Conditioning Multimodal Information for Smart Environments

D. Looney*
and N. Ur Rehman
and D. Mandic
Imperial College London
Emails: {david.looney06,
naveed.rehman07,
d.mandic}@ic.ac.uk

T. M. Rutkowski
RIKEN Brain Science Institute,
Saitama, Japan
Emails: tomek@brain.riken.jp

A. Heidenreich
and D. Beyer
Siemens AG, Germany
Emails: {alla.heidenreich,
dagmar.beyer}@siemens.com

***Abstract*—This study aims at providing signal processing solutions for the conditioning of multimodal information in audio-aided smart camera environments. A novel approach is introduced for processing audio and video within a unified 'data fusion via fission' framework. This is achieved using empirical mode decomposition (EMD), a fully data-driven algorithm which facilitates analysis at multiple time-frequency scales. Its adaptive nature makes it suitable for processing real-world data and allows, for example, signal conditioning (denoising, illumination invariant video) and robust feature extraction. Furthermore, complex extension of the EMD algorithm are used to quantify shared dynamics between the conditioned modalities facilitating multimodal fusion. The proposed collaborative approach is used to model human-human interaction.**

***Index Terms*—communication atmosphere, multimodal analysis, empirical mode decomposition (EMD), data fusion via fission**

I. Introduction

The growing interest in human computer interfaces (interactive virtual environments, surveillance) has highlighted the need for smart environments capable of accurately detecting and modeling human activity. A key application is the analysis of human communication in which the goal is to estimate parameters within the framework of the so-called communication atmosphere. In this way, any conversation episode can be defined based on the quality of human interaction and common understanding between the communicators. The applications of such technology can be used to monitor real or distance lectures, and group discussion meetings.

The evaluation of the communication atmosphere is primarily dependent on robust feature extraction from data modalities (detection of facial expressions and body language in video or speech in audio). Feature extraction, however, is often critically sensitive to external "ambient" phenomena such as noise. Data conditioning is prerequisite in achieving accurate information segregation and low error rates. This is made difficult in practice as standard signal processing algorithms, such as the discrete Fourier transform (DFT) or discrete cosine transform (DCT), make assumptions of linearity and stationarity which do not hold for modalities such as video and audio. A robust framework should therefore use feature extraction algorithms suitable for real world data.

As illustrated by the multimodal approaches in [1], [2], [3], fusion of features from both video and audio enables a significantly more accurate evaluation of the communication atmosphere than can be achieved by the features of one modality alone. Although a multimodal approach facilitates enhanced analysis, it introduces new challenges. Features are typically obtained through the use of carefully selected algorithms that are suitable only for a specific modality (non-negative matrix factorization for video and mel-frequency cepstral coefficients for audio) so that they are not directly compatible. The estimation of synchronised activity between the modality features is crucial to the evaluation of the communication atmosphere. In practice, this is achieved using linear approximations of entropy by calculating feature co-variance matrices [4] so as to estimate mutual information within the framework of information theory [5]. Often, however, crucial information is contained in higher order (nonlinear) signal statistics and linear approaches used in previous studies are thus not adequate to fully establish shared feature dynamics [6].

To address these concerns, a unified framework to perform conditioning, feature extraction and information fusion for multimodal data in communication events using empirical mode decomposition (EMD) [7] is proposed. EMD is a fully data driven algorithm which decomposes data into a set of oscillations, known as intrinsic mode functions (IMFs). Unlike Fourier or wavelet methods, EMD makes no prior assumptions of the data [7] facilitating the analysis of nonstationary and nonlinear signals [8], [9]. The IMFs are narrow band by design facilitating highly localised analysis in time and frequency. Thus the algorithm is suitable for the considered modalities as features (speech, noise, texture, incident illumination) correspond to specific variations in temporal/spatial frequencies. Furthermore it is shown how recent complex extensions of the algorithm [10], [11], which enables an IMF by IMF comparison for a pair of sources, can be used to assess shared dynamics between the modalities in time and frequency to establish the quality of communication. Simulations on audio and video for a communication event support the analysis.

978-1-4244-4620-9/09/$25.00

II. Communication Atmosphere

The communication atmosphere refers to a recently developed framework that interprets multimodal features to model communication situations [1], [2], [3], [12]. Specifically, this paper considers face-to-face conversation scenarios between two participants. Two sensory modalities, video and audio, are obtained for a given communication episode and conditioning/feature extraction is used to obtain communication-related features. In the case of video, features are extracted in regions of interest (ROIs) such as the face from the smoothed difference between consecutive frames (visual flow) and, in the case of audio, speech features are obtained.

Synchronised activity between the modality features can be evaluated to identify the roles of the communication members as either a primary information sender (speaker) or as a receiver (listener). Furthermore, shared modality activity can also be used to evaluate the quality of information flow between the communicators, known as communication efficiency [13], [14]. Communication is defined as efficient when the intended message sent by the sender is transmitted and received by the receiver. It reflects the ability of the participants to interact and is a crucial qualitative measure of a communication episode. The framework is illustrated in Fig. 1.

