# Algorithm for Computing Spatiotemporal Coordination

*Adriano V. Barbosa*[1], *Hani C. Yehia*[2], *Eric Vatikiotis-Bateson*[1]

[1]Department of Linguistics, University of British Columbia, Vancouver, Canada
[2]Department of Electronics, Federal University of Minas Gerais, Belo Horizonte, Brazil

adriano.vilela@gmail.com, hani@cefala.org, evb@interchange.ubc.ca

## Abstract

This work presents an algorithm that allows the coupling between auditory and visual concomitants of spoken communication to be computed as it evolves through time, thus allowing coordination of events to be examined computationally in the time domain.

**Index Terms**: spatiotemporal coordination, synchronization, multimodal events.

## 1. Overview

Some degree of spatial and temporal coordination can be computed for almost any combination of biological events either within or between organisms. At the same time and regardless of perception, strict synchronization (with or without a temporal offset) between events is almost non-existent physically in coordinated behavior (huge exception – British fans of Freddie Mercury at Live Aid, 1985). For example, for esthetic reasons, if for no other, musicians in an ensemble do not all play their notes at precisely the same moment. Rather, they work their way around the synchronization point in such a way that the result sounds synchronized, but is not quite. It is also the case that within an organism, there may be nearly constant offsets between events in one domain and related (e.g., consequential) events in another, such as the neurophysiological link between muscle activity and speech articulation.

Previous work has shown strong correspondences between face motion and spectral acoustics [1], and between fundamental frequency (F0) and rigid body motion of the head during speech production [2]. However, these estimations have always been made from sample-by-sample correspondences averaged over sentence-sized temporal spans. As a result, the temporal organization contributes only indirectly to the assessment of multimodal behavior and temporal offset, be it zero or otherwise, is treated as a constant for the entire measurement span.

The algorithm described here computes correspondences between signals (e.g., head motion and RMS amplitude of the acoustics) recursively as time-series of instantaneous correlations. Unlike the first version of this algorithm [3], the filter now includes past and future samples in estimating the correlation at a moment in time and the user sets not only the filter sharpness, $\eta$, but also the range of temporal offsets (zero) between the compared signals. The visualized result is a 2D correlation map where temporal offsets between two compared signals are plotted as a function of correlation (degree and direction shown by color) over time.

The mathematical formulation of the instantaneous correlation algorithm is presented in Section 2. Both the 1D and the 2D correlation signals are discussed. The techniques developed in Section 2 are then applied to both synthetic and audiovisual speech data, and the results are presented and discussed in Section 3. The summary is presented in Section 4.

## 2. Instantaneous correlation algorithm

### 2.1. The causal filter approach

The instantaneous correlation coefficient $\rho(k)$ between signals $x(k)$ and $y(k)$ can be defined as [4]

$$\rho(k) = \frac{S_{xy}(k)}{\sqrt{S_{xx}(k)\,S_{yy}(k)}}\,, \tag{1}$$

where the instantaneous covariance $S_{xy}(k)$ between $x(k)$ and $y(k)$ is computed as

$$S_{xy}(k) = \sum_{l=0}^{\infty} c\,e^{-\eta\,l}\,x(k-l)\,y(k-l)\,, \tag{2}$$

with $S_{xx}(k)$ and $S_{yy}(k)$ defined similarly. In Equation 2, $\eta$ is a small positive number. From this equation it can be seen that the covariance $S_{xy}(k)$ at any point $k = k_0$ in time is simply a weighted mean of the signal $v(k) = x(k)\,y(k)$ computed over the interval $\{k : k \leq k_0\}$. This is a *causal* system, because only past samples are used to compute $S_{xy}(k)$ at any point in time. The weights decay exponentially with time in such a way that older samples receive smaller weights. Figure 1 shows a graphical representation of the computation of $S_{xy}(k)$ as given in Equation 2.

The constant $c$ in Equation 2 is a normalization factor which ensures that the sum of all weights is 1. Thus, the value of $c$ can be calculated by doing

$$\sum_{l=0}^{\infty} c\,e^{-\eta\,l} = 1\,, \tag{3}$$

which yields

$$c = 1 - e^{-\eta}\,. \tag{4}$$

The signal $S_{xy}(k)$ can be seen as the output of a first-order low-pass linear filter excited by the signal $v(k) = x(k)\,y(k)$. The $z$-transform representation of this linear filter is given by

$$H(z) = \frac{c}{1 - e^{-\eta}z^{-1}}, \quad |z| > e^{-\eta}\,. \tag{5}$$

If $\mathcal{F}_c\{\cdot\}$ is used to denote the operation performed by the linear filter in Equation 5, then Equation 2 can be rewritten as

$$S_{xy}(k) = \mathcal{F}_c\left\{x(k)\,y(k)\right\}\,. \tag{6}$$

The definition of $S_{xy}(k)$ as given in Equation 2 is limited to those cases where $x(k)$ and $y(k)$ are zero mean signals. In

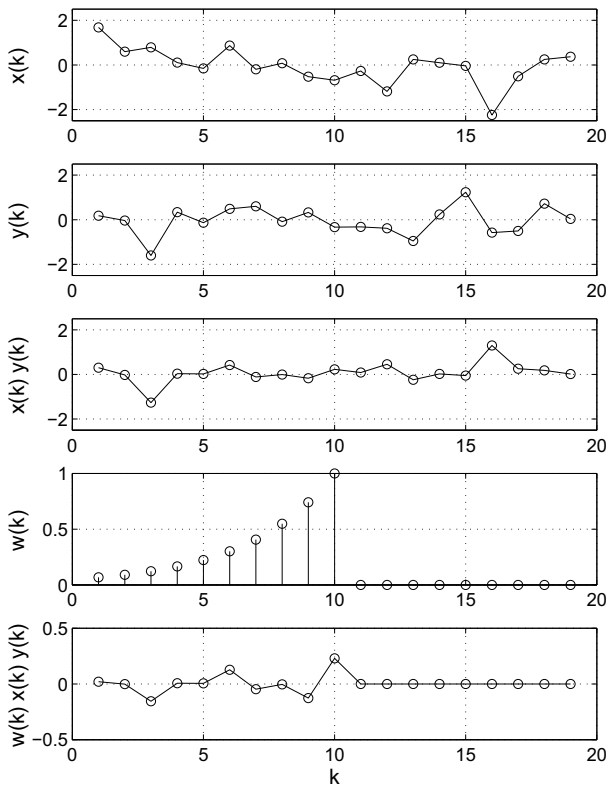


Figure 1: *Computation of the instantaneous covariance $S_{xy}(k)$ between zero-mean signals $x(k)$ and $y(k)$ at time $k_0 = 10$. From top to bottom: signal $x(k)$; signal $y(k)$; the product $x(k)\,y(k)$; the exponential weighting function $w(k)$; and the weighted product $w(k)\,x(k)\,y(k)$. $S_{xy}(k_0)$ is computed by summing the samples of the signal in the bottom panel and then multiplying the result by the normalization factor $c$ in Equation 3.*

the more general case of nonzero, and even time-varying, mean values, Equation 2 can be redefined as

$$S_{xy}(k) = \sum_{l=0}^{\infty} c\, e^{-\eta\, l}\, (x(k-l) - \bar{x}(k))\, (y(k-l) - \bar{y}(k))\,, \quad (7)$$

where the instantaneous means $\bar{x}(k)$ and $\bar{y}(k)$ are computed as

$$\bar{x}(k) = \sum_{l=0}^{\infty} c\, e^{-\eta\, l}\, x(k-l)\,, \quad (8)$$

$$\bar{y}(k) = \sum_{l=0}^{\infty} c\, e^{-\eta\, l}\, y(k-l)\,. \quad (9)$$

Again, we note that the signals $\bar{x}(k)$ and $\bar{y}(k)$ can be seen as the output of the linear filter in Equation 5 excited by the signals $x(k)$ and $y(k)$, respectively. Thus

