODER

The functioning of this coder is based on the classic principle of analysis-synthesis (frame by frame) of the LPC coders (figure 1). First, the coefficients of the filter 1/A(z) modelling the vocal tract are obtained by

means of an LPC analysis [5]. The prediction error is extracted by inverse filtering of the speech signal through A(z). This error can be replaced by a white noise or a pulse train whose parameters are transmitted to the decoder.

The audio parameters used in this study are the Log Area Ratio, or LAR coefficients, because they are more adapted to this application and avoid the instability of the filter. They can easily be deduced from PARCOR coefficients k_i (partial correlation) :

$$lar_i = \log\left(\frac{1+k_i}{1-k_i}\right)$$

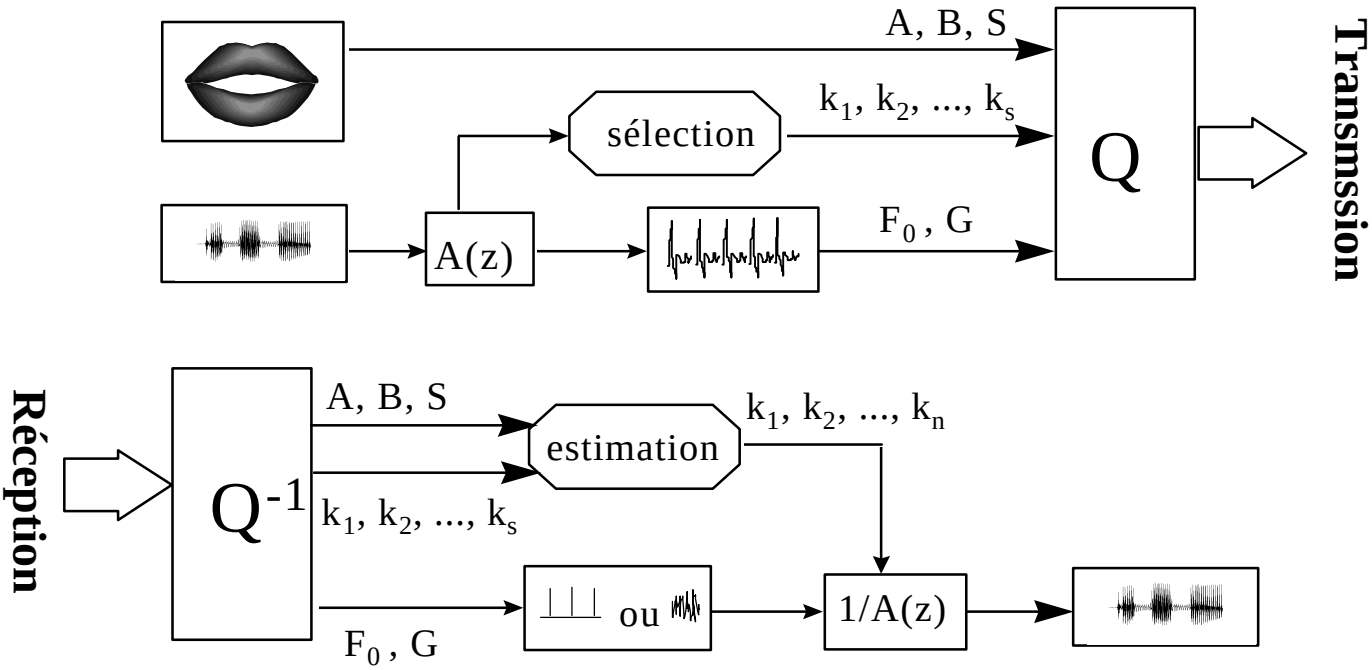

Figure 1 : Principle of the audio-visual vocoder

As for the parameters describing the speaker's lip movements, we use the face processing system developed at the ICP [4] which allows to automatically extract three basic parameters of the labial contour : interolabial width (A), height (B) and area (S). These parameters are transmitted as well as a subset of audio LPC parameters of the filter 1/A(z), selected here to reduce the bit rate. In the decoder, the filter is reestimated from selected audio parameters and video coefficients in accordance with the method described in the next section. The speech signal is then synthesised by filtering of the modelled source through 1/A(z), estimated from audio-visual transmitted coefficients.

3 ESTIMATION OF THE FILTER 1/A(z) FROM TRANSMITTED COEFFICIENTS

The basic problem is to estimate the filter 1/A(z) from transmitted coefficients, that is to say video parameters (V) and the subset of audio parameters (A1). The concatenation of these vectors is called AV=[A1 V]. We have to estimate, from this bimodal vector AV, the vector A2 of audio coefficients which have not been transmitted. The reduction of the transmission rate is obtained thanks to the fact that the transmission of the video vector V needs less bits than the one of A2. To solve this problem of estimation, linear regression method is used, because of its simplicity and efficiency concerning our application. Its efficiency has already been shown in different applications of audio-visual speech signal processing such as recognition and noisy enhancement [2] [7]. So, we can suppose first that it would be also efficient in the field of speech coding.

Here is its principle : the product between the vector AV and a matrix M leads to an estimation of the vector A2 of non-transmitted audio coefficients. M is determined during a learning phase by using the linear regression method between two matrix built by concatenation of a set of learning vectors of both spaces "non-transmitted audio" and "transmitted audio + video". The principle of this method is to minimize the squared error :

$$e = \| (AV)_{learn} * M - (A2)_{learn} \|.$$

We must notice that the importance of the audio coefficients in respect with the spectral envelop description is increasing as the coefficients order decreases. That explains that the first coefficients are generally coded with higher precision than the last ones. That also explains that the vector A1 contains the first coefficients while A2 is composed with the last ones. Different cases have been studied, in particular concerning the number of audio parameters transmitted to evaluate the quality of speech signals obtained after the decoding phase. As the size of A1 is variable, the vector AV changes too. We have studied 10 different configurations so there are different 10 associators (from 0 to 9 audio coefficients transmitted) .

4 PROTOCOL OF EXPERIMENTATION

The corpus, on which we have tested the audio-visual vocoder, have been chosen voluntarily simple in order to evaluate first the feasibility and the efficiency of the proposed audio-visual coder. This kind of project is really innovating so it is important to work first on simple speech signals to see what kind of difficulties we meet before testing our coder with more complex speech signals.

In order to show the efficiency of the audio-visual coder, we use a corpus containing 10 stationary vowels [a e ɛ i ∅ œ o u y] and a corpus containing vowel-consonant transitions of the following form $V_1CV_2CV_1$ taken from the two groups [a i u y] and [p t k b d g]. Each one of these corpus is divided in a learning corpus used to build the audio-visual associators, thanks to the method of linear regression detailed in section 3, and a test corpus used for the following experiences.

First, we intended to show that there is more information, in terms of intelligibility, in the three video parameters [A B S] than in the first three audio parameters [lar_1 lar_2 lar_3]. At this point, the associator is relating the visual only and audio spaces. A recognition test has been done with 10 listeners who have heard three times the 10 stationary vowels i.e. 30 stimuli.

The second test aims to show a reduction of the bit rate while keeping the quality of the coding unchanged compared with a classic vocoder. This test has been performed on the same corpus as before. This time, the visual parameters and some first audio coefficients are transmitted. The number of the audio coefficients transmitted is increasing until the quality of the coded signal reach the same as the one coded by a classic vocoder.

An auditory test of the quality between different speech signals has been performed with 10 listeners who have heard one time the signal coded with a classic vocoder. Then, they have designed, among the different speech signals coded with the audio-visual vocoder, the one whose quality was the same as the one of the signal coded with the classic coder. We have then noted the number of audio coefficients transmitted with the visual parameters needed to reach the same quality of coding as our reference coder.

5 RESULTS

The figure 2 presents the results obtained for the test of intelligibility on the corpus of stationary vowels and for which the LPC filter has been estimated only from the video parameters. It can be seen that the same number of transmitted coefficients, three video parameters, allow to rebuild an intelligible signal whereas only the first three audio parameters cannot.

A test of intelligibility has been performed on the corpus of stationary vowels (figure 2) and on the corpus of vowel-consonant transitions and the conclusions are the same : with the same number of transmitted coefficients, we have more information in three video parameters than in three audio parameters.

The contribution of video information is evident but uncomplete : some vowels, for example [e] and [i], or some consonants, for example [p] and [b], are described by the same lip movements and so the same video parameters. Thus, two different LPC filters cannot be estimated from the same video coefficients : there are ambiguities. To avoid this kind of error, we must transmit some audio parameters with the video coefficients together. It permits to discriminate the " identical visual cases ".

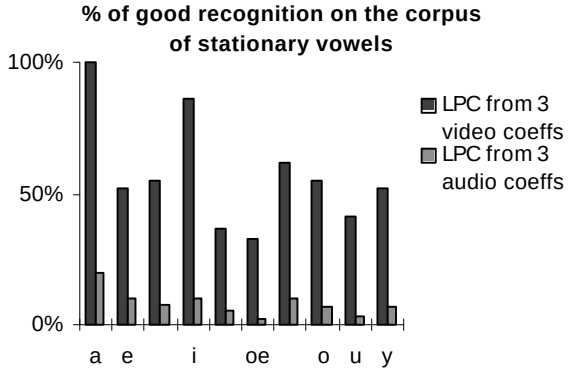

Figure 2

Concerning the second test, which consists to keep the quality of coding unchanged compared to a classic

vocoder while reducing the bit rate, the results obtained with the corpus of stationary vowels appear on figure 3.

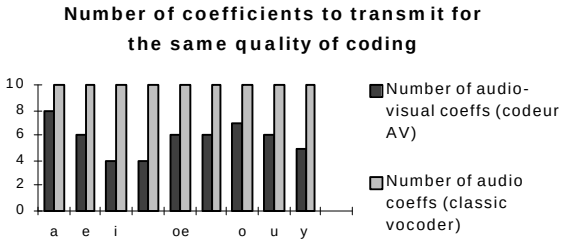

Figure 3

From the tests done on the corpus of vowel-consonant transitions, we can conclude that, over and above the visual parameters, four audio parameters are needed to be transmitted to reach the same quality of coding as a classic vocoder.

The transmission rate can be expressed as a function of the number of audio-visual coefficients transmitted to estimate the LPC filter. Such a description appears on figure 4. For example, instead of transmitting 10 LPC coefficients as for a classic vocoder, 4 audio coefficients and 2 visual parameters i.e. 6 audio-visual coefficients are necessary to be transmitted for the audio-visual coder while keeping the quality unchanged. In this case, taken in account the bit allocation of the different parameters, the use of audio-visual parameters reduces the bit rate of 20%.

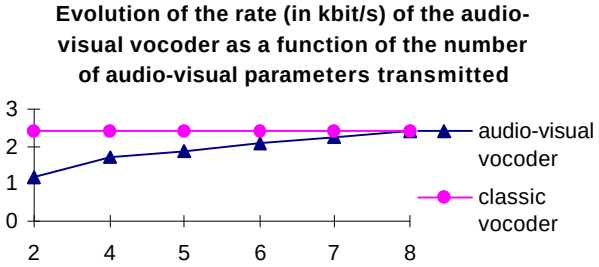

Figure 4

6 CONCLUSION AND PERSPECTIVES

We have shown, during these experiences, that there is evidently some information in terms of intelligibility in visual parameters, and that the bimodality of speech has to be exploited in the coding systems, to decrease the transmission rate of a coder chosen as reference.

The innovating aspect of this project involves prudence and such simple studies to evaluate first the feasibility and the efficiency of audio-visual systems. Thus, these results are still preliminaries but they confirm the interest of an audio-visual coding system and the use of the bimodality of speech signal.

The perspectives concern the methods to build the audio-visual associators and particularly the use of neural networks which should permit to realise non-linear associations. They could be more adapted to describe the audio space as a function of the video space, for example when a linear audio/video relation does not exist. A first test has been performed in this field with non-linear associator and the results are encouraging [3].

Finally, it would be interesting to study the suppression of the redundancy between some audio parameters and video coefficients to make our system as efficient as possible. Vector quantization could be a solution to this problem [1], by exploiting the correlation between audio and video in order to reduce efficiently the bit rate with the same quality of coding.

7 REFERENCES

[1] GERSHO A., GRAY R.M. (1992), Vector Quantization and Signal Compression, Kluwer academic Publishers.

[2] GIRIN L., FENG G., SCHWARTZ JL., (1998), “ Fusion of auditory and visual information for noisy speech enhancement : a preliminary study of vowel transitions”, Proc. ICASSP’98, Seattle, USA.

[3] GIRIN L., VARIN L., FENG G., SCHWARTZ JL., (1998), “Débruitage audiovisuel de parole : apport d’une association vidéo-audio non-linéaire par des réseaux de neurones.”, CORESA 1998.

[4] LALLOUACHE M.T. (1991), “ Un poste "visage-parole" couleur. Acquisition et traitement automatique des contours des lèvres ”, Thèse de doctorat, Institut National Polytechnique de Grenoble.

[5] MARKEL J.D., GRAY A.H. Jr (1976), Linear Prediction of Speech, Springer-Verlag, NY.

[6] ROBERT-RIBES J. (1995), Modèles d'intégration audiovisuelle de signaux linguistiques : de la perception à la reconnaissance automatique des voyelles ”, Thèse doctorale, INPG, Grenoble.

[7] ADJOUDANI A. (1997), “ Reconnaissance automatique de la parole audiovisuelle, Stratégies d'intégration et réalisation du LIPTRACK, labiomètre temps-réel ”, Thèse doctorale, INPG, Grenoble.