Auditory-Visual Speech Processing
2005 (AVSP'05)
British Columbia, Canada
July 24-27, 2005

A CODING METHOD FOR VISUAL TELEPHONY SEQUENCES

*Edson Bárcenas, Mauricio Díaz, Rafael Carrillo, Ricardo Solano, Carolina Soto,
Luis Valderrama , Javier Villegas and Pedro Vizcaya*

Pontificia Universidad Javeriana
Bogotá-Colombia

ABSTRACT

Usually the design of a vector quantizer involves the minimization of a distortion measure such as the MSE. In this paper we present a new paradigm applied to the design of a visual telephony coding scheme: the synthesis of credible image sequences. In other words, the creation of smooth and coherent transitions between the images to be reproduced. This new paradigm requires the redefinition of the samples representative of each class in the Lloyd-Max algorithm. In our case the design criterion is the minimization of the maximum error within the class samples and their representative. The results obtained from the proposed method are compared to those obtained from the Lloyd-Max original algorithm.

Video transmission over asynchronous networks without real time control frequently suffers from information losses that can cause the loss of entire images. The present paper introduces a method based on the interpolation of the received images to estimate the lost images. The interpolation uses the search of credible image sequences by means of the Viterbi algorithm and a modified sequence distance.

1. INTRODUCTION

Current videoconference systems needs large bandwidth to transmit the information. It has been shown that in multimodal communication the audio speech is perceptually more important than the visual information [1]. However the video teleconferencing standards, such as MPEG 1/2/4 [1], uses more channel capacity to transmit the video than to transmit the audio. In this paper we present a coding method where the information capacity used by the video is less than the one used by the audio, being coherent with the human perceptual system [3]. Two algorithms are used in the coding method: an image vector quantizer and a video sequence interpolation system.

The classic paradigm in the design of a vector quantizer is to maximize the signal to noise ratio [4]. The set of images that maximize the PSNR is selected from the training video and each new image is represented using the minimum distance criterion. The process described bellow is the Lloyd Max algorithm [5][6] which has two iterative steps: (1) the representative image is selected as the momentum center (centroide) of the images subset within class, and (2) each image of the training set is classified using the minimum distance to the centroides. The Lloyd-Max algorithm assigns more centers to the regions of the space containing more samples, because these regions have higher probability of occurrence. The atypical images, which have low occurrence probability, are not well represented by the codebook set. In this paper we present a new paradigm applied to the design of a visual telephony coding scheme: the synthesis of credible image sequences including atypical images in the codebook. The coding scheme will be explained in section 2.

An improvement to the coding process is to reduce the amount of information to be transmitted. We did some experiments which show that it is possible to reconstruct a visual speech sequence sampling the original one and then interpolating it. The improvement implemented in this work consists in the development of an interpolation algorithm using the Viterbi search. The paradigm of the interpolation system is the construction of credible sequences, which means generating sequences with smooth transitions. This algorithm could be useful in other applications such as visual speech synthesis from text or from speech [7]-[10]. In section 3 we will present a detailed description of the interpolation method. Finally sections 4 and 5 present some experimental results and concluding remarks.

2. CODER TRAINING

The coder design is a process that could be realized in two steps: the first one is the parameterization of the images; the objective of this process is to reduce the signal space dimension and so to reduce the computational complexity of the problem. The second step is the classifier training consisting in the selection of a subset of representative images, or codebook. The training step is done in the parameter domain. The codebook is chosen from a complete video of the visual telephony scene of about half a minute. A complete video is one that contains all the visemes of the language and some transitions between them.

In this paper we present a different method to design the codebook of a visual speech synthesis classifier. The basic criterion is to include atypical images in the codebook and discard similar images in order to generate credible sequences and maintain a reduced codebook size. On the other hand, the Lloyd-Max algorithm includes similar images as representatives. In the context of this work, credible sequences are sequences with smooth transitions between them, specifically in the mouth region [9],[11].

In order to achieve smooth sequences the optimization condition is to minimize the maximum error within the class samples and their representative image. This is done using a modified version of the Lloyd-Max algorithm changing the definition of the representative sample.

2.1 Modifications to Lloyd-Max algorithm

The new center could be defined as the geometrical center of the minimum volume solid that contains every sample in the class. The coherence between the calculation of the representative (center of the minimum volume solid) and the distance measure is needed to guarantee the convergence of the algorithm to the expected limit, which is to minimize the maximum error within classes. The distance measure used defines the form of the solid whose surface contains all the points with the same distance to the center [12]. Here we briefly analyze three different distance measures with their corresponding solids.

The surface of an L-dimensional cube is the region that contains all the points with the same L_∞ distance to its center. The calculus of the center is straightforward because the side length of the solid is the distance between the furthest samples and the calculus of the solid center in each dimension is simple. The solid defined using the Euclidean distance, L_2, is an L-dimensional sphere. The calculus of the L-dimensional sphere with minimum radius that contains all the points is an open problem and it depends on the vector space dimension.

The Manhattan distance, L_1, defines an L-dimensional rhombus as the solid whose surface contains all the points with the same L_1 distance to the center. To determine the volume of the solid the distance between the furthest samples in the subset is calculated and to determine the center, the rhombus is moved iteratively until it includes all the samples in the subset. A special characteristic obtained using the L_1 distance is that the volume does not change along the iterations.

The computer complexity of the calculation of the solids and their center is high, so that, a sub optimal but less complex algorithm was implemented. The algorithm, called for us the *minmax* algorithm is described in the next.

2.2 *Minmax* algorithm

Let $X=\{\mathbf{x}_n\}_{n\in 1,\dots,N}$ be the training set, where all the samples are vectors in a M-dimensional space

$$\mathbf{x}_n=\left(x_n(1),\dots,x_n(m),\dots,x_n(M)\right),\ \mathbf{x}_n\in\Re^M;$$

let $Y=\{\mathbf{y}_k\}_{k\in 1,2,\dots,K}$ be the codebook set with K classes; let $Y_k=\{\mathbf{x}_{k,i}\}_{i\in 1,2,\dots,N_k}$ be the subset of N_k samples in the training set assigned to the *k-th* class (all samples nearest to $\mathbf{y}_k$), where $\mathbf{x}_{k,i}\in X$, $X=\bigcup_{k=1}^{K}Y_k$ and $Y_k\cap Y_l=\varnothing$ for each $k\neq l$, the set $\{Y_k\}_{k\in 1,\dots,K}$ is a partition of X.

The conditions of the quantizer are the following:

1. The new center, $\mathbf{y}_k$ is defined as:

$$\mathbf{y}_k=\arg\min_{\mathbf{x}_i\in Y_k}\left\{\max_{\mathbf{x}_j\in Y_k}\left[d_{L_1}\left(\mathbf{x}_i,\mathbf{x}_j\right)\right]\right\}.$$

2. The *i-th* class defined as :

$$Y_i=\left\{\mathbf{x}_n\in X\,\middle|\,d_{L_1}(\mathbf{x}_n,\mathbf{y}_i)=\min_{k\in 1,\dots,K}\left\{d_{L_1}(\mathbf{x}_n,\mathbf{y}_k)\right\}\right\},$$

$$\text{with } d_{L_1}(\mathbf{x},\mathbf{y})=\sum_{i=1}^{M}\left|x(i)-y(i)\right|.$$

The algorithm described bellow uses the L_1 distance but it could be used with any other distance. The L1 distance achieves better results in visual speech re-synthesis than the $L\infty$ and L_2. The *minmax* algorithm is sub-optimal because when the number of samples is small the algorithm does not yield the optimal solution and when the training set is large the solutions is optimal but time consuming.

3. INTERPOLATION AND SELECTION OF INTERMEDIATE IMAGES

A typical approach to create intermediate images is to calculate a direct linear interpolation, pixel by pixel, between the two target images. However this method generates new images that are not real, e.g. in the **Figure 1** it is shown a lineal interpolation of four images between an open and a closed mouth image: two of them clearly show both mouth positions, which is unrealistic.