$$\bar{x}(k) = \mathcal{F}_c\,\{x(k)\}\,, \quad (10)$$

$$\bar{y}(k) = \mathcal{F}_c\,\{y(k)\}\,. \quad (11)$$

Using the definitions in equations 8 and 9, Equation 7 can be rewritten as

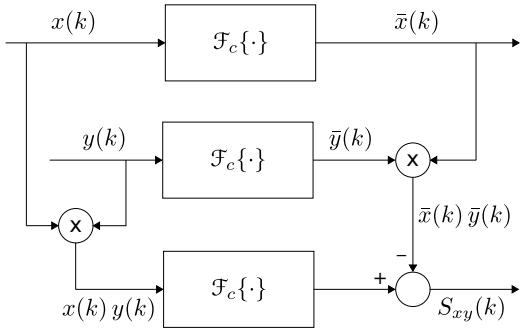


Figure 2: *Block diagram representation of the computation of the instantaneous covariance between signals $x(k)$ and $y(k)$ according to Equation 13.*

$$S_{xy}(k) = \sum_{l=0}^{\infty} c\, e^{-\eta\, l}\, x(k-l)\, y(k-l) - \bar{x}(k)\, \bar{y}(k)\,, \quad (12)$$

or, using the filter operator defined in Equation 6

$$S_{xy}(k) = \mathcal{F}_c\,\{x(k)\, y(k)\} - \mathcal{F}_c\,\{x(k)\}\; \mathcal{F}_c\,\{y(k)\}\,. \quad (13)$$

The block diagram in Figure 2 shows the computation of the instantaneous covariance $S_{xy}(k)$ from signals $x(k)$ and $y(k)$ as given in Equation 13.

### 2.2. The non-causal filter approach

The causal system, in which only past samples are used, is useful for computing the instantaneous covariance in real time. However, in those cases where real time processing is not a requirement, both past and future samples can be used to compute the covariance. For zero-mean signals, this can be done by redefining Equation 2 as

$$S_{xy}(k) = \sum_{l=-\infty}^{\infty} c\, e^{-\eta\,|l|}\, x(k-l)\, y(k-l)\,. \quad (14)$$

Thus, in the non-causal approach, the covariance at time $k = k_0$ is a weighted mean computed to both sides (before and after) of $k_0$. Figure 3 shows a graphical representation of the computation of $S_{xy}(k)$ for two zero-mean signals as given in Equation 14.

Again, the constant $c$ in Equation 14 ensures that the sum of all weights is 1 and can be computed by doing

