IMPACT OF SAMPLE SIZES ON INFORMATION THEORETIC MEASURES FOR AUDIO-VISUAL SIGNAL PROCESSING

Ivana Arsic, Ninoslav Marina and Jean-Philippe Thiran

Signal Processing Institute, Ecole Polytechnique Fédérale de Lausanne (EPFL)
CH-1015 Lausanne, Switzerland
email:{Ivana.Arsic, Ninoslav.Marina, JP.Thiran}@epfl.ch
http://itswww.epfl.ch

ABSTRACT

In this paper we aim to explore what is the most appropriate number of data samples needed when measuring the temporal correspondence between a chosen set of video and audio cues in a given audio-visual sequence. Presently the optimal model that connects statistics of audio and video signals does not exist since one does not know the most appropriate features to be extracted in order to analyze their correlation. Previous approaches assumed simple parametric and non-parametric models for the joint distribution for capturing the complex signal relationships. The main problem in using standard information theoretic quantities, such as entropy and mutual information, is to accurately estimate the probability density function from a limited number of data samples. The main idea is to project the data into a statistically sufficient low-dimensional subspace, suitable for density estimation. Then using a simple parametric model based on assumption of Gaussianity, mutual information is estimated and applied as a measure of correspondence. We exploit how the choice of the sample size affects the reliability of the correspondence measure (mutual information) between selected features of the two modalities, audio and video.

1. INTRODUCTION

Due to the expansive and rapid growth of new technologies and human-computer interaction devices, information is becoming available through different media in various forms, such as audio, video and text to name a few. The fact that we can exploit different modalities makes multi-modal signal processing an important and challenging research area.

Extensive research done in the area of joint processing of audio and video signals shows that the help of one modality can be beneficial in cases when the other one is corrupted by noise. Moreover, even when the environmental conditions are not changed the system performance can be enhanced by using common properties of the two modalities. Such applications are for example audio-visual speech and speaker recognition, ([1],[2]), as well as speaker localization [3].

Most of the work done so far exploring cross-modal relationships between audio and visual features was for the purpose of measuring audio-visual synchrony. Recent work by Hershey and Movellan [4] can be considered as an introduction in a new study area where the mutual information can be used as a measure of correlation between two modalities. They use a model based on the assumption of Gaussian distribution for audio and visual cues. Slaney and Covell [5] introduce a more general method based on Canonical Correlation Analysis using the information from all pixels. In [6], based on information theory, Fisher et al. present a non-parametric approach that models the relationship between audio and visual data and finds the most informative subspaces by learning joint statistical models. Butz and Thiran in [7] use a model based on Markov chains for audio and visual signals and the maximization of mutual information for measuring audio-visual consistency.

In this paper we aim to investigate how the sample size affects the possibility of applying mutual information for measuring the correlation between audio and visual features. First, in Section 2 we recall the mathematical background regarding information theoretic quantities, such as entropy and mutual information, followed by an introduction of a parametric method for density estimation using the assumption of Gaussianity. Further on, the extraction step of audio and visual features of interest is described in Section 3. The experimental framework and obtained results are presented in Section 4, followed by the conclusions in Section 5.

2. INFORMATION THEORETIC APPROACH

Given a practical audiovisual system, there is a possibility to use the common information of the audio and the video component. The main problem is what features have to be extracted from both modalities in the joint signal. An intuitive solution is that we have to choose those features from the two modalities that have maximal mutual information. How to calculate the mutual information from different features is a practical problem in itself, due to the finite number of samples that are available in practice.

Let us consider the problem from the theoretical point of view. Assume first that we have two memoryless processes X and Y with joint probability density function p_{XY}. The mutual information between them is

$$\begin{aligned} I(X;Y) &= \sum_{x,y} p_{XY}(x,y) \log_2 \frac{p_{XY}(x,y)}{p_X(x)p_Y(y)} && (1) \\ &= E_X[D(p_{Y|X}||p_Y)]. && (2) \end{aligned}$$

where p_X, p_Y and $p_{Y|X}$ are the marginal probabilities of X and Y respectively and $p_{Y|X}$ is the conditional probability of Y given X, while $D(\cdot||\cdot)$ represents Kullback-Leibler distance, [8].