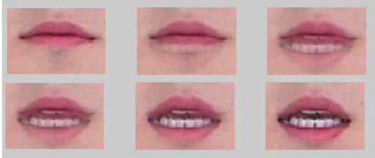


Figure 1. Image sequence generated using a linear interpolation pixel by pixel between two images.

Other approach to the interpolation is to use a sigmoidal pixel by pixel interpolation function. In this kind of interpolation the intermediate images are more real than with the previous method. An example of sigmoidal interpolation using the hyperbolic tangent function is shown in **Figure 2**.

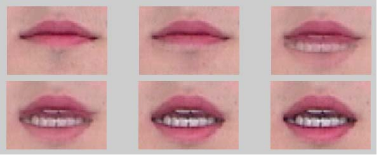


Figure 2. Interpolation between two images using a sigmoidal interpolation pixel by pixel.

Other technique to generate intermediate images is optical flow, which achieve better results than the previous methods but its computational complexity is much higher[8],[13].

A new interpolation method is presented in this work. The basic difference between this method and the previous ones is that the intermediate images are looked for in a database of images instead of generating new images. The main idea is to find the images that create a credible sequence between a couple of images. A modified version of the Viterbi algorithm is used to find appropriate images and a modification to the sequence distance is proposed here to assure an improvement in the credibility of the sequences.

From the study of natural sequences of different people and synthesized videos that seemed non credible, it was concluded that the credibility of a visual speech sequence is related to its smoothness.

The purpose of the modification to the distance measure is to help in the selection of smother sequences. The change applied to the distance measure is explained in the following paragraphs with an example.

In **Figure 3** is shown a problem of finding the smoothest sequence between a couple of images. The nodes in the graph represent a set of images in the database and the arcs between them represent the distance between images. The total distance between two original images is defined as the sum of the individual distances of the intermediate steps (nodes) used to rich the final one. The graph shows three different paths between image 1 and 5, each one with three intermediate images. The goal is to select the most credible sequence between them, i.e., the smoothest path.

In order to choose the smoothest path we look for the path with minimum distance, in this case the second path, that has a total distance of 5 (1st path 6, 2th path 8). However, the second path is not a smooth path, because it contains a big difference between the images 8 and 5.

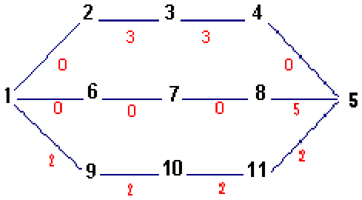


Figure 3.Example of the distance measure of sequences for the selection of smoother sequences

If the distance between images is raised to the power p the selection of the smoothest sequence could change. The **Table 1** contains the sequence distances for p=1.5, p=2 and p=10 and the path selected as smoothest path in red.

	p=1	p=1.5	p=2	p=10
1st path	6	10.4	18	118098
2th path	5	11.2	25	9765625
3th path	8	11.3	16	4096

Table 1 Example of the influence of the p factor in the selection of smoother sequences

It can be seen in this example that the variation of the power factor changes the path selected as shortest path. The smoothest sequence is selected as the larger power factor is used. The power factor apparently depends on the number of intermediate images that will be generated and the availability of images in the database. With an appropriate factor smooth sequences are observed. The power factor is heuristically adjusted to obtain credible sequences. In our work we have found that this factor should be proportional to the number of intermediate images.

Formally the proposed algorithm is the following:

Let $I = \{i_n\}_{n\in 1,\dots,M}$ be a set of images (data-base), $i_n, i_m \in I$ the pair of target images, $s^{(k)}$ the *k-th* possible sequence of *N* images *(N<M)* between i_n and i_m and $D(s^{(k)})$, the weight of the *k-th* sequence

$$D(s^{(k)}) = \left[d(i_n, s_1^{(k)})\right]^p + \sum_{l=1}^{N-1}\left[d(s_l^{(k)}, s_{l+1}^{(k)})\right]^p + \left[d(s_N^{(k)}, i_m)\right]^p$$

where *p* is the power factor and $s_l^{(k)}$ is the *l-th* image in $s^{(k)}$. We select the smoothest sequence as the sequence with the minimum weight.

$$s^{(opt)} = \underset{s^{(k)},k=1,\dots,M^{N-1}}{\arg\min}\left\{D(s^{(k)})\right\}$$

The modification of the sequence distance assures that smooth sequences will be obtained. The problem of finding a sequence of N images, in a database of M images, between two images has M^{N-1} possible solutions. The direct evaluation of all possible sequences is time consuming, however the Viterbi search is a computational efficient method to perform this search. The Viterbi algorithm is a dynamic programming technique to find the optimal path between two nodes in a graph in an efficient way [14].