$$\sum_{l=-\infty}^{\infty} c\, e^{-\eta\,|l|} = 1\,, \quad (15)$$

which yields

$$c = \frac{1 - e^{-\eta}}{1 + e^{-\eta}}\,. \quad (16)$$

Equation 14 can be expanded into

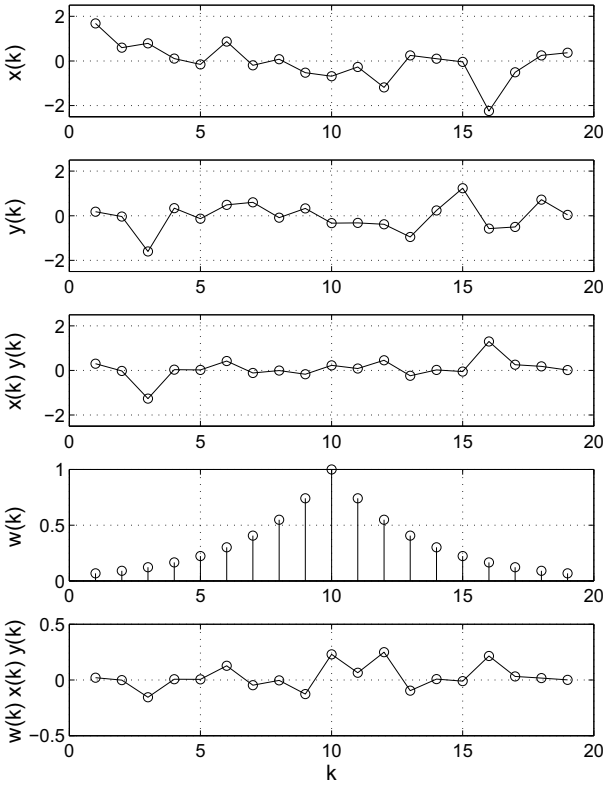


Figure 3: *Computation of the instantaneous covariance $S_{xy}(k)$ between zero-mean signals $x(k)$ and $y(k)$ at time $k_0 = 10$. From top to bottom: signal $x(k)$; signal $y(k)$; the product $x(k)\,y(k)$; the exponential weighting function $w(k)$; and the weighted product $w(k)\,x(k)\,y(k)$. $S_{xy}(k_0)$ is computed by summing the samples of the signal in the bottom panel and then multiplying the result by the normalization factor $c$ in Equation 16.*

$$
\begin{aligned}
&S_{xy}(k) = \\
&= \sum_{l=-\infty}^{0} c\,e^{\eta\,l}\,v(k-l) + \sum_{l=0}^{\infty} c\,e^{-\eta\,l}\,v(k-l) - c\,v(k-l) \\
&= \sum_{l=0}^{\infty} c\,e^{-\eta\,l}\,v(k+l) + \sum_{l=0}^{\infty} c\,e^{-\eta\,l}\,v(k-l) - c\,v(k-l)\,,
\end{aligned}
\tag{17}
$$

where $v(k) = x(k)\,y(k)$. Both the first and second terms in Equation 17 can be obtained with the linear filter in Equation 5. The second term is simply the response of the filter to the input signal $v(k)$ (see equations 2 and 6). The first term, in turn, can be obtained by reversing the signal $v(k)$ in time, inputing this *mirrored* signal to the filter, and then reversing the filter's response back to its original time course. Defining an operator $\mathcal{P}\{\cdot\}$ for this mirroring operation, the first term in Equation 17 can be written as

$$
\sum_{l=0}^{\infty} c\,e^{-\eta\,l}\,v(k+l) = \mathcal{P}\left\{\mathcal{F}_c\left\{\mathcal{P}\left\{v(k)\right\}\right\}\right\}\,. \tag{18}
$$

Thus, Equation 17 becomes

$$
S_{xy}(k) = \mathcal{P}\left\{\mathcal{F}_c\left\{\mathcal{P}\left\{v(k)\right\}\right\}\right\} + \mathcal{F}_c\{v(k)\} - c\,v(k)\,. \tag{19}
$$

We can think of the combined computations in Equation 19 (i.e., the forward and backward filtering of $v(k)$ in the first two terms and the summing of all three terms) as being performed by a single non-causal filter $\mathcal{F}_{nc}\{\cdot\}$,

$$
\mathcal{F}_{nc}\{v(k)\} = \mathcal{P}\left\{\mathcal{F}_c\left\{\mathcal{P}\left\{v(k)\right\}\right\}\right\} + \mathcal{F}_c\{v(k)\} - c\,v(k)\,. \tag{20}
$$

Equation 19 can then be rewritten as

$$
S_{xy}(k) = \mathcal{F}_{nc}\{v(k)\} = \mathcal{F}_{nc}\{x(k)\,y(k)\}\,, \tag{21}
$$

which is the non-causal equivalent of Equation 6 in Subsection 2.1.

The definition of $S_{xy}(k)$ as given in Equation 14 applies only to zero-mean signals. For signals with non-zero, or even time-varying, mean values, Equation 14 can be redefined as

$$
S_{xy}(k) = \sum_{l=-\infty}^{\infty} c\,e^{-\eta\,|l|}\,\left(x(k-l) - \bar{x}(k)\right)\left(y(k-l) - \bar{y}(k)\right)\,, \tag{22}
$$

where the instantaneous means $\bar{x}(k)$ and $\bar{y}(k)$ are computed as

$$
\bar{x}(k) = \sum_{l=-\infty}^{\infty} c\,e^{-\eta\,|l|}\,x(k-l)\,, \tag{23}
$$

$$
\bar{y}(k) = \sum_{l=-\infty}^{\infty} c\,e^{-\eta\,|l|}\,y(k-l)\,. \tag{24}
$$