An alternative way of expressing the mutual information is in terms of marginal entropies $H(X)$ and $H(Y)$ and the joint entropy $H(X,Y)$, namely

$$I(X;Y) = H(X) + H(Y) - H(X,Y). \qquad (3)$$

The work is supported by the Swiss National Science Foundation through the National Center of Competence in Research on Interactive Multimodal Information Management (IM2).

For both expressions we need the exact joint probability density function p_{XY} in order to calculate the mutual information. Given a practical data set we are able only to find an approximate estimation for p_{XY}. How far we are from the correct probability density function depends on the number of samples N, that are available for modeling the joint process.

A very simple idea is to use an approximation for the joint probability density function. For example one possibility is to take the joint Gaussian probability density function

$$g_{XY}(x,y) = \frac{1}{2\pi\sqrt{|K|}} \cdot e^{-\frac{1}{2}[x,y]K^{-1}[x,y]^T}. \tag{4}$$

We use the notation $(X,Y) \sim N(\mu, K)$ to indicate that X and Y are jointly Gaussian with mean vector μ and covariance matrix $K = \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix}$.

Since the mean does not bring any uncertainty, without a loss of generality we may assume that $\mu = 0$. Thus, since $H(X) = 0.5\log_2 2\pi e\sigma^2$ and $H(X,Y) = 0.5\log_2(2\pi e)^2|K|$, we obtain

$$\begin{aligned} I(X;Y) &= \frac{1}{2}\log_2 \frac{\sigma_X^2\sigma_Y^2}{\sigma_X^2\sigma_Y^2 - \sigma_{XY}^2} \\ &= -\frac{1}{2}\log_2\left(1 - \rho^2(X,Y)\right). \end{aligned} \tag{5}$$

where ρ represents Pearson's correlation coefficient. The value of ρ can be between -1 and 1 denoting the existence of positive or negative correlation, and 0 when the random variables are uncorrelated.

2.1 Correlation coefficient and window size

For estimation of mutual information in terms of the correlation coefficient ρ we use the same principle as in [4]. The difference is that we assume Gaussianity holds for two signals locally over some short time window. Thus, at each time t_k and pixel position (x,y) we have a joint sequence of audio and visual feature vectors $(A(t_k), V(x,y,t_k))$ having Gaussian form of $p(A,V)$ in a given time window of size w (number of frames). In the described model both feature vectors are defined as sets of audio and visual cues in the previous frames at times $(t_k, ..., t_{k-w-1})$.
We can show the importance of choosing the right window size by performing a simple experiment with two uncorrelated signals $x(t)$ and $y(t)$ having normal distribution, each consisting of 1500 samples. We get different estimates of ρ and $I(X;Y)$ by changing the window size ranging from two to a maximum of 1500 samples. The histograms of the estimated parameters are shown in Figure 1. We can see that the left histogram of a two sample window size leads to the wrong conclusion of a linear relationship between x and y. With increasing window size the estimated value of ρ will come closer to the true value of 0, as shown on the right.

3. FEATURE EXTRACTION

3.1 Audio-visual database

The work presented here utilizes material from the Clemson University audio-visual speech corpus called CUAVE [9]. It is a speaker-independent database of 36 speakers of different gender and origin (19 male and 17 female) pronouncing both isolated and continuous digit strings. For our experiments we consider those video clips where only one person is present in the scene frontally facing the camera with no rule regarding the position of the speaker in the frame.

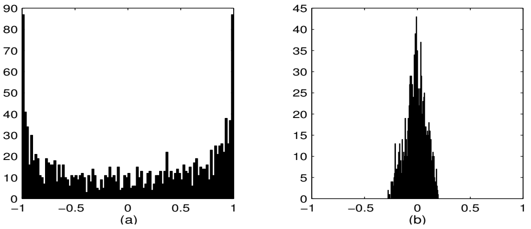


Figure 1: Correlation coefficient histograms using different window sizes (a) $w = 2$, and (b) $w = 100$.

