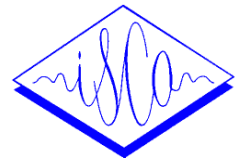


Auditory-Visual Speech Processing
(AVSP'98)
Terrigal – Sydney, Australia
December 4-6, 1998

AUDIOVISUAL SPEECH CODER : USING VECTOR QUANTIZATION TO EXPLOIT THE AUDIO/VIDEO CORRELATION

Elodie FOUCHER, Laurent GIRIN & Gang FENG

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

ABSTRACT

Visual information can help listeners to better understand what is said. In the speech coding domain, it will be shown that it allows to reduce the transmission rate of a classic vocoder (1,9 kbit/s instead of 2,4 kbit/s) by estimating audio parameters from video ones. In addition, vector quantization seems to be a good method to reduce the redundancy between some audio and visual coefficients. With the vector quantization, we can reduce again the bit rate while decreasing the quantization error.

1. INTRODUCTION

Speech is multimodal : it is both auditory and visual. Thus, seeing the speaker's face can help the listeners to better understand the message. The lip movements are strongly correlated with the acoustical signal, and audio and video information are complementary. The visual information, in particular the lip shape, has already been exploited in different fields, as speech recognition and noisy speech enhancement, in which significant improvements have been obtained [1] [2] [3]. 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.

In this paper, the original study of an audio-visual vocoder is presented which exploits visual information to reduce the transmission rate of a classic vocoder. In a first part, the audio-visual coder is described in more details and the first results of its efficiency [4]. In a second part, a method to use the correlation of some audio and video parameters is exposed, using vector quantization, to reduce the bit rate.

In order to illustrate the contribution of visual information, work on a very low bit rate coder, which privileges the aim of intelligibility of the transmitted speech signal instead of its quality, appeared us to be a good solution. 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, a vocoder based on linear prediction coding (LPC) with a bit rate of 2,4 kbit/s is used.

2. THE AUDIO-VISUAL VOCODER

In a first study [4], an audio-visual vocoder has been proposed, using information linked to the lip shape movements. It is described here succinctly.

2.1. Structure

The proposed audio-visual 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. As for the parameters concerning the speaker's lips movement, a face processing system developed at ICP [6] is used, 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 selection of audio

parameters LPC of the filter 1/A(z), selection destined 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 synthesized by filtering of the excitation signal through 1/A(z).