Equations 22-24 are the non-causal equivalents of equations 7-9. The signals $\bar{x}(k)$ and $\bar{y}(k)$ can be seen as the response of the generalized filter $\mathcal{F}_{nc}\{\cdot\}$ to the input signals $x(k)$ and $y(k)$, respectively. Thus

$$
\bar{x}(k) = \mathcal{F}_{nc}\{x(k)\}\,, \tag{25}
$$

$$
\bar{y}(k) = \mathcal{F}_{nc}\{y(k)\}\,. \tag{26}
$$

In the same way that Equation 7 could be rewritten as Equation 12, Equation 22 can be rewritten as

$$
S_{xy}(k) = \sum_{l=-\infty}^{\infty} c\,e^{-\eta\,|l|}\,x(k-l)\,y(k-l) - \bar{x}(k)\,\bar{y}(k)\,, \tag{27}
$$

or, using the filter operator $\mathcal{F}_{nc}\{\cdot\}$ defined in Equation 20

$$
S_{xy}(k) = \mathcal{F}_{nc}\left\{x(k)\,y(k)\right\} - \mathcal{F}_{nc}\left\{x(k)\right\}\,\mathcal{F}_{nc}\left\{y(k)\right\}\,, \tag{28}
$$

which is the non-causal counterpart of Equation 13. By replacing the operator $\mathcal{F}_c\{\cdot\}$ with $\mathcal{F}_{nc}\{\cdot\}$ in Figure 2, the same block diagram can be used to represent the computation of the non-causal covariance as given in Equation 28.

It should be noted that all computations involved in obtaining the instantaneous covariance, using either the causal or the non-causal approach, can be performed using the linear filter in Equation 5. This is important when using Matlab, as all calculations can be carried out by using Matlab's `filter()` function, which makes the computation of $S_{xy}(k)$ much faster than simply implementing the loops associated with the summations.

Summarizing, we have defined two ways to compute the instantaneous covariance $S_{xy}(k)$ between signals $x(k)$ and $y(k)$. The first method, given by Equation 13, computes the covariance at any point $k = k_0$ in time as a weighted mean over the samples before $k_0$. The second method, given by Equation 28, computes the covariance at $k = k_0$ as a weighted mean over the samples both before and after $k_0$. Both methods can be applied to signals with non-zero, time-varying mean values. The instantaneous correlation coefficient $\rho(k)$ can then be easily computed from the instantaneous covariances using Equation 1.

### 2.3. The 2D correlation map

In the previous subsections, two methods were defined for computing the correlation coefficient between two signals as a function of time. In the case of offline processing of pre-recorded signals, there is nothing to prevent us from computing the correlation coefficient between time-shifted versions of the signals as well. This is useful in many scenarios. For example, in audiovisual speech analysis, we may be interested in looking at the synchronization between a speaker's speech acoustics and his/her accompanying gestures. We know that signals in these two domains are related somehow, but they are not necessarily synchronized. So, corresponding events in these two domains which are even slightly out of sync will not show up as high correlation values in the instantaneous correlation signal. However, by computing the correlation between the signals for a range of temporal offsets between them, a high correlation value, indicating synchronization, may emerge for some offset. It then becomes possible to express the correlation as a function of both time and temporal offset

$$\rho(k,d) = \frac{S_{xy}(k,d)}{\sqrt{S_{xx}(k,d)\, S_{yy}(k,d)}}\,, \tag{29}$$

where $S_{xy}(k,d)$ is the instantaneous covariance between signals $x(k)$ and $y(k-d)$

$$S_{xy}(k,d) = S\left(x(k), y(k-d)\right)\,. \tag{30}$$

Typically, a maximum value $d_{max}$ is defined for the offset and then all values in the range $\{d : |d| \leq d_{max}\}$ are considered.

## 3. Validation and results

In this section, we present and discuss the results of applying these techniques to both synthetic and audiovisual speech signals. The use of synthetic signals is particularly important in validating the techniques, since the instantaneous correlation results for very simple signals can be easily validated visually (at least in qualitative terms).

It should be emphasized that the computation of the instantaneous correlation coefficient (and also of the instantaneous means), for both the causal and the non-causal approaches, depend ultimately on the linear filter in Equation 5. This is an adjustable filter where the exponential decay can be changed by varying the value of the parameter $\eta$. The larger $\eta$ is, the steeper the exponential is, which means past samples are "forgotten" faster. This corresponds to a short memory filter, which is more sensitive to local changes in the input signal. The smaller $\eta$ is, the less steep the exponential is, which means past samples are "forgotten" more slowly. This corresponds to a long memory filter, which is less sensitive to local changes in the input signal. Thus, the value of $\eta$ can be used to make the filter less or more sensitive to rapid changes in the correspondence between

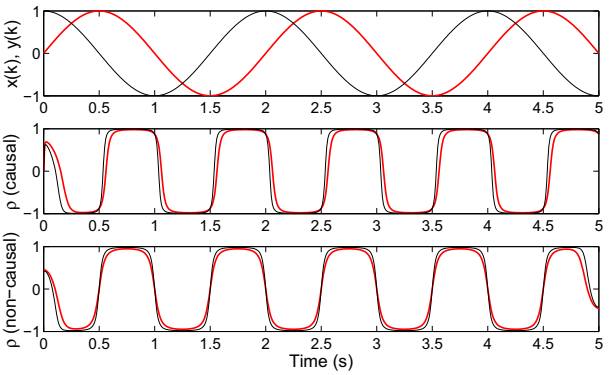


Figure 4: *Instantaneous correlation coefficient between signals* $x(t) = \sin(\pi t)$ *and* $y(t) = \cos(\pi t)$. *Top: signals* $x(t)$ *(red thick line) and* $y(t)$ *(black thin line). Middle: instantaneous correlation using the causal approach with* $\eta = 0.3$ *(red thick line) and* $\eta = 0.5$ *(black thin line). Bottom: instantaneous correlation using the non-causal approach with* $\eta = 0.3$ *(red thick line) and* $\eta = 0.5$ *(black thin line).*

signals and, therefore, to adjust the scale at which we want to analyze the underlying events.

Figure 4 shows the instantaneous correlation coefficient between the synthetic signals $x(t) = \sin(\pi t)$ and $y(t) = \cos(\pi t)$. From the figure, it is obvious that the correlation coefficient oscillates between -1 and +1, which is consistent with the fact that the two signals either move in the same direction at the same time, or in opposite directions at the same time. Also clearly visible is the effect the value of $\eta$ has on how fast the correlation coefficient goes from one extreme to the other; a larger value means a more abrupt change between the two extremes. This is true for both the causal and non-causal filters. Figure 4 also shows the difference between the causal and non-causal approaches. The non-causal filter approach results in a correlation coefficient signal which is completely symmetric along the time axis (i.e., $\rho(t) = \rho(-t)$), except at the beginning and end of the signal, where there are border effects. However, the same is not true for the causal filter approach. This is consistent with the fact that the non-causal approach computes the correlation at every point in time by taking into account both sides of $k = k_0$, whereas in the causal approach only the signal before $k = k_0$ is considered.

Figure 5 shows the instantaneous correlation coefficient between the signals of Figure 4 as a function of both time and temporal offset between the signals. Since we now have a function of two variables, the dependent variable (the correlation coefficient) is represented as color ranging from dark blue ($\rho = -1$) to dark red ($\rho = +1$). This 2D signal (whose graphical representation we call a correlation map) provides much more information about how the two signals are related than the 1D correlation signal of Figure 4. This becomes obvious when we consider real audiovisual signals (below); for now we note that the 1D correlation signal is simply the 2D correlation signal at offset zero.