III. Empirical Mode Decomposition

Empirical mode decomposition [7] is a technique which adaptively decomposes a given signal into a finite set of AM/FM modulated components. These components, called "intrinsic mode functions" (IMFs), represent the oscillation modes embedded in the data. The IMFs act as a naturally derived set of basis functions for the signal; EMD can thus be seen as an exploratory data analysis technique. In fact, EMD and the Hilbert-Huang transform comprise the so-called "Hilbert spectral analysis" [7]; a unique spectral analysis technique employing the concept of instantaneous frequency. In general, the EMD aims at representing an arbitrary signal via a number of IMFs and the residual. More precisely, for a real-valued signal $x(t)$, the EMD performs the mapping

$$x(t) = \sum_{i=1}^{M} c_i(t) + r(t) \tag{1}$$

where the $c_i(t)$, $i = 1, \ldots, M$ denote the set of IMFs and $r(t)$ is the trend within the data (also referred to as the last IMF or residual). By design, an IMF is a function which is characterized by the following two properties: the upper and lower envelope are symmetric; and the number of zero-crossings and the number of extrema are exactly equal or they differ at most by one.

The first IMF is obtained as follows [7].

1) Let $\tilde{x}(t) = x(t)$;
2) Identify all local maxima and minima of $\tilde{x}(t)$;
3) Find an "envelope," $e_{min}(t)$ (resp. $e_{max}(t)$) that interpolates all local minima (resp. maxima);
4) Extract the "detail," $d(t) = \tilde{x}(t) - (1/2)(e_{min}(t) + e_{max}(t))$;
5) Let $\tilde{x}(t) = d(t)$ and go to step 2); repeat until $d(t)$ becomes an IMF.

Once the first IMF is obtained, the procedure is applied iteratively to the residual $r(t) = x(t) - d(t)$ to obtain all the IMFs. Following the sifting process, the Hilbert transform can be applied to each IMF separately. This way, it is possible to generate analytic signals, having an IMF as the real part and its Hilbert transform as the imaginary part, that is $x + j\mathcal{H}(x)$ where $\mathcal{H}$ is the Hilbert transform operator. Equation (1) can therefore be augmented to its analytic form given by

$$X(t) = \sum_{i=1}^{M} a_i(t) \cdot e^{j\theta_i(t)} \tag{2}$$

where the trend $r(t)$ is purposely omitted, due to its overwhelming power and lack of oscillatory behavior. Observe from (2), that now the time dependent amplitude $a_i(t)$ can be extracted directly and that we can also make use of the phase function $\theta_i(t)$. Furthermore, the quantity $f_i(t) = \frac{d\theta_i}{dt}$ represents the instantaneous frequency [15]; this way by plotting the amplitude $a_i(t)$ versus time t and frequency $f_i(t)$, we obtain a time-frequency-amplitude representation of the entire signal called the Hilbert spectrum. It is this combination of the concept of instantaneous frequency and EMD that makes the framework so powerful as a signal decomposition tool.

For convenience, we refer to a matrix which represents the time-frequency-amplitude activity of a set of IMFs in matrix form as a Hilbert matrix, $\mathbf{H}_{[N_F \times N_T]}(i, j)$, where the matrix dimension is $N_F \times N_T$ where N_F and N_T represent the resolution of the matrix in frequency and time respectively. In other words, matrix entry $\mathbf{H}_{[N_F \times N_T]}(i, j)$ denotes the summation of all IMF amplitudes of frequency i at time j.

A. Complex Extensions of EMD

Several extensions of EMD to the field of complex numbers have been recently developed. These include "Complex Empirical Mode Decomposition" [16], "Rotation Invariant Empirical Mode Decomposition (RIEMD)" [10] and "Bivariate Empirical Mode Decomposition (BEMD)" [11]. However, only RIEMD and BEMD operate directly in $\mathbb{C}$ making them suitable in practical applications [17]. In particular, BEMD facilitates enhanced local mean estimation compared to RIEMD [17] and was used in analysis.

In order to obtain a set of M complex/bivariate IMFs, $\gamma_i(t)$, $i = 1, \ldots, M$, from a complex signal $z(t)$ using bivariate EMD, the following procedure is adopted [11]:

1) Let $\tilde{z}(t) = z(t)$;
2) To obtain K signal projections, given by $\{p_{\theta_k}\}_{k=1}^{K}$, project the complex signal $\tilde{z}(t)$, by using a unit complex number $e^{-\jmath\theta_k}$, in the direction of θ_k, as

$$p_{\theta_k} = \Re\big(e^{-\jmath\theta_k}\tilde{z}(t)\big),\ k = 1, \ldots, K \tag{3}$$

where $\Re(\cdot)$ denotes the real part of a complex number, and $\theta_k = 2k\pi/K$;