4. RESULTS

The results are presented in two main parts. The first one contains the results obtained with the selection of the codebook using the *minmax* criterion and the second part contains the results of the interpolation method.

4.1 Codebook Selection

In this section a couple of examples of the vector quantizer design with the *minmax* criterion applied to the Lloyd-Max algorithm using L_1 distance are shown. The first example is made with data generated and the second is made with images.

The vector quantizer was trained with 1000 samples, 200 from distribution f_1 and 800 from distribution f_2. $f_i(\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}_i, \mathbf{C}_i)$, $f_i(\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}_i, \mathbf{C}_i)$ where

$\boldsymbol{\mu}_1 = (-5, 2)$, $\mathbf{C}_1 = 1.6 * I$, $\boldsymbol{\mu}_2 = (1, -3)$ and $\mathbf{C}_2 = 5 * I$

The training of the vector quantizer was made using the original Lloyd-Max algorithm with the distance L_2 and using the momentum center of the samples in the class as the class representative, the results of this simulations are in the Figure 4. The vector quantizer was also trained using the *minmax* center and the L1 distance, the results of this process are in the Figure 5. In both cases the classifier has 4 classes.

It can be seen in Figure 4 and Figure 5 that in the Lloyd-Max algorithm using the *minmax* criterion the distance between class centers is larger than the obtained with classic definition of the center. The numerical results for the distance L_1 and L_2 are in Table 2 .

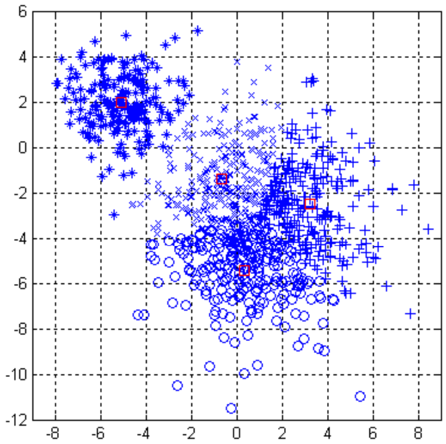


Figure 4. Clustering with the classical Lloyd-Max algorithm

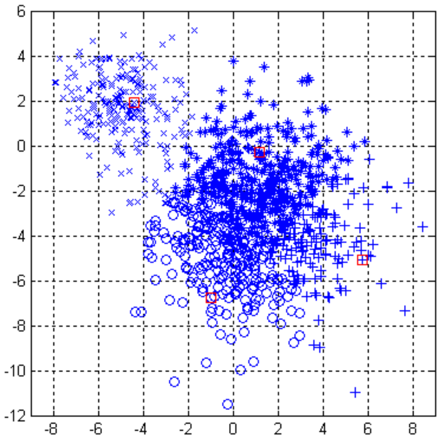


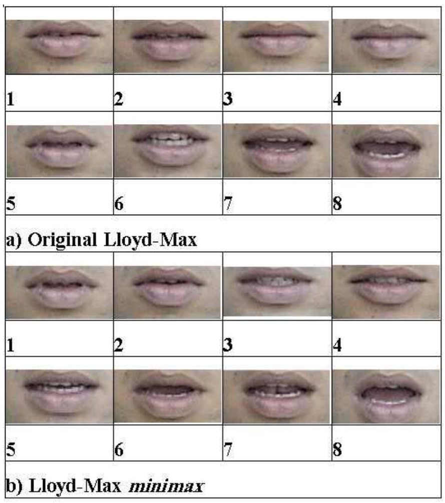
Figure 5 Clustering with the Lloyd-Max *minimax.*

	L1 Original	L1 minmax	L2 Original	L2 minmax
Class 1	6.8954	5.9235	5.2735	5.0235
Class 2	6.5143	5.5341	4.9701	4.7439
Class 3	9.3289	6.2328	6.6044	5.9340
Class 4	10.6798	5.9515	7.5576	4.7635

Table 2 Maximum L_1 and L_2Distance within Classes.

