Look at Who's Talking: Voice Activity Detection by Automated Gesture Analysis

Marco Cristani[1,2], Anna Pesarin[1], Alessandro Vinciarