

Objective Viseme Extraction and Audiovisual Uncertainty: Estimation Limits between Auditory and Visual Modes

Javier Melenchón, Jordi Simó, Germán Cobo, Elisa Martínez

Communications and Signal Theory Department
Enginyeria i Arquitectura La Salle, Universitat Ramon Llull
08002 Barcelona, Spain

{jmelen, jsimo, gcobo, elisa}@salle.url.edu

Abstract

An objective way to obtain consonant visemes for any given Spanish speaking person is proposed. Its face is recorded while speaking a balanced set of sentences and stored as an audio-visual sequence. Visual and auditory modes are segmented by allophones and a distance matrix is built to find visually similar perceived allophones. Results show high correlation with tedious subjective earlier evaluations regardless of being in English. In addition, estimation between modes is also studied, revealing a tradeoff between performances in both modes: given a set of auditory groups and another of visual ones for each grouping criteria, increasing the estimation performance of one mode is translated to decreasing that of the other one. Moreover, the tradeoff is very similar (< 7% between maximum and minimum values) in all observed examples.

Index Terms: Audiovisual processing, Viseme extraction, Auditory visual uncertainty

1. Introduction

It is well known that speech is experienced by humans from the beginning of their lives as a bimodal activity [1], i.e. with two related communication channels: the visual and the auditory ones. Predicting one mode from the other one has shown to be useful in lip reading activities [2] and applications like visual telephony for the hearing impaired [3]. It has been shown that visual mode can be estimated linearly from the auditory one with a precision of about 65% [4]. Nonlinear methods, as artificial neural networks and hidden markov models, have been proposed [5, 6] to account for the remaining 35%.

About the particular units of each mode, auditory data is grouped into allophones (similar auditory realizations of the minimum abstract meaning unit called phoneme) [7], while visual data is clustered into similar visual realizations of phonemes (actually, allophones), also known as homophonous sounds, visual phonemes or visemes [8]. The first viseme proposals emerged in the seventies [9] and eighties [10] and were obtained through long and tiring subjective testing processes. In fact, it has been shown that it is possible to specify different viseme groupings with different degrees of similarity between visual realizations [11]. Current research works are based on those tedious preliminary studies, like the viseme specification of MPEG-4 [12], which is based on that of [9].

This paper deals with auditory and visual data clustering. First, a new objective viseme grouping method using the Bhattacharyya distance [13] is proposed in section 2 to overcome the subjective testing difficulties of previous works and allow using natural speech when obtaining visemes. Second, an analysis of

different auditory and visual clustered data detailed in section 3 reveals that better data clustering in one mode is translated to a worse one in the other domain (section 4). Moreover, their tradeoff appears to be nearly constant (section 4.2)

2. Viseme grouping

A new viseme grouping method is stated in this section. It is based on simulating the experiments carried out by real people in preliminary works like [8, 9, 10]. Simplifying their procedures, tested people in those works were used as expert systems when classifying visual appearances of different phonemes. The resulting groups were supposed to be visually distinguishable among them but not inside them, i.e., each group consisted of visual appearances so similar that it was impossible to classify them into different groups. In this paper, it is proposed to change the expert measurement given by a person by an objective one obtained numerically by a computer. In order to achieve this aim, a particular codification of visual appearances is proposed in section 2.1 and a measure of distance $d(\cdot)$ between two visual appearances is provided and stated in section 2.2.

2.1. Data representation

Audiovisual sequences containing the face of different people (one at a time) speaking a set of balanced sentences in Spanish at f_{ps} frames per second and sampling frequency f_s are taken as input in this work. Synchronized audio and video channels were then extracted and processed separately for each person.