Table 2 shows the results of the *minmax* criterion compared to those obtained with the original Lloyd-Max algorithm. In Table 2 it can be seen how the maximum L_1 and L_2distance within samples and the class representative is larger for the original Lloyd-Max algorithm and smaller for the proposed *minmax* algorithm..

The second experiment has as training samples a set of 1000 images from the mouth region with a resolution of 64 X 128 pixeles. The images are transformed to a lower dimensional space using a decimation of 4 in the two dimensions, this is done applying a low pass filter and then sampling the original image to obtain a 16 X 32 new image. The DCT of the sampled image is calculated and the first 8 X 16 coefficients are taken. In this experiment the codebook is a set of 8 images. In the two methods to train the classifier are compared when the input samples are the set described bellow.



a) Original Lloyd-Max

b) Lloyd-Max *minimax*

Figure 6 Image code-books .

In the Figure 6(a) it can be seen that most of the codebook images obtained from the original Lloyd-Max are closed mouth (image 1-image 5), a common image in the database. In the Lloyd-Max modified algorithm the codebook images are more different between them and the codebook only has one closed mouth image.

The method used in the second experiment is being used at the moment in a visual telephony system having a database of around 1000 images and a codebook size of 256.

4.2 Image Interpolation

The results shown in this section were obtained using a natural visual speech sequence of 856 images in AVI format without compression, 720 X 480 pixels resolution and 24-bits RGB color. Half of the video was used as a data-base and the other half was used as test sequence.

In order to reduce the signal dimension the mouth region was segmented using a rectangle of 75 X 110 pixels. The mouth region was transformed with a Bi-dimensional DCT and the first hundred coefficients were kept.

In the Figure 7 is shown a sequence between a closed mouth and an open mouth images. The sequence was generated with a power p equal to 2.

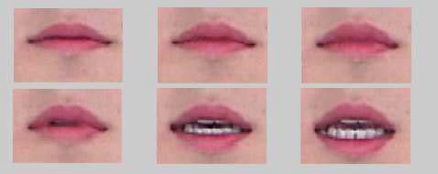


Figure 7 Interpolation between two images using the proposed algorithm with p=2

In order to evaluate the interpolation method performance it was made an experiment consisting in the simulation of the periodic loss of images, the reconstruction of the sequence and the measure of the distortion between the videos.

The periodic loss of images was simulated using a decimation process

$$y[n] = x[kn]$$

where k is the decimation factor, *x[n]* is the input video sequence and *y[n]* is the output sequence.

The reconstruction of the sequences is done using the interpolation algorithm, which search the *k-1* intermediate images in the data-base. The interpolation was done with different p factors.

As a distortion measure it was defined, the quality of reconstruction index

$$iqor = 10\log\left(\frac{1}{N}\sum_{i=1}^{N}\frac{\operatorname{var}(x_i)}{\operatorname{var}(x_i - y_i)}\right) \text{ (dB)} \quad ,$$

where N is the number of reconstructed images in the video sequence, x_i is the *i-th* lost image in the original sequence and y_i is the image in the reconstructed sequence that represents x_i.

Although the *iqor* is not related with the credibility of the synthetic sequence, it shows the results of the system in terms of traditional distortion measures.

The results of the experiment for different decimation factors are shown in Figure 8. As it was expected the higher *iqor* are obtained with low decimation factors. However, the subjective quality of the videos with a decimation factor of 4 is still good. In the Figure 8 it also can be seen the influence of the *p* factor in the quality of reconstruction index, where the maximum *iqors* were obtained with *p* near to the decimation factor *k*.