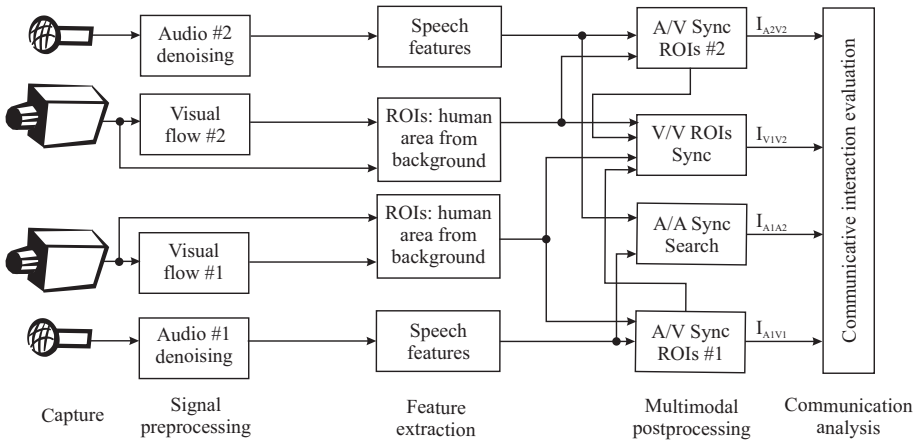


Fig. 1. Framework for evaluation of the communication atmosphere.

3) Find the locations $\{t_j^k\}_{k=1}^K$ corresponding to the maxima of $\{p_{\theta_k}\}_{k=1}^K$;
4) Interpolate (using spline interpolation) between the maxima points $[t_j^k,\ \tilde{z}(t_j^k)]$, to obtain the envelope curves $\{e_{\theta_k}\}_{k=1}^K$;
5) Obtain the mean of all the envelope curves, $m(t)$, and subtract from the input signal, that is, $d(t) = \tilde{z}(t) - m(t)$. Let $\tilde{z}(t) = d(t)$ and go to step 2); repeat until $d(t)$ becomes an IMF.

Similarly to real-valued EMD, once the first IMF is obtained, $\gamma_1(t)$, the procedure is applied iteratively to the residual $r(t) = z(t) - d(t)$ to obtain all the IMFs. For more detail, refer to [11].

B. Trivariate EMD

The recently introduced Trivariate EMD (TEMD) [18] further extends the theory of EMD to consider trivariate and three dimensional (3D) signals. As with bivariate EMD [11], a major challenge is local mean estimation of the input data. In the case of TEMD, local mean estimation is achieved by projecting the data in 3D space using suitable direction vectors governed by a unit sphere. The extrema of these signal projections can then be interpolated using a component-wise spline interpolation, yielding 3D envelope curves which can be conveniently represented by quaternion algebra. Quaternion algebra is a well established mathematical framework for modeling 3D rotations and is used extensively in adaptive filtering and robotics. For more detail, refer to [18].

IV. EMD and Information Fusion

Data and information fusion is the approach whereby data from multiple sensors or components is combined to achieve improved accuracies and more specific inferences that could not be achieved by the use of only a single sensor [19]. Its principles have been employed in a number of research fields including information theory, signal processing and computing [20], [19], [21], [22]; an overview can be found in [23].

Recent work [6] demonstrates that the decomposition nature of EMD provides a unifying framework for "information fusion via fission," where fission is the phenomenon by which observed information is decomposed into a set its components (see Fig. 2). More specifically, the stages of Signal Processing, Feature Extraction and Situation Assessment from the waterfall model, the well established fusion model, can all be achieved by EMD.

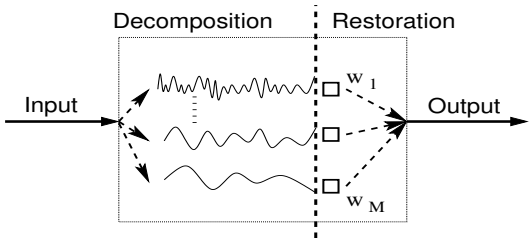


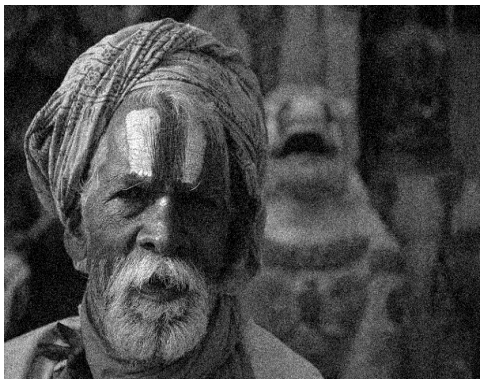
Fig. 2. Spatio-temporal fusion

To illustrate the decomposition and fusion properties of EMD, consider the original noisy video frame given in Fig. 3 (a). Decomposition of the video frame is achieved by applying the algorithm to an image vector constructed by concatenating either the frame rows or its columns. The IMFs for the vector can be converted into the 2D form of the original frame, producing a set of M scale images. The summation of the first 5 IMFs is given in Fig. 3 (b) and the remaining IMFs is given in Fig. 3 (c). Note how each of the scales represents different properties of the image. The higher index scales contain high frequency detail such as noise and the image edges, while slowly oscillating effects such as illumination are contained within the low index scales.