3.2 Visual feature extraction

The first step in the process of extracting meaningful visual features is in the pre-processing of facial images. In order to determine the facial region we rely on a model based on anthropometric measures of the human head and face as in [10].

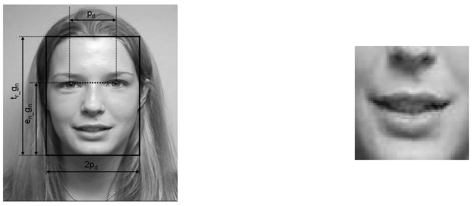


Figure 2: Anthropometric measures used for locating the lower face region from the pupils' distance (a), and the cropped ROI (b).

The region of interest (ROI) includes the lower part of the face, mouth area, jaws and chin, and is located from the pupils-facial middle distance, as shown in Figure 2 (a). Due to the lack of an adequate eye-tracker, centers of eyes are manually marked and the distance between them is denoted p_d. Once the ROI is obtained, it is downsampled to the size of 64×64 pixels as represented in Figure 2 (b). We consider a pixel-based approach such that every pixel in the raw image is an element of a feature vector.
Besides the raw pixel intensity values, we also consider those obtained after image averaging in 2×2 and 3×3 kernels. Those features were of particular interest due to the fact that the value of the correlation coefficient ρ depends on the level of the image noise that can be decreased by performing averaging in the neighborhood region. In order to include the temporal component, delta images (pixel intensity change over time) are also considered, for both original and enhanced images.

3.3 Audio feature extraction

The audio pre-processing step results in the extraction of the commonly used Mel Frequency Cepstral Coefficients (MFCCs) and also in the average audio energy per frame, as well as their first order derivatives. Thus, for each speech frame a 26-dimensional feature vector is obtained.

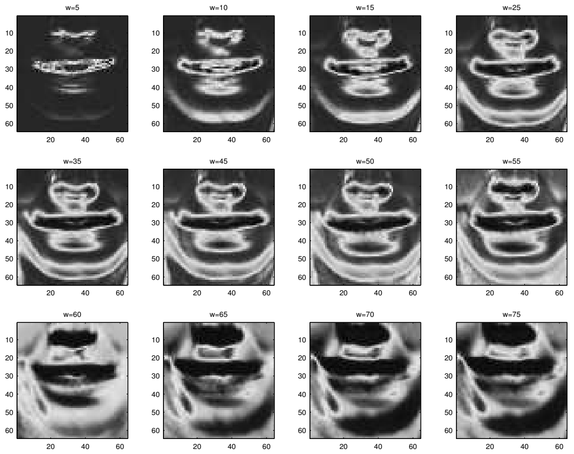


Figure 3: The effect of changing window size from 5 to 75 samples.

4. EXPERIMENTS AND RESULTS

For conducting the experiments we use an MPEG-2 video clip from the audio-visual database CUAVE, [9]. The original clip consists of both a "still" and a "moving" part, but we consider only the first one, where the subject pronounces digits from "zero" to "nine", five times in repetition in connected manner without changing the orientation of the head significantly. Since the video frame rate is 29.97 fps, the number of acquired frames from this chosen part of the video sequence is 1225. Using previously described feature extraction methods, visual cues from 64×64 ROI and accompanying audio features are obtained and normalized to the interval of $(0,1)$.

Then, we use the simple parametric model presented in Section 2 (and exploited in the work of Hershey and Movellan, [4]) and apply it to our audio-video sequence. The extracted video frames are ordered in time and for each one there is a corresponding audio frame. Mutual information is calculated between the two time-series from the intensity change of each pixel at the location (x,y) over time and each audio energy value and Mel-Frequency Cepstral Coefficients. The estimated mutual information between pixel and audio features is expressed in the form of gray scale pixel brightness, and scaled such that the full color map scale range is covered.