The codification proposed for the visual appearances is based on eigenspaces [14] and it is obtained with the principal component analysis (PCA) [15] of the vectorized set of aligned image mouth regions (simultaneously computed following [16]). Therefore, a vectorized mouth region \mathbf{m} of P pixels can be approximated by (1):

$$\mathbf{m}_{P \times 1} \approx \mathbf{U}_{P \times K} \mathbf{c}_{K \times 1} + \bar{\mathbf{m}}_{P \times 1} \quad (1)$$

where \mathbf{U} is an orthonormal matrix which identifies the eigenspace (its K columns are the largest K principal components of the given set of mouth images), $\bar{\mathbf{m}}$ is the mean mouth image and \mathbf{c} is the projection of $\mathbf{m} - \bar{\mathbf{m}}$ with respect to basis \mathbf{U} or video vector. Dimensionality reduction can be optimally obtained thanks to PCA, with $K \ll P$; particularly, K can be selected so the eigenspace accounts for a specific amount of singular value energy (85% in this work), obtaining high compression rates with low perceptual losses [14]. In this case, $K = 12$ was selected. Faster comparisons can be made using this compact visual representation of mouth images. Moreover, since visual redundancy is minimized through PCA, fewer dimensions

are obtained, allowing the computation of covariance matrices with fewer examples. Only $K + 1$ mouth images are necessary to obtain an estimation of its $K \times K$ covariance matrix Σ .

A labelling process was carried out over the auditory information in order to obtain the temporal labels t_n of the different uttered allophones. This task was automatically achieved using the HTK toolkit [17]. From t_n , video frames could be obtained as $i = \text{round}(t_n * f_{ps})$. Unfortunately, due to little random asynchronies given by the recording webcam, manual supervision of video channel was needed to avoid them.

Auditory information was windowed using audio frames of 20 ms, centering them at sample $s = \text{round}(t_n * f_s)$, where temporal labels are represented by t_l . Next, the audio frames were parameterized with 12 linear spectral frequencies (LSF), obtaining audio vectors \mathbf{a} . LSF were selected since they are closely related to the vocal tract geometry [18] and are also used in extended standards like GSM [19].

2.2. Distance between groups

In order to find the distance between different sets of visual appearances labelled with the same allophone, or visual sets, they must be quantified. It is proposed to approximate each set r with a multivariate normal distribution \mathcal{N}_r , i.e., providing it with a mean or centroid μ_r and a covariance matrix Σ_r . Let \mathbf{C}_r be a matrix which columns are the video vectors \mathbf{c} (see section 2.1) related to those mouth regions where the particular allophone r is uttered. Then μ_r and Σ_r can be computed as:

$$\mu_r = \frac{1}{N_r} \mathbf{C}_r \cdot \mathbf{1} \quad (2)$$

$$\Sigma_r = \frac{1}{N_r} (\mathbf{C}_r - \mu_r \cdot \mathbf{1}^T) (\mathbf{C}_r - \mu_r \cdot \mathbf{1}^T)^T \quad (3)$$

where $\mathbf{1}$ is a column vector of ones. Next, the Bhattacharyya distance [13] between normal distributions \mathcal{N}_r and \mathcal{N}_l (4) can then be used to find the similarity between sets of visual appearances related to allophones r and l :

$$B(\mathcal{N}_r, \mathcal{N}_l) = \frac{1}{8} B_1 + \frac{1}{2} B_2 \quad (4)$$

$$B_1 = (\mu_r - \mu_l)^T \left(\frac{\Sigma_r + \Sigma_l}{2} \right)^{-1} (\mu_r - \mu_l)$$

$$B_2 = \ln \frac{|\Sigma_r + \Sigma_l|}{\sqrt{|\Sigma_r| |\Sigma_l|}}$$

Therefore, a graph containing the distances among all visual sets can be defined and represented as a symmetric matrix \mathbf{D}_v , where $\mathbf{D}_v(r, l) = B(\mathcal{N}_r, \mathcal{N}_l)$ (see figure 1).