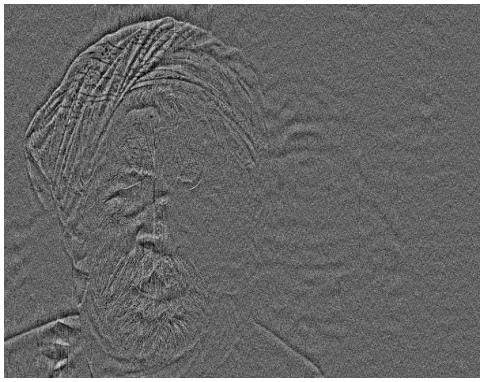
V. Illumination Invariance using EMD

Before introducing the general feature extraction EMD framework for audio and video, a conditioning procedure unique to video is described, that of illumination invariance.

Light on a surface produces complex artifacts which cause local variations in illumination intensity and colour, thus making it difficult to determine the original 'surface image'



(a) Original video frame



(b) High frequency detail contained in $\sum_{i=1}^{5} c_i$

(c) Low frequency detail contained in $\sum_{i=6}^{19} c_i$

Fig. 3. Illustration of the sifting process via EMD for a frame of video (a). Note how each of the IMFs c_1 - c_{19} represents the frequency scales within the image. The higher index IMFs contain high frequency detail such as noise and the image edges while slowly oscillating effects such as illumination are contained within the low index IMFs.

and reducing the performance accuracy for a number of rudimentary vision operations that can accurately estimate image features. Once the regions of interest (the faces of the subjects) have been detected, it is proposed to use a novel EMD-based methodology to remove the effects of illumination. EMD is a natural choice as illumination is likely to be captured in one or more low frequency scales. Indeed, it was illustrated in [24] that the algorithm can be used as a preprocessing step to achieve significant normalization of facial images subject to illumination variation. Although effective, the approach only considers grayscale images.

In the computer vision community, a significant challenge is shadow removal of colour images which requires simultaneous analysis across all colour channels. However, as a consequence of its data driven nature, standard EMD is not directly suitable for the task as it suffers from the problem of uniqueness. That is, there is no guarantee that decompositions of different sources are matched either in number or their properties (frequency). This makes it difficult to compare decompositions from multiple sources or in the case of shadow removal - multiple colour channels. It is thus proposed to use trivariate EMD (TEMD) [18], a recently developed extension of EMD for trivariate data.

Trivariate EMD facilitates the simultaneous decomposition of three dimensional signals into a single set of three dimensional IMFs (κ_i, for $i = 1 \dots M$) using quaternion algebra. Thus it is proposed to perform the following operation

$$\kappa = \mathbf{TEMD}(r, g, b) \tag{4}$$

and obtain a set of M trivariate IMFs for a trivariate signal constructed from the red (r), green (g) and blue (b) components of an image. Separating the IMF components gives three sets of real valued IMFs which are matched in frequency and number [18], thus facilitating a robust comparison and analysis between the colour channels [18].

Incident shadow will dominate one or more of the low frequency IMF components, the effects of which can be reduced by replacing the IMF (or residue) with a uniform function. The analysis was performed on the original images shown in Fig. 4.

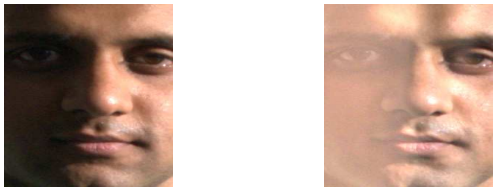
VI. Feature Extraction

It is proposed to use EMD to extract features from both video and audio. The differences in dimensionality can be addressed by applying Hilbert analysis, so that a $1 \times T$ feature vector is obtained for each modality where $t = 1, \dots, T$ is the total number of video frames.

In the case of video, the ROI for each differential frame $\mathbf{d}_{[h \times w]}(t)$ (see Appendix) for the $\#k$ communication member and frame $\#t$ is decomposed using EMD and a Hilbert matrix, $\mathbf{Hv}_{[N_F \times N_T]}(i, j, t, k)$, is obtained. The feature vector is thus given by

$$V_k(t) = \sum_{i=\alpha_1}^{\alpha_2} \sum_{j} \mathbf{Hv}_{[N_F \times N_T]}(i, j, t, k) \tag{5}$$

(a) Original (b) Shade Removed



(a) Original (b) Shade Removed

Fig. 4. Shadow removal for colour images using trivariate EMD.

where α_1 and α_2 are empirically determined parameters which define the instantaneous frequency range used to construct the feature vector. For example, high frequency components can contain noise while very low frequency components contain little information relevant to facial features.

For the case of audio, the raw audio streams for each communication member are decomposed using EMD and the Hilbert matrices, $\mathbf{Ha}_{[N_F \times N_T]}(i,j,k)$, are determined. The audio feature vector for communication member $\#k$ is thus given by

$$A_k(t) = \sum_{i=\epsilon_1}^{\epsilon_2} \sum_{j=\rho(t-1)+1}^{\rho t} \mathbf{Ha}_{[N_F \times N_T]}(i,j,k) \tag{6}$$

where, as before, ϵ_1 and ϵ_2 are spectrum parameters (in this instance determined by the spectrum range of speech) and $\rho = \frac{f_s}{f_{ps}}$, that is the ratio between the sampling frequency of the audio, f_s, and the frames per second of the video, f_{ps}. The feature vectors for the different modalities are shown in Fig. 5. Note the alignment of features for relevant modalities.