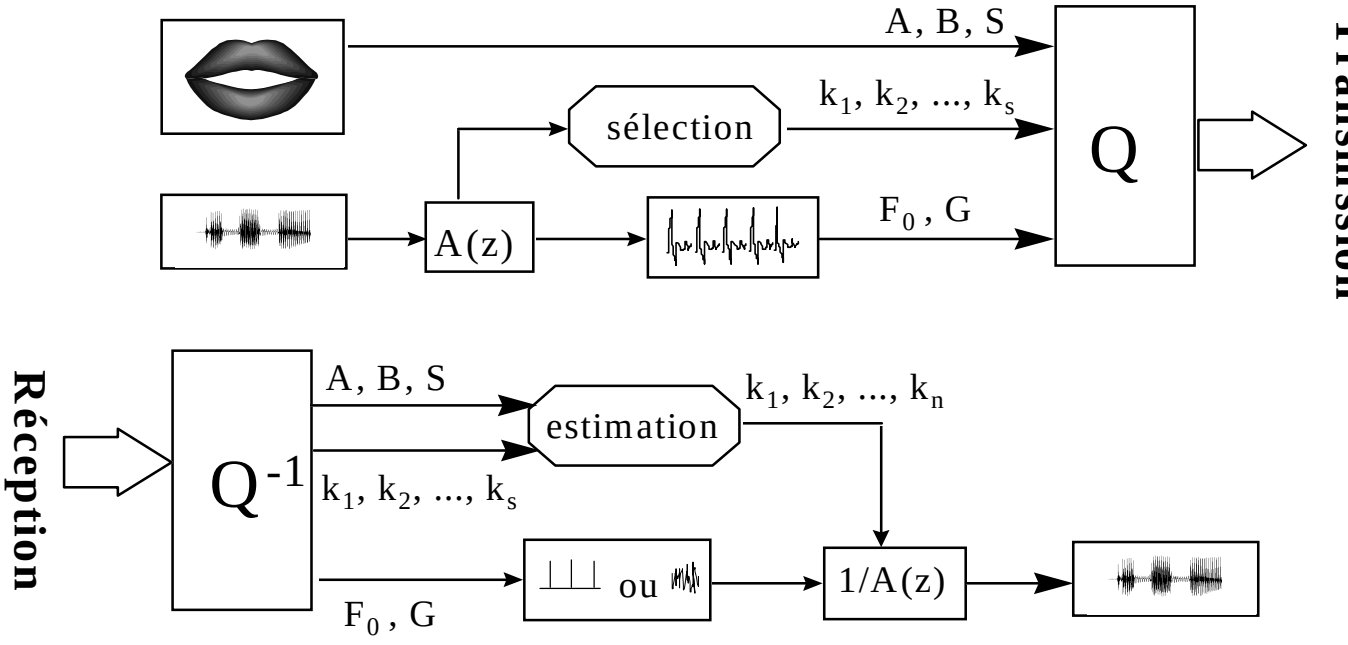


Figure 1 : Principle of the audio-visual vocoder

2.2. Estimation of 1/A(z)

In the decoding phase, we have to estimate the filter 1/A(z) from the transmitted coefficients, i.e. a vector of video coefficients (V) and a vector of a selection of audio coefficients (A1). The concatenation of the video vector (V) and the audio one (A1) in an audio-visual vector is noted AV=[A1 V]. We aim, from this multimodal vector, to estimate the vector of audio and non transmitted coefficients called A2 and which, associated to A1, will permit to determine completely the synthesis filter. To solve this problem, linear regression method is used by its simplicity and efficiency concerning our application : the product between AV and a matrix M leads to an estimation of the vector A2. M is determined during a learning phase by using the linear regression method between two matrix built by concatenation of vectors of both spaces "non-transmitted audio" and "transmitted audio + video".

2.3. First results

In order to evaluate the efficiency of the proposed audio-visual coder, a corpus containing stationary vowels and a corpus containing vowel-consonant transitions are used. Two kinds of tests have been performed [4]. It has been shown that, with the same number of transmitted coefficients, three video parameters allow to rebuild an intelligible signal whereas three audio coefficients cannot. Also, it has been shown that the use of audio-visual coefficients permits to reduce the bit rate of a classic vocoder with the quality of the coding unchanged. In fact, only two video parameters and four audio coefficients permit to code speech signals as well as ten audio coefficients do it in the case of a classic vocoder. The reduction of bit rate can reach 20 %, concerning the two corpus used.

This first study clearly shows the gain of the video information for speech coding systems and the use of the bimodality of speech signal in a such field.

3. CODER BASED UPON VECTOR QUANTIZATION

Two kinds of information, audio and video, enter separately in the coding system. They have to be encoded together and restored simultaneously after the decoding phase. Thus, we attempt to use the bimodality of speech signal, in particular the correlation between some audio and video coefficients, in the coding step, to reduce the bit rate. Working on the quantization block and using vector quantization seemed us to be a good method to exploit as well the redundancy between some of the parameters. This kind of quantization is applied to audio-visual vectors. It aims to determine a *codebook* of vectors able to replace real audio-visual data during the coding phase (the index of the datum is stocked / transmitted rather than the datum itself) from an optimal cartography reflecting the data distribution. This cartography should permit to profit by the redundancy and the natural complementarity between audio and video information.
A learning phase is needed to build the codebook of vectors using the LBG algorithm (Linde Buzo and Gray) [7]. First the codebook is initialised thanks to the *splitting step* and then optimised with the Lloyd algorithm [7]. Then, to code audio-visual data by a vector of the codebook, the nearest neighbour is calculated according to a definition of distance.

3.1. Introduction of an audiovisual distance

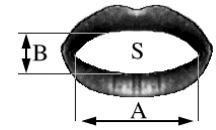
An audio-visual distance has to be defined, on which the built of the codebook and the quantization itself depends.
This distance must be adapted to the kind of coefficients, that's why the Itakura distance [8] is associated to the audio parameters and the squared distance to the video coefficients.

$$d_A = d_{ITAKURA} = \frac{a'^T Ra'}{a^T Ra}$$

aR model of reference associated to $A(z)$
$a'R'$ model associated to $A'(z)$

$$d_V = (A - A')^2 + (B - B')^2$$

with A, B and A', B', visual parameters of two different sounds defined as following :



These both distances are weighted to give the audio-visual distance :

$$d_{AV} = \alpha d_A + (1 - \alpha) d_V$$

where α is a coefficient to determine to minimise the global audio-visual quantization error.