2.3. Grouping similar visual sets

The visual sets defined in section 2.2 can be grouped together using the similarity information stored in matrix \mathbf{D}_v . The similarity of visual set r to the other ones is represented by the r -th column of \mathbf{D}_v or \mathbf{d}_v^r . It can be said that given two similar visual sets \mathbf{d}_v^a and \mathbf{d}_v^b (5), if \mathbf{d}_v^a is similar or unsimilar to \mathbf{d}_v^c (6), so must be \mathbf{d}_v^b w.r.t. \mathbf{d}_v^c (7) (8):

$$|\mathbf{d}_v^a - \mathbf{d}_v^b| = |\mathbf{d}_v^b - \mathbf{d}_v^c| = \epsilon \quad (5)$$

$$|\mathbf{d}_v^a - \mathbf{d}_v^c| = |\mathbf{d}_v^c - \mathbf{d}_v^a| = S \quad (6)$$

$$|\mathbf{d}_v^b - \mathbf{d}_v^c| < |\mathbf{d}_v^b - \mathbf{d}_v^a| + |\mathbf{d}_v^a - \mathbf{d}_v^c| \Rightarrow |\mathbf{d}_v^b - \mathbf{d}_v^c| < S + \epsilon \quad (7)$$

$$|\mathbf{d}_v^a - \mathbf{d}_v^c| < |\mathbf{d}_v^a - \mathbf{d}_v^b| + |\mathbf{d}_v^b - \mathbf{d}_v^c| \Rightarrow |\mathbf{d}_v^b - \mathbf{d}_v^c| > S - \epsilon \quad (8)$$

Consequently, when obtaining G visemes, the columns of \mathbf{D}_v can be clustered into G classes, yielding similar visual sets into the same class. Since the resulting classes consist of similar visual appearances of allophones, they can be called visemes, following the original definition of [8].

3. Clustering

Once a set of visemes is found with a method like that of section 2.3, it is desired to know how well they can be estimated from auditory information. Bayesian estimation [20] is used in this work to obtain this goodness measurement (see section 3.3). As can be seen in the results of section 4, viseme estimation from audio data is far from being optimal, since auditory information is clustered following a visual criteria, e.g., visually different allophones [n] and [m] are acoustically similar, making it difficult to distinguish among them with auditory information. However, if allophones are clustered following an auditory criteria of similarity (see section 3.1), an estimation improvement is obtained (see section 4) for them. Different audio clusters are shown in figure 2 when grouped by visemes or sets of similar allophones. In addition, another grouping criteria is presented in section 3.2, which does not take into account the segmentation of section 2.1. The goodness measure defined in section 3.3 can be used to compare all clustering processes.

3.1. Audio clustering

The distance measurements of section 2.2 and the set grouping process of section 2.3 can be extended to auditory information with the redefinition of matrix \mathbf{C}_r to include audio vectors instead of video ones. Identities (2) and (3) can be used to obtain their respective normal distributions. Next, a symmetric matrix \mathbf{D}_a can be generated using the Bhattacharyya distance (4) among the resulting normal distributions (see figure 1) like in \mathbf{D}_v . Taking \mathbf{D}_a , the G groups of more similar allophones can then be found following the same process stated in section 2.3 for obtaining visemes.

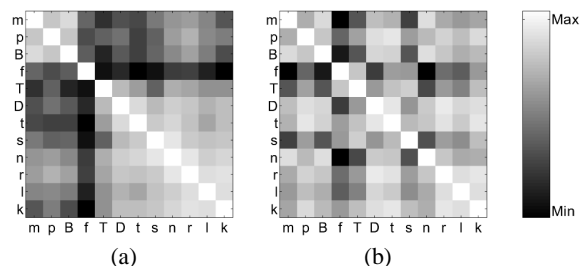


Figure 1: Distance graphs: (a) among sets of visual mouth appearances grouped by allophones; (b) among sets of audio vectors grouped also by allophones. White color is related to maximum similarity while black corresponds to a minimum one.