VII. Evaluation of the Communication Atmosphere

Once the video and audio features have been determined, it is proposed to use EMD to evaluate the level of shared information between each modality so as to evaluate the communication atmosphere. Mutual information can be defined as shared dynamics across frequency and time, information which is defined locally by the instantaneous amplitudes of the IMF components.

However, as discussed previously, standard EMD is not appropriate in scenarios where it is desirable to compare IMFs from multiple sources. Decompositions, even for signals of similar statistics, are often different in number and properties [17]. Instead it is proposed to use complex extensions of EMD, for example bivariate EMD, which decompose pairs of

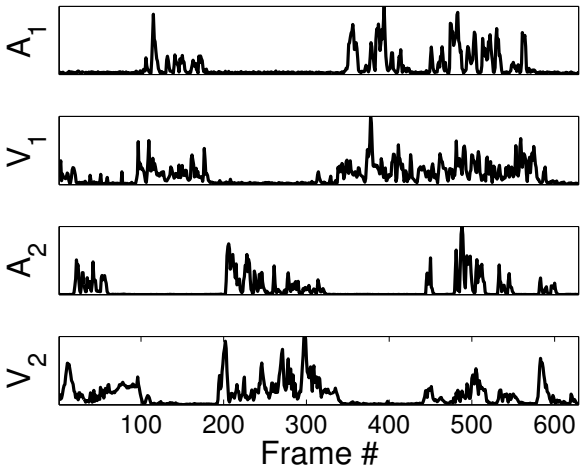


Fig. 5. The feature vectors for the different modalities.

signals simultaneously as a single complex entity. This guarantees that the IMFs are matched in number and properties, thus finding a set of common scales which are unique to the sources being analysed and facilitating fusion [17].

Shared dynamics between modalities V_1 and A_1 are evaluated as follows. Firstly, a complex entity is constructed from the feature modalities and decomposed using bivariate EMD [11]

$$\sum_{i}^{M} \gamma_i = (V_1 + jA_1) \tag{7}$$

The instantaneous amplitudes for the $i = 1, \ldots, M$ IMFs are denoted by $\Re\{a_i\}$ and $\Im\{a_i\}$ for the real and imaginary components respectively. Synchronised activity between each IMF component can be evaluated as

$$\rho_i(t) = \Re\{a_i(t)\}' \times \Im\{a_i(t)\}' \tag{8}$$

where $\Re\{a_i(t)\}'$ and $\Im\{a_i(t)\}'$ denote, respectively, $\Re\{a_i(t)\}$ and $\Im\{a_i(t)\}$ normalised between 0 and 1. Similar to the manner in which the Hilbert transform matrix is constructed, the 'time-frequency-synchronised activity' information can be represented by a $N_F \times N_T$ matrix $\mathbf{I}_{V_1A_1[N_F \times N_T]}(i,t)$ where, as before, N_F and N_T represent the resolution of the matrix in frequency and time respectively. [1]

The level of synchronised activity can thus be evaluated as

$$I_{V_1A_1}(t) = \sum_{i=\delta_1}^{\delta_2} \mathbf{I}_{V_1A_1[N_F \times N_T]}(i,t) \tag{9}$$

where δ_1 and δ_2 are frequency range parameters. In a similar fashion, the level of synchronised activity for communicator $\#2$ can be evaluated as

$$I_{V_2A_2}(t) = \sum_{i=\delta_1}^{\delta_2} \mathbf{I}_{V_2A_2[N_F \times N_T]}(i,t) \tag{10}$$

[1]For accuracy, low levels of white Gaussian noise was introduced into each component of the complex vector $(V_1 + jA_1)$ and the methodology was averaged over several simulations. A noise assisted decomposition can facilitate a more natural separation of the scale components. For more detail we refer to [25].

The level of simultaneous activity in video only can be estimated as:

$$I_{V_1V_2}(t) = \sum_{i=\delta_1}^{\delta_2} \mathbf{I}_{V_1V_2[N_F \times N_T]}(i,t) \tag{11}$$

where V_1, V_2 are the video feature vectors extracted respectively from ROIs of communicator $\#1$ and communicator $\#2$. Similarly audio activities are evaluated as

$$I_{A_1A_2}(t) = \sum_{i=\delta_1}^{\delta_2} \mathbf{I}_{A_1A_2[N_F \times N_T]}(i,t) \tag{12}$$

Quantities $I_{A_1V_1}$ and $I_{A_2V_2}$ evaluate the local synchronicity between the audio (speech) and video (mostly facial movements) flows. It is expected that the sender should exhibit higher synchronicity due to the associated higher activity. Quantity $I_{V_1V_2}$ is related to possible crosstalks in the video modalities. It is useful in detecting possible overlapping in the activity of communicators that can impair the quality of the evaluation of communicators' role. The level of synchronised activity between the modalities (video and audio) as estimated by bivariate EMD is plotted for 620 frames in Fig. 6.