Figure 6 shows the instantaneous correlation coefficient between two signals representing different techniques for measuring orofacial motion; namely, marker tracking and optical flow (see [5, 3] for details). Again, how the value of $\eta$ influences the smoothness of the resulting correlation signal is evident. A smaller value of $\eta$ (longer memory, less sensitive filter) clearly

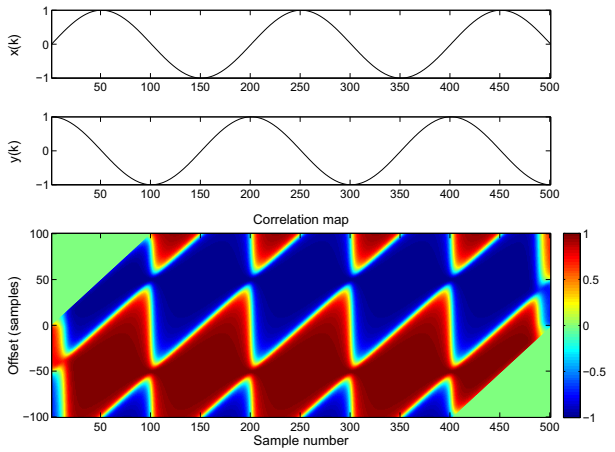


Figure 5: *Instantaneous correlation coefficient between the signals in Figure 4 for various temporal offsets between them. Top: signal $x(k)$. Middle: signal $y(k)$. Bottom: instantaneous correlation coefficient between $x(k)$ and $y(k)$ as a function of both the sample number and the temporal offset between the signals. A non-causal filter with $\eta = 0.3$ was used.*

results in a smoother correlation signal for both the causal and non-causal filters. It can also be seen (by comparing the middle and the bottom panels) that, for the same value of $\eta$, the non-causal filter results in a smoother correlation signal. This happens because, in the non-causal approach, the correlation coefficient at each point in time is computed over a larger number of samples (both left and right sides of $k = k_0$, rather than only the left side ).

Figure 7 shows the 2D correlation map for the signals in the top panel of Figure 6 for three different values of $\eta$ (0.1, 0.3, 0.5). The influence of $\eta$ on the correlation results becomes even more evident from this figure. As we move from a smaller

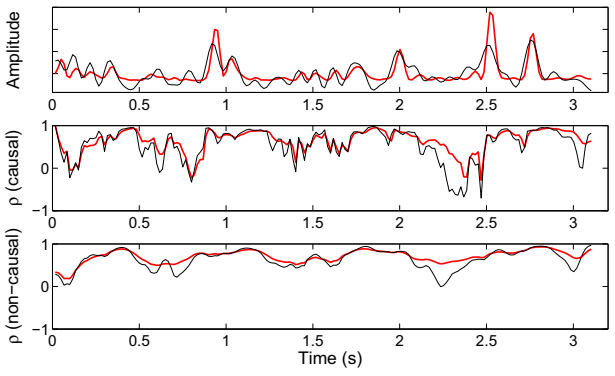


Figure 6: *Instantaneous correlation coefficient between two time-series of orofacial motion for the English sentence "Those windows are do dirty I can't see anything outside". Top: facial motion measured with marker tracking (summed marker positions – thick red line) and with optical flow (summed optical flow values – thin black line) (for details, see [6]). Middle: instantaneous correlation using the causal approach with $\eta = 0.3$ (red thick line) and $\eta = 0.5$ (black thin line). Bottom: instantaneous correlation using the non-causal approach with $\eta = 0.3$ (red thick line) and $\eta = 0.5$ (black thin line).*

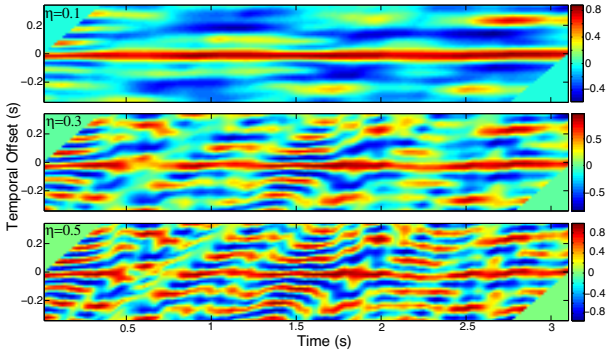


Figure 7: *Instantaneous correlation of signals shown in the top panel of Figure 6 at a range of temporal offsets (+/- 0.33 s) for three values of $\eta$ (0.1, 0.3, 0.5). Note the strong positive correspondence at zero offset for all three $\eta$ values.*