3.2. First results

In order to test the efficiency of the vector quantization and to valid the audio-visual distance, different tests have been performed on the corpus of vocalic transitions such as V1V2V1, where V1 and V2 are vowels issues from the group [a i o u]. We aim to show that vector quantization of audio-visual parameters is more efficient than vector quantization of audio parameters and vector quantization of video parameters separately in terms of transmission rate and even in terms of quality (decrease of the quantization error).

First, audio and video have been quantized separately, respectively on 5 bits (32 different spectrum) and 4 bits (16 lip shapes). Then, the audio-visual data are quantized on 6 bits for $\alpha = 0$ to 1. The different quantization errors are shown in figure 2.
We can see that the audio-visual quantization is better than the audio and video quantizations joint in terms of transmission rate (6 bits for audio-visual instead of 4 bits for video plus 5 bits for audio = 9 bits i.e. 30% less) and in terms of quality (quantization error smaller in each field of video and audio for some values of α) .

In order to determine more precisely a value of the parameter α, a small part of the figure 2 is shown in figure 3.
Vector quantization appears to be a good method to exploit efficiently the correlation between some audio and video parameters in order to increase the quality of quantization (quantization error smaller) and to decrease the transmission rate.

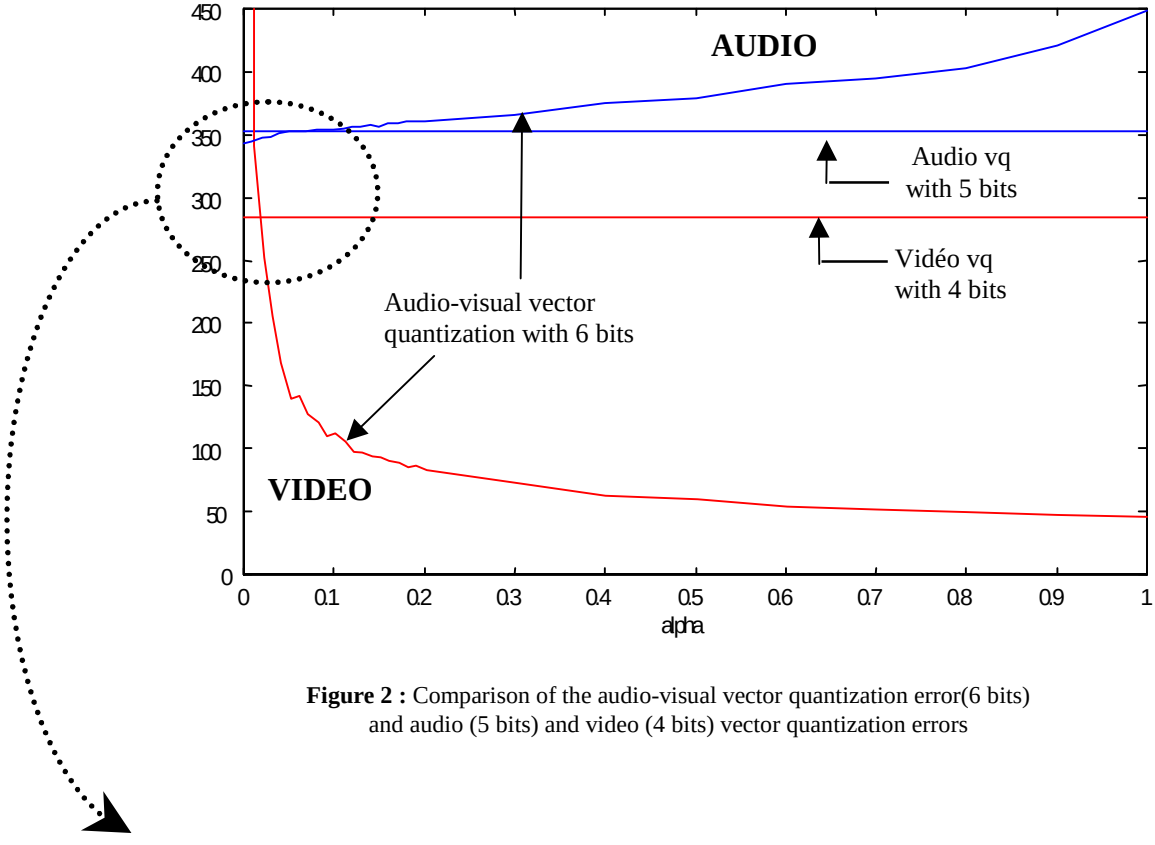


Figure 2 : Comparison of the audio-visual vector quantization error(6 bits) and audio (5 bits) and video (4 bits) vector quantization errors