Combining these estimates of the dynamics shared between the respective modalities can give a temporal measure of the communication efficiency [2], [3]:

$$C(t) = \left(1 - \frac{I_{V_1V_2}(t) I_{A_1A_2}(t)}{2}\right) \left| I_{A_1V_1}(t) - I_{A_2V_2}(t) \right|, \tag{13}$$

The estimated information sender and the communication efficiency for the first 620 frames are plotted in Fig. 7. Note that the sender has been estimated accurately (see Fig. 7(a)). For example, communicator $\#2$ is initially the sender while communicator $\#1$ becomes the sender at approximately frame $\#100$. There is no clear information sender after approximately frame $\#400$. This is caused by both communicators attempting to communicate at once for a duration of approximately 200 frames. This ineffective communication is also reflected by the low communication efficiency for the same time period (see Fig. 7(b)).

VIII. Conclusion

A novel unified framework for the evaluation of the communication atmosphere using empirical mode decomposition (EMD) has been presented. EMD determines the natural oscillations inherent to the data, making it ideal for establishing temporal and spatial frequency scales in both audio and video. Its data driven nature is suitable for nonlinear and nonstationaty signals and facilitates robust conditioning, feature extraction and illumination invariance (video). Within the proposed framework, complex extensions of the algorithm are used to establish the level of shared dynamics between the modality features and establish the communication efficiency, which reflects the quality of interaction between the communicators for a communication episode.

APPENDIX

A. Differential Flow

The visual flow of video is established as follows. Consider two conditioned consecutive video frames $\mathbf{f}_{[h\times w]}(t-1)$ and $\mathbf{f}_{[h\times w]}(t)$, where $h \times w$ denotes pixel dimension. The temporal gradient $\mathbf{G}$, that is, a smoothed difference between the images convolved with a two-dimensional Gaussian filter $\mathbf{g}$ with adjusted standard deviation σ, can be evaluated at pixel (n,m) for time t as:

$$\mathbf{G}_{[h\times w]}(t,n,m) = \left| \sum_{i=1}^{x} \sum_{j=1}^{y} \mathbf{d}_{[h\times w]}(t, n-i, m-j) \mathbf{g}_{[x\times y]}(\sigma, i, j) \right| \tag{14}$$

Variable $\mathbf{d}_{[h\times w]}(t)$ represents the difference between the two consecutive frames, and is given by

$$\mathbf{d}_{[h\times w]}(t) = \mathbf{f}_{[h\times w]}(t) - \mathbf{f}_{[h\times w]}(t-1). \tag{15}$$

References

[1] T. M. Rutkowski and D. Mandic, *Human Computing*, vol. 4451 of *Lecture Notes in Artificial Intelligence*, chapter Modelling the Communication Atmosphere - A Human Centered Multimedia Approach to Evaluate Communicative Situations, pp. 155–169, Springer Berlin & Heidelberg, 2007.

[2] T. M. Rutkowski, V. V. Kryssanov, A. Ralescu, K. Kakusho, and M. Minoh, "An audiovisual information fusion approach to analyze the communication atmosphere," in *Proceedings of the Symposium on Next Generation Approaches to Machine Consciousness - Interaction - Situational and Environmental Information Enforcing Involvement in Conversation, AISB'05*, Hatfield, UK, 12–15 April 2005, University of Hertfordshire, pp. 113–120.

[3] Tomasz Rutkowski, Danilo Mandic, and Allan Barros, "A multimodal approach to communicative interactivity classification," *The Journal of VLSI Signal Processing*, vol. 49, no. 2, pp. 317–328, 2007.

[4] T. M. Rutkowski, S. Seki, Y. Yamakata, K. Kakusho, and M. Minoh, "Toward the human communication efficiency monitoring from captured audio and video media in real environments," in *Proceedings of 7th International Conference on Knowledge-Based Intelligent Information and Engineering Systems KES 2003, Part II*, Oxford, UK, September 3–5 2003, vol. 2774 of *Lecture Notes in Computer Science*, pp. 1093–1100, Springer-Verlag Heidelberg.

[5] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley & Sons, 2001.

[6] D. P. Mandic, M. Golz, A. Kuh, D. Obradovic, and T. Tanaka, *Signal Processing Techniques for Knowledge Extraction and Information Fusion*, Springer, 2008.

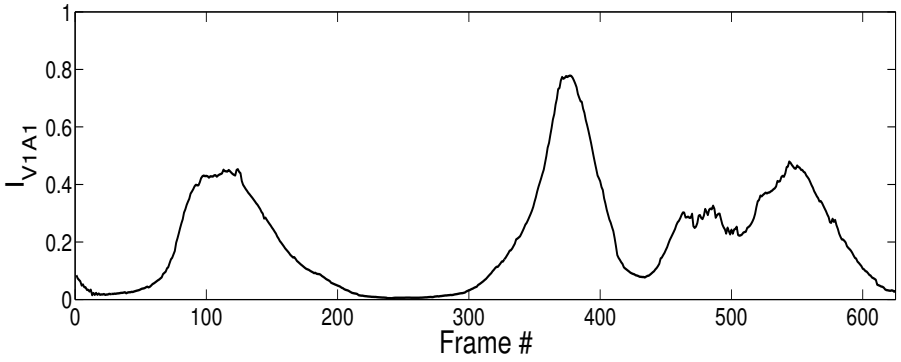
[7] N. E. Huang, Z. Shen, S. R. Long, M. L. Wu, H. H. Shih, Z. Quanan, N. C. Yen, C. C. Tung, and H. H. Liu, "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," *Proceedings of the Royal Society A*, vol. 454, pp. 903–995, 1998.