(top panel) to a larger (bottom panel) value of $\eta$, the 2D correlation signal becomes less smooth, gaining spatial resolution. Thus, the value of $\eta$ can be used to control the degree of spatial resolution of the 2D correlation signal and, therefore, the scale at which we want to look at the correspondence between events in the two signals. For example, the larger value of $\eta$ in the bottom panel results in a more finely grained examination of the spatial and temporal coordination between the signals. Furthermore, it is interesting to note that, regardless of the value of $\eta$, a roughly straight red line can be seen along offset zero, what stresses the high synchronization between the two signals.

We believe the visual inspection of the correlation map provides a flexible way of exploring the spatiotemporal coordination between channels of behavioral data. The actual computation of the correlation signal is quite fast which allows one to quickly try different values of $\eta$ and choose a suitable one. This means the correlation algorithm can be tuned to the behavioral data under analysis. This is important, since a too sensitive function can be confused by high frequency noise in the signals, whereas a less sensitive function can miss subtle changes in the spatiotemporal coordination between signals.

## 4. Summary

This paper presented an algorithm for the computation of the instantaneous correlation coefficient between two signals as a function of both time and the temporal offset between the signals. The computation of the instantaneous correlation can be ultimately seen as a linear filtering operation, and two types of filters were formulated, namely, a causal and a non-causal filter. The main advantage of the causal approach is that it can be used in real time applications. In turn, the non-causal approach has the advantage of using a symmetric filter which produces a correlation signal where each sample is equally affected by its two (left and right) vicinities. Regardless of the type of filter used (causal or non-causal), the algorithm produces a 2D correlation signal that, when plotted, allows rapid visualization of the correspondence between the signals with points of high correlation (either positive or negative) emerging as deep color regions.

The filter used by the correlation algorithm is adjustable, making it possible to tune the filter to the data under analysis. By controlling the value of a single parameter, it is possible to obtain either a short memory filter that is more sensitive to high frequency changes in the correspondence between signals,

or a long memory filter that is more sensitive to low frequency changes.

We believe the techniques presented in this work constitute a useful tool for the analysis of audiovisual speech in that they provide a flexible means of exploring the spatiotemporal coordination between channels of audiovisual speech data. Although the results presented here focus on orofacial motion data, the techniques can be applied to any type of signal.

## 5. References


[1] H. C. Yehia, P. Rubin, and E. Vatikiotis-Bateson, "Quantitative association of vocal-tract and facial behavior," *Speech Communication*, vol. 26, no. 1–2, pp. 23–43, October 1998.

[2] H. C. Yehia, T. Kuratate, and E. Vatikiotis-Bateson, "Linking facial animation, head motion and speech acoustics," *Journal of Phonetics*, vol. 30, pp. 555–568, 2002.

[3] A. V. Barbosa, H. C. Yehia, and E. Vatikiotis-Bateson, "Temporal characterization of auditory-visual coupling in speech," in *Proceedings of Meetings on Acoustics*, vol. 1, 2007, pp. 1–14.

[4] R. M. Aarts, R. Irwan, and A. J. E. M. Janssen, "Efficient tracking of the cross-correlation coefficient," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 6, pp. 391 – 402, September 2002.

[5] A. V. Barbosa and E. Vatikiotis-Bateson, "Video tracking of 2D face motion during speech," in *Proceedings of the IEEE International Symposium on Signal Processing and Information Technology – ISSPIT'2006*, Vancouver, Canada, August 2006, pp. 791–796.

[6] A. V. Barbosa, H. C. Yehia, and E. Vatikiotis-Bateson, "Linguistically valid movement behavior measured non-invasively," in *Proceedings of the International Conference on Auditory-Visual Speech Processing – AVSP 2008*, Tangalooma, Australia, September 2008.