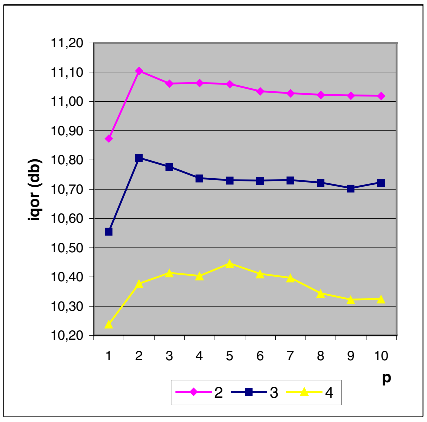


Figure 8. Quality of reconstruction with different decimation and power factors

5. CONCLUSIONS

In the present article has been shown that the proposed modification to the Lloyd-Max algorithm is useful to generate code-books with the criterion of minimize the maximum error, instead of minimize the MSE. The use of L_1 distance and the *minmax* criterion to calculate the class center are valid alternatives to design a vector quantizer and they allow efficient calculation of the representative sample. The codebook selected with this algorithm has less redundancy in the images, i.e. similar images are clustered all in the same class and the codebook is not biased by the frequency of the images.

In the article has also been shown an effective method to interpolate images searching them over a data-base, instead of generating novel images. The Viterbi algorithm working with a modified distance guarantees computational efficiency and soft transition between images leading to the creation of credible sequences.

In the modified distance the power factor *p* and the data-base quality determine the smoothness of the sequences. If the power *p* is increased up to infinite the algorithm will converge to the sequence with minimum distance between adjacent images using the images in the database. This could be useful in real time applications.

Acknowledgments

This work has been partially funded by COLCIENCIAS with contract RC. No. 35-2003.

6. REFERENCES

[1]. O'neill, J.J. "Contributions of the visual components of oral symbols to speech comprehension". In: Journal of Speech Hearing research, Vol 19, 1954.

[2]. ITU-T Recommendation H.263 Version 2 (H.263+). Video coding for low bitrate communication, January 1998.

[3]. Chen T y Rao R, "Audio-visual integration in multimodal communication", Proceedings of IEEE, Special Issue on Multimedia Signal Processing, 837-852, 1998.

[4]. A. Gersho and R. M. Gray, Vector Quantization and Signal Compression. Norwell, MA: Kluwer, 1992.

[5]. Linde, Y., A. Buzo, R. Gray, "An Algorithm for Vector Quantizer Design", IEEE T-COM, Vol .28, No. 1, pp. 84-95, January, 1980.

[6]. Nasrabadi, N., R. King, "Image Coding Using Vector Quantizatization: a Review", IEEE Trans.Commun., vol 36, No. 8, pp 957-971, august, 1988.

[7]. Ezzat, E., G. Geiger, T. Poggio, "Trainable Videorealistic Speech Animation", ACM Transactions on Graphics, 21(3): p. 388-398, 2002.

[8]. Bárcenas, E., Galán, J., Soto, C., Urbina, J., Vásquez, S., Vizcaya, P., (2001), "Visual speech synthesis in Spanish using an optical flow algorithm". In: Proceedings of IASTED International Conference on Visualization, Imaging and Image Processing, 577-583, 2001.

[9]. Machado, J., Vizcaya, P., Santa, D., "Visual Speech Synthesis Using A Real Time Parametric Approach", Memorias del VII Simposio de Tratamiento de Señales, Imágenes y Visión Artificial, pp. 104-109, Bucaramanga, Colombia, noviembre, 2002.

[10]. Beskow, J., "Talking heads – communication, articulation and animation", TMH-QPRS 2, Swedish Phonetics Conference, Nasslingen, 1996.

[11]. Soto, C., *Generación de Corpus para Síntesis de Voz Visual*, Master Thesis. Universidad Javeriana, 2004.

[12]. Duda R.O, Hart P.E. and Stork D.G., *Pattern Classification, 2nd ed.*, John Wiley & Sons, New York, 2001.

[13]. T. Ezzat and T. Poggio. "Visual speech synthesis by morphing visemes". In *MIT A.I Memo No. 1658*, May 1999.

[14]. Proakis, J., *Digital Communications*, McGraw-Hill, 4th edition, 2001.