[8] M. Datig and T. Schlurmann, "Performance and limitations of the Hilbert–Huang transformation (HHT) with an application to irregular water waves," *Ocean Engineering*, vol. 31, pp. 1783–1834, October 2004.

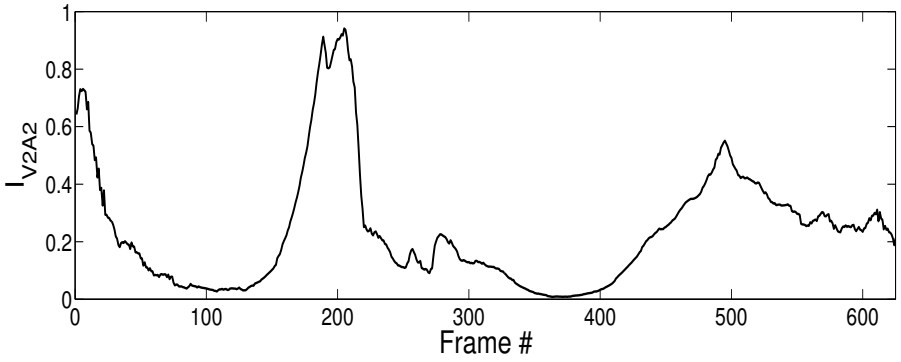
[9] T. M. Rutkowski, R. Zdunek, and A. Cichocki, "Multichannel EEG brain activity pattern analysis in time-frequency domain with nonnegative matrix factorization support," in *International Congress Series*, 2007, vol. 1301, pp. 266–269.

[10] M. Umair Bin Altaf, T. Gautama, T. Tanaka, and D. P. Mandic, "Rotation invariant complex empirical mode decomposition," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2007, vol. 3, pp. 1009–1012.

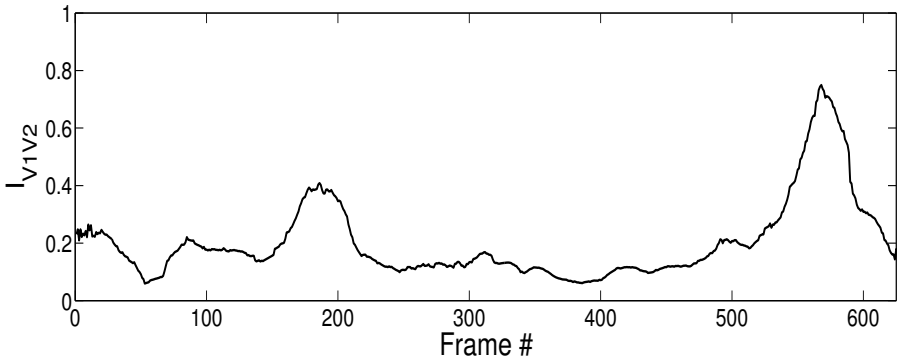
[11] G. Rilling, P. Flandrin, P. Goncalves, and J.M. Lilly, "Bivariate empirical mode decomposition," *IEEE Signal Processing Letters*, vol. 14, pp. 936–939, 2007.