In order to show the importance of choosing the appropriate sample size we vary it from five to 75. The obtained results are represented in Figure 3. What we can see is that in the case window size being very small e.g. five samples (shown in the upper left corner), mutual information is not a good measure of correspondence since high correlation is indicated at the facial parts that are not assumed to be connected to speech. This can be seen in the upper left picture of Figure 3 where the dark pixels (high mutual information) are attributed to the areas of supposedly low connectivity. We can consider this fact from the statistical point of view and observe only Pearson's correlation coefficients for a given number of samples. The example is shown in Figure 4 for the frame number 45 when the window size is 10. The dark pixel regions in the right image (high correlation) that correspond to the light pixel regions in the left image (high mutual information) do not show a true relation between audio and video. Supposedly high correlation coefficient scores are less than 0.632 which is a critical value for 10 samples in a 95% confidence interval and therefore the hypothesis of the existing correlation can be rejected.

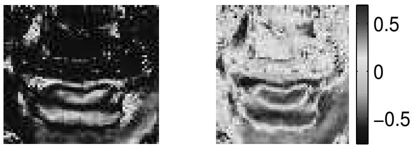


Figure 4: Mutual information image (left) and correlation image using Pearson's correlation coefficient (right) for a window size of 10 frames.

On the other hand, if we consider setting larger window sizes (more than 75) mutual information values would be incorrect due to the fact that the real audio and visual cues do not have Gaussian distribution, so the assumption and proposed model do not any longer hold in a given time window.

Satisfactory results when using a mutual information based measure are obtained only when sample sizes are between 60 and 75 (equal to window size in frames). Examples showing the correlation at the expectably most speech associated facial parts (facial muscles around the lips area and jaws) when using good sample sizes can be seen in the bottom row of Figure 3. These values are for a given set of audio and visual features, in this case audio energy and pixel intensities. Figure 5 displays the examples showing how the facial parts having highest mutual information scores are changing with the increase of window size.

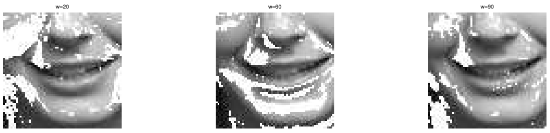


Figure 5: Mutual information scores superposed on the original image when window size is 20, 60 and 90.

The same set of experiments were performed using different previously described visual features. There was no significant change regarding the window size in cases when the pixel values were averaged in a 2×2 or 3×3 region. The difference can be noticed only when the temporal component is also taken into account, and the corresponding results are presented in Figure 6. Satisfactory results in terms of high mutual information scores that can be observed around the mouth region and jaws are obtained for a window sizes ranging from 45 (darker gray pixels around lips) to 65 (lightest pixels).

Another question of interest is if the chosen sample size will be appropriate in the case when facial movements are uncorrelated with the audio. For this purpose we alternated the original audio signal with the one taken from another speaker in the database. The obtained results for a window sizes of 10 and 65 frames are represented in Figure 7. When the number of samples is small, mutual information will show a relationship between video features and unassociated audio. In the case of the window size being correctly chosen e.g. $w = 65$ highest mutual information values (lightest pixels) are detected around lip region and jaw line of the

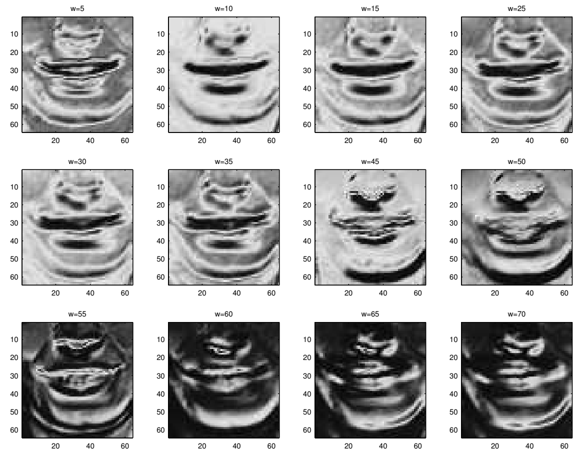
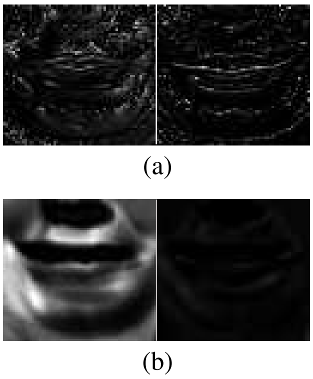


Figure 6: The effect of changing window size from 5 to 70 samples when using delta images with pixel intensities averaged in a 3×3 pixel region.

person related to audio, and are almost zero in the case of an unassociated sound source.