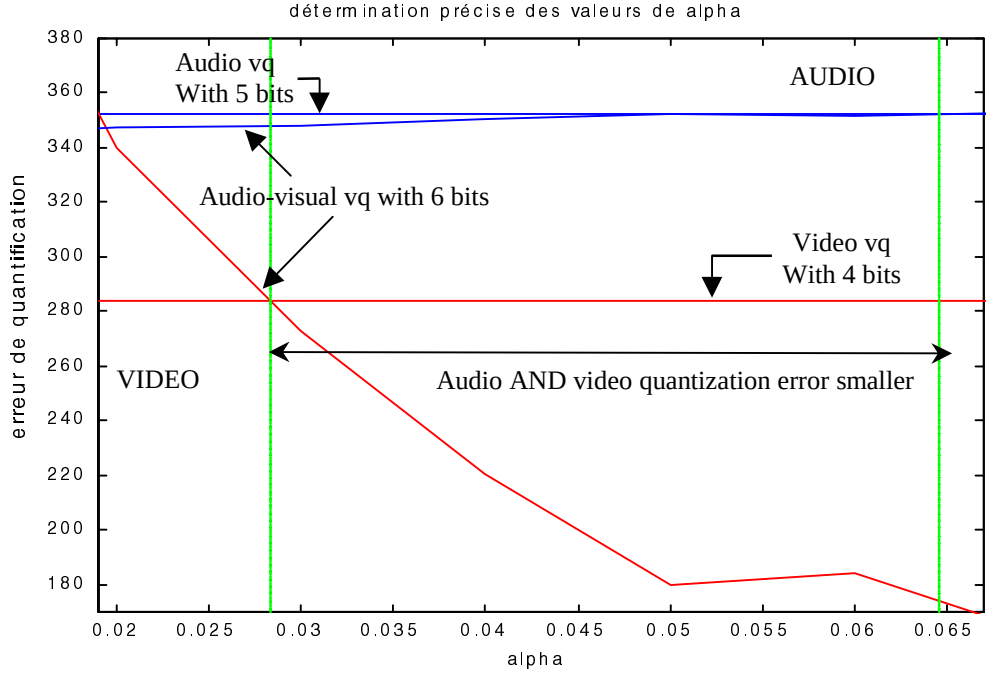


Figure 3 : Quantization error (zoom of the figure 2)

To prove that this diminution of the quantization error does not depend on the bigger size of the coded vectors but is really due to the correlation between some of audio and video parameters, mixed audio-visual data are quantized : audio and video coefficients do not correspond to the same frame of speech signal. The results are shown in figure 4 :