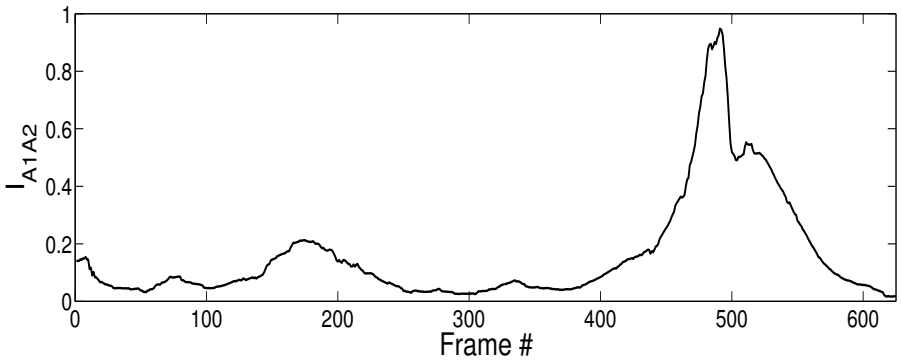
(a) $I_{V_1A_1}$



(b) $I_{V_2A_2}$

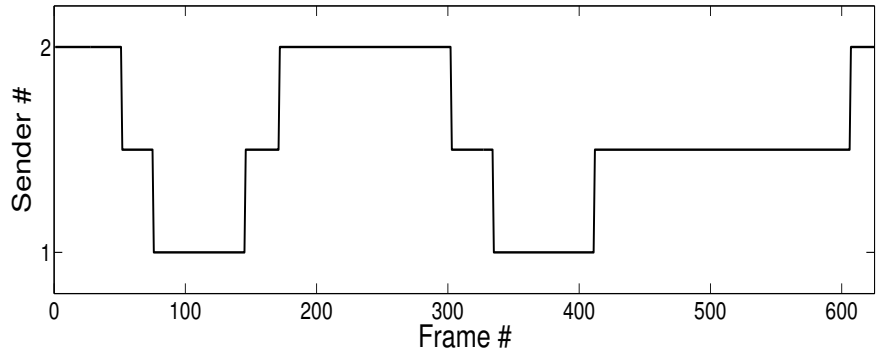


(c) $I_{V_1V_2}$

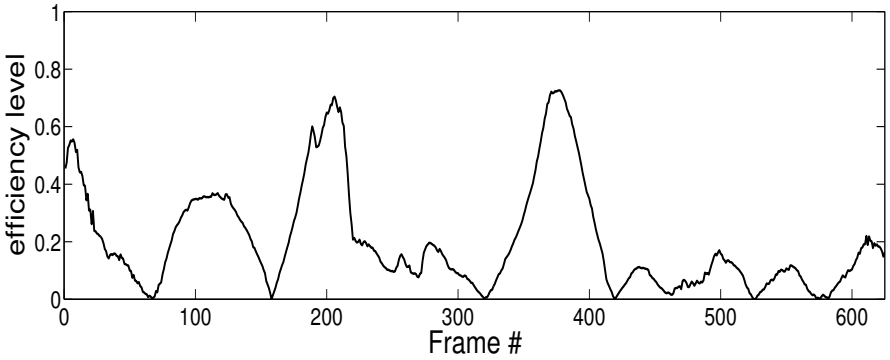


(d) $I_{A_1A_2}$

Fig. 6. The level of synchronised activity between the modalities (video and audio) as estimated by bivariate EMD.



(a) Communication role estimation



(b) Estimation of communication efficiency

Fig. 7. Analysis of the communication atmosphere ('Who is the information sender?', 'Is the communication efficient?') for 620 frames.

[12] T. M. Rutkowski and D. Mandic, "Communicative interactivity - a multimodal communicative situation classification approach," in *Artificial Neural Networks: Formal Models and Their Applications - ICANN 2005, 15th International Conference*, W. Duch, J. Kacprzyk, E. Oja, and S. Zadrozny, Eds. 2005, vol. 3697 of *Lecture Notes in Computer Science*, pp. 741–746, Springer Berlin & Heidelberg.

[13] T. M. Rutkowski, S. Seki, Y. Yamakata, K. Kakusho, and M. Minoh, "Toward the human communication efficiency monitoring from captured audio and video media in real environments," in *Proceedings of 7th International Conference on Knowledge-Based Intelligent Information and Engineering Systems (KES 2003)*, Oxford, UK, September 3–5 2003, pp. 1093–1100, Springer Verlag.

[14] T. M. Rutkowski, K. Kakusho, V. V. Kryssanov, and M. Minoh, "Evaluation of the communication atmosphere," in *Proceedings of 8th International Conference on Knowledge-Based Intelligent Information and Engineering Systems KES 2004, Part I*, Wellington, New Zealand, September 20–25 2004, vol. 3215 of *Lecture Notes in Computer Science*, pp. 364–370, Springer-Verlag Heidelberg.

[15] L. Cohen, "Instantaneous anything," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1993, vol. 5, pp. 105–108.

[16] T. Tanaka and D. P. Mandic, "Complex empirical mode decomposition," *IEEE Signal Processing Letters*, vol. 14, no. 2, pp. 101–104, Feb 2007.

[17] D. Looney and D. Mandic, "Multi-scale image fusion using complex extensions of EMD," *IEEE Transactions on Signal Processing*, vol. 57, no. 1, pp. 1626–1630, 2009.

[18] N. Rehman and D. Mandic, "Empirical mode decomposition for trivariate signals," *IEEE Transactions on Signal Processing (accepted)*, 2009.

[19] D. L. Hall and J. Llinas, "An introduction to multisensor data fusion," *Proceedings of the IEEE*, vol. 85, no. 1, pp. 6–23, 1997.

[20] D. F. Group, "Functional description of the data fusion process," Technical report, Office of Naval Technology, 1992.

[21] L. Wald, "Some terms of reference in data fusion," *IEEE Transactions on Geosciences and Remote Sensing*, vol. 37, no. 3, pp. 1190–1193, 1999.

[22] E. Waltz and J. Llinas, *Multisensor Data Fusion*, Artech House, 1990.

[23] D. P. Mandic, D. Obradovic, A. Kuh, T. Adali, U. Trutschell, M. Golz, P. De Wilde, J. Barria, A. Constantinides, and J. Chambers, "Data fusion for modern engineering applications: An overview," in *Proceedings of the IEEE International Conference on Artificial Neural Networks (ICANN'05)*, 2005, pp. 715–721.

[24] R. Bhagavatula and M. Savvides, "Analyzing facial images using empirical mode decomposition for illumination artifact removal and improved face recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'07)*, 2007, vol. I, pp. 505–508.

[25] Z. Wu and N. E. Huang, "Ensemble empirical mode decomposition: A noise-assisted data analysis method," Tech. Rep. 193, Center for Ocean-Land-Atmosphere Studies, 2004.