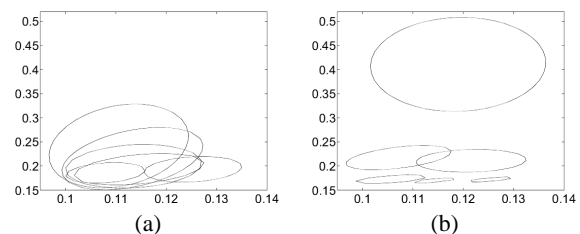


Figure 2: Auditory information clustering into six classes: (a) grouped by visemes; (b) grouped by sets of similar allophones. Clusters of (a) are more similar among them than those of (b).

3.2. Source data

The proposed clustering process stated in section 2.3 for video information and used in 3.1 for audio is known in this paper as *grouped data clustered by video* and *grouped data clustered by audio*, respectively. A strong restriction is imposed in both cases: vectors labelled with a particular allophone cannot be assigned to different groups in the clustering process. This restriction can be avoided if audio and video data are clustered regardless of their corresponding allophone. In these cases, *single data clustered by audio* and *single data clustered by video* are obtained, respectively.

3.3. Goodness measurement

Different sets of labels are provided by the four clustering processes defined in section 3.2. They depend on the source data (audio or video) and if segmentation information is considered (grouped or single). Moreover, a set of clusters can be defined from each set of labels for both the audio and video data subspaces (see table 1). In order to measure the goodness of the clustered data in each case, the percentage of correctly classified data vectors will be obtained with a maximum a posteriori (MAP) Bayesian estimation technique [20]. Each cluster can be approximated by a multivariate normal distribution $\mathcal{N}(\mu, \Sigma)$ from its own data as in section 2.2. The MAP technique finds the estimated cluster $\hat{\theta}_k$ which maximizes the following a posteriori probability of data vector \mathbf{x}_k from distributions \mathcal{N}_i and the a priori probabilities $p(\theta_i)$:

$$\hat{\theta}_k = \underset{\theta_i}{\operatorname{argmax}} \{p(\theta_i|\mathbf{x}_k)\} = \underset{\theta_i}{\operatorname{argmax}} \{p(\theta_i)p(\mathbf{x}_k|\theta_i)\} \quad (9)$$

$$p(\mathbf{x}_k|\theta_i) = \frac{1}{(2\pi|\Sigma_i|)^{M/2}} e^{-\frac{1}{2}(\mathbf{x}_k - \mu_i)\Sigma_i^{-1}(\mathbf{x}_k - \mu_i)} \quad (10)$$

where μ_i and Σ_i are the mean and covariance matrix of cluster i and M is the dimension of vector \mathbf{x}_k . Let θ_k be the real cluster of \mathbf{x}_k . If $\hat{\theta}_k = \theta_k$, then \mathbf{x}_k is correctly classified.

When evaluating the clustering goodness in the visual subspace, \mathbf{x}_k are the video vectors. On the contrary, they are the auditory vectors when the evaluation is carried out in the auditory subspace. The data sets used to build the distributions \mathcal{N} are the same as the test data sets; although this results in global higher estimation rates, it is not considered important in this work because relative comparisons between clustering processes are desired rather than absolute performance results.

Table 1: *Obtained cluster sets (CS). Labels are obtained from grouped or single data from both modes and then they can be used to cluster video and audio information.*

Clustered data	Single Data		Grouped Data	
	by audio	by video	by audio	by video
Audio	CS 1	CS 3	CS 5	CS 7
Video	CS 2	CS 4	CS 6	CS 8

4. Experimental results

Three audiovisual sequences of three minutes each were recorded at 25 frames per second with a sampling frequency of 16000 Hz and image resolution of 320×240 pixels. A set of twelve sentences was uttered in Spanish by three non professional speakers. Then, video vectors were extracted from the

visual channel and audio vectors were obtained from the auditory one following the process described in section 2.1.

The sentence set was balanced for the Spanish language and included varied prosody because analysis of real natural data was desired. As a result, some allophones appeared too few times (< 12) and were not taken into account in the analysis. The considered allophones appeared between 19 and 199 times each and are shown in table 2. The discarded ones were [v], [b], [g], [m], [j], [p], [n], [c], [z], [ʎ], [r], [s], [x], [d]. In Spanish, [δ] symbol is used instead of [ð] because the former is more dental than interdental that the latter.