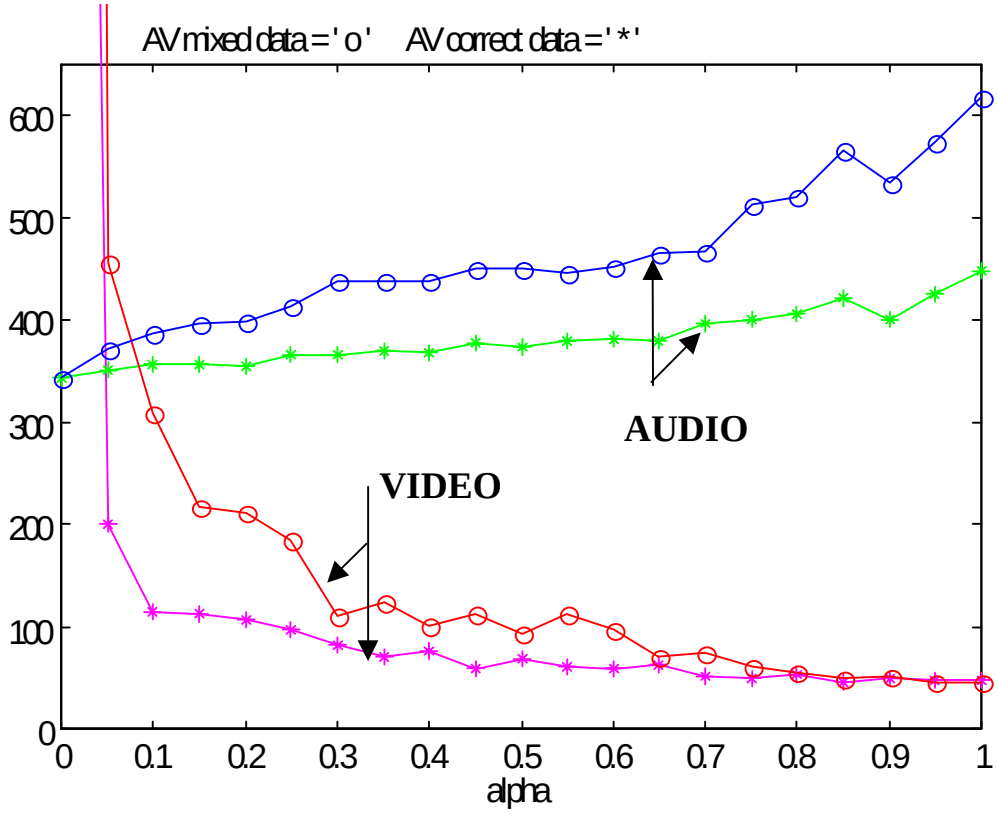


Figure 4 : Quantization error for AV mixed data and AV correct data

The quantization error is bigger for audio and video parameters when the audio-visual data are mixed : that shows clearly that the correlation between audio and video coefficients exists and is well exploited by this kind of quantization.

4. CONCLUSION

This first study shows clearly the interest of an audio-visual coding system.

First, we can see, thanks to the results, that there is evidently some information in terms of intelligibility in visual parameters, and that the bimodality of speech has to be exploited to decrease the transmission rate of a coder chosen as reference.

Then, vector quantization appears to be a good method to use efficiently the redundancy of audio-visual parameters in order to reduce the bit rate and even to increase the quality of the coded speech signal.

Other tests should be done, particularly on the coefficient α and with different allocations of bit to code audio-visual parameters.

This method should be tested on other corpus and may permit to quantize precisely the correlation between audio and video parameters.

Finally, other kinds of visual distance are studied to find the one which permits to better code the video coefficients according to their own importance and

stability [9] and according to the quantization error permitted.

5. REFERENCES

1. Girin L., Feng G., Schwartz JL, " Fusion of auditory and visual information for noisy speech enhancement : a preliminary study of vowel transitions", *Proc. ICASSP'98*, Seattle, USA, 1988.

2. Adjoudani A., " Reconnaissance automatique de la parole audiovisuelle, Stratégies d'intégration et réalisation du LIPTRACK, labiomètre temps-réel ", Thèse doctorale, INPG, Grenoble, 1997.

3. Robert-Ribes J., Modèles d'intégration audiovisuelle de signaux linguistiques : de la perception à la reconnaissance automatique des voyelles ", Thèse doctorale, INPG, Grenoble, 1995.

4. Foucher E., Feng G., Girin L.," Une étude préliminaire de la réduction du débit d'un vocodeur LPC par utilisation de paramètres vidéo ", *JEP 1998.*

5. Markel J.D., Gray A.H. Jr, *Linear Prediction of Speech*, Springer-Verlag, New-York, 1976.

6. Lallouache M.T., " Un poste "visage-parole" couleur. Acquisition et traitement automatique des contours des lèvres ", Thèse de doctorat, Institut National Polytechnique de Grenoble, 1991.

7. Linde Y., Buzo A., Gray R.M., "An algorithm for vector quantizer Design", *IEEE Trans. On Comm*, vol. COM-28, no.1, pp. 84-95,1989.

8. Gray A.H., Markel J.D., "Distance measures for speech processing", *IEEE Transactions on Acoustics Speech and Signal Processing*, vol assp-24, NO.5, october 1976.

9. Le Goff B., Guiard-Marigny T., Benoit C., "Read my ips...and my jaw ! How intelligible are the components of a speaker's face ? ", *Proc. ot the 4th European Conf. on speech communication and technology*, Madrid, Spain, pp 291-294, 1995.