(a)

(b)

Figure 7: Mutual information images when using the correct (left) and alternated audio signal (right) for a window size $w = 10$ (a), and $w = 65$ (b).

The drawback of choosing the window size between 60 and 75 is the increased delay between the input and the output. To cope with this problem we performed the same test with the correct and uncorrelated audio using the same sample size, but taking into account $w/2$ before and $w/2$ number of samples after the observed frame at time t_k. Obtained results for a window size of two seconds duration are shown in Figure 8 displaying similar properties as in previous cases.

5. CONCLUSIONS AND FUTURE WORK

The use of mutual information as a statistical measure of dependence between two random variables highly depends on the correct probability density estimation. No matter if the approach is based on parametric or non-parametric assumptions the question that remains is how to correctly choose the number of data samples such that the empirically estimated distribution gives satisfactory results. By performing various experiments using a parametric model with Gaussian distribution we show that size of the sample set does have an impact on correct estimation of correlation and mutual information. Moreover, the window size should be carefully chosen. If too small, the data would be insufficient to reliably measure the correlation, while if too big, the Gaussianity assumption will not hold and the used statistical model will not be appropriate. The future work is to explore theoretical concepts for finding the right window size (number of samples) instead of choosing it in some heuristic manner (for example using ”method of types”, [11]).

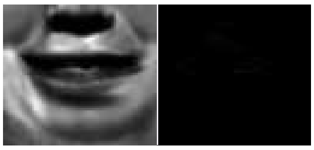


Figure 8: Mutual information images when using the correct (left) and alternated audio signal (right) for a window size $w = 60$ taking $w/2$ past and $w/2$ future samples.

REFERENCES

[1] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, “Audio-visual automatic speech recognition: An overview,” *Issues in Visual and Audio-Visual Speech Processing* (G. Bailly, E. Vatikiotis-Bateson, and P. Perrier, eds.), MIT Press, 2004.

[2] G. Potamianos, C. Neti, G. Gravier, and A. Garg, “Automatic recognition of audio-visual speech: recent progress and challenges,” *Proceedings of the IEEE*, vol. 91, no. 9, Sep. 2003.

[3] H. J. Nock, G. Iyengar, and C. Neti, “Speaker localization using audio-visual synchrony: an empirical study,” in *Proc. of the 10th ACM International Conference on Multimedia*, 2002.

[4] J. Hershey and J. Movellan, “Audio-vision: using audio-visual synchrony to locate sounds,” in *Proc. of NIPS*, vol. 13, 2000.

[5] M. Slaney and M. Covell, “FaceSync: a linear operator for measuring synchronization of video facial images and audio tracks,” in *Proc. of NIPS*, vol. 12, 1999.

[6] J. W. Fisher III, T. Darrell, W. T. Freeman, and P. Viola, “Learning joint statistical models for audio-visual fusion and segregation,” in *Proc. of NIPS*, vol. 13, 2000.

[7] T. Butz and J.-P. Thiran, “From error probability to information theoretic (multi-modal) signal processing,” *Signal Processing*, vol. 85, no. 5, pp. 875-902, 2005.

[8] T. M. Cover and J. A. Thomas, *Elements of Information Theory.* John Wiley & Sons, Inc., 1991.

[9] E. Patterson, S. Gurbuz, Z. Tufekci, and J. Gowdy, “Moving talker, speaker-independent feature study and baseline results using the CUAVE multimodal speech corpus,” *EURASIP Journal on Applied Signal Processing*, vol. 2002, no. 11, pp. 1189-1201, 2002 .

[10] L. G. Farkas, *Anthropometry of the Head and Face.* Raven Press, 1994.

[11] I. Csiszar, “The method of types [information theory],” *IEEE Transactions on Information Theory*, vol. 44 , no. 6 , pp. 2505-2523, Oct. 1998.