Table 2: *Considered Spanish allophones.*

Allophones
[f], [θ], [p], [β], [δ], [k],
[m], [t], [l], [r], [n], [s]

4.1. Viseme sets

Following the allophone grouping process in the visual domain stated in section 2.3, the visemes obtained are summarized in table 3 for each studied subject. The result with six groups is given following the number of groups used in the literature [10] to compare with. Note that for each subject the viseme consisting of allophones [m], [p], [β], the one consisting of [f] and the one corresponding to allophone [θ] are found in separated groups. However, there is weak agreement with other groups involving non visible articulators of the vocal tract.

Table 3: *Visemes obtained for three subjects and six groups.*

Visemes	Subject 1	Subject 2	Subject 3
Viseme 1	[m], [p], [β]	[m], [p], [β]	[m], [p], [β]
Viseme 2	[θ]	[θ]	[θ]
Viseme 3	[f]	[f]	[f]
Viseme 4	[δ], [t]	[δ], [t], [s]	[δ], [k]
Viseme 5	[n], [r]	[n], [l], [k]	[n], [r], [l]
Viseme 6	[s], [l], [k]	[r]	[t], [s]

4.2. Uncertainty measurement

The goodness measure defined in section 3.3 is obtained for every cluster set of table 1, with a number of groups ranging from 5 to 14. Their mean values through all three people can be seen in table 4. Note that increasing the goodness of one mode is translated to decreasing that of the other one. Moreover, the computed geometric mean between the goodness in both modes for each column of table 4 is similar through all columns ($< 7\%$ of difference between maximum and minimum value).

Table 4: *Goodness of each clustered data subspace of video and audio with the geometric mean in the last row.*

Clustered data	Grouped Data		Single Data	
	by audio	by video	by audio	by video
Audio	0.682	0.630	0.916	0.326
Video	0.386	0.443	0.331	0.923
Geometric mean	0.513	0.528	0.549	0.551

5. Discussion

Dimensionality reduction of section 2.1 has been a key point in this work because of the need of estimation of covariance matrix: image vectors of thousand of pixels have been reduced to 12 parameters thanks to PCA preserving 85% of the energy.

The results about viseme grouping of section 4.1 follow earlier studies (the reader can take a look at [10] for a good review) in two ways: *i*) allophone groups involving visible articulators are the same than those in the previous works but taking into account Spanish instead of English: viseme [m], [p], [β], viseme [f] and viseme [θ]; *ii*) as in the literature, less agreement is obtained for the other allophone groups involving internal articulators, with only few common allophones like [n] or [δ]. Furthermore, the results have been obtained with natural speech rather than with the nonsense utterances of previous works. This fact strengthens both previous and current work and suggests a new objective way to obtain visemes from natural speech.

Another interesting result is the observed uncertainty in section 4.2 between audio and video for a given clustering criteria. Increasing performance in one mode seems to decrease that of the other one. Moreover, their geometric mean remains similar in all four kinds of clustering. This fact shows numerically the idea stated in previous works about the limitations of working with one isolated mode (remember that each clustering process depends only of data of one mode). Particularly, the lower video performance when using segmented data (grouped data) can be explained by the fact that segmentation was made using auditory information. When no segmented data was used (single dada), the two options are nearly symmetric.

6. Concluding remarks and future work

This work has proposed an objective way to find viseme groups from segmented audiovisual sequences with natural speech in Spanish. The obtained results are highly correlated to earlier works about viseme definition in English using nonsense utterances and subjective long evaluation processes.

The method to find viseme groups has been generalized and other clustering criteria have been applied to audio and video modes. In order to keep objectivity, a goodness measurement has been defined to obtain the quality of each clustered data obtained in the different grouping processes. It has been noted a tradeoff in the quality between clustered audio and video information for each of the four given clustering options. Moreover, this tradeoff seems to be near a specific value, which remains similar in all observed examples.

Future work includes using more audiovisual sequences with professional acquisition devices to further confirm the obtained visemes. The proposed viseme grouping method can also be useful when customizing personalized talking heads.

7. References

- [1] Chen, T., "Audiovisual Speech Processing: Lip Reading and Lip Synchronization", IEEE Signal Processing Magazine, 18:9-31, 2001.
- [2] Campbell, R., Dodd, B. and Burnham, D., "Hearing by eye II: Advances in the psychology of speechreading and auditory-visual speech", UK: Psychology Press, 1998.
- [3] Massaro, D.W., Beskow, J., Cohen, M.M., Fry C.L. and Rodriguez, T., "Picture My Voice: Audio to Visual Speech Synthesis using Artificial Neural Networks", Proc. of AVSP, Santa Cruz, CA. pp. 133-138, 1999.
- [4] Yehia, H., Rubin, P. and Vatikiotis-Bateson, E., "Quantitative Association of Vocal-tract and Facial Behaviour", Speech Communication, 26:23-43, 1998.
- [5] Hong, P., Wen, Z. and Huang., T., "Real-Time Speech-Driven Face Animation With Expressions Using Neural Networks". IEEE Transactions on neural networks, 13:916-927, 2002.
- [6] Choi, K. and Hwang, J., "Automatic Creation of a Talking Head from a Video Sequence". IEEE Transactions on Multimedia 7:628-637, 2005.
- [7] Huang, X., Hon, H.W. and Reddy, R., "Spoken Language Processing: A Guide to Theory, Algorithm, and System Development". USA: Prentice Hall PTR, 2001.
- [8] Fisher, C.G., "Confusions Among Visually Perceived Consonants", JSHR, 11:796-804, 1968.
- [9] Binnie, C., Montgomery, A. and Jackson, P., "Auditory and Visual Contributions to the Perception of Consonants", JSHR, 17:619-630, 1974.
- [10] Owens, E. and Blazek, B., "Visemes Observed by Hearing-Impaired and Normal-Hearing Adult Viewers", JSHR, 28:381-393, 1985.
- [11] Summerfield, A.Q., "Some Preliminaries to a Comprehensive Account of Audio-Visual Speech Perception", Hearing by Eye: The psychology of lip-reading, USA: Lawrence Erlbaum Ass., pp. 3-51, 1987.
- [12] Tekalp, A.M. and Ostermann, J., "Face and 2-D Mesh Animation in MPEG-4", Signal Processing: Image Communication, 15:387-421, 2000.
- [13] Fukunaga, K., "Introduction to Statistical Pattern Recognition" USA: Academic Press, 2nd edition, 1990.
- [14] Kirby, M., "Geometric Data analysis: An Empirical Approach to Dimensionality Reduction and the Study of Patterns", USA: John Wiley & Sons Inc., 2001.
- [15] Jolliffe, I.T., "Principal Component Analysis", USA: Springer-Verlag, 1986.
- [16] Melenchón, J., Iriondo, I. and Meler, L., "Simultaneous and Causal Appearance Learning and Tracking", ELCVIA, 3:44-54, 2005.
- [17] Young, S., Evermann, G., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P. "The HTK Book (for HTK Version 3.2.1)", Cambridge University Engineering Department, 2003.
- [18] Yehia, H. and Itakura, F., 1994, "Determination of Human Vocaltract Dynamic Geometry from Formant Trajectories Using Spatial and Temporal Fourier Analysis", Proc. IEEE ICASSP, pp. 477-480, 1994.
- [19] GSM 06.90 version 7.2.1. "Digital cellular telecommunications system (Phase 2+); Adaptive Multi-Rate speech transcoding", 1998.
- [20] Moon, T.K. and Stirling, W.C., "Mathematical Methods and Algorithms for Signal Processing", USA: Prentice-Hall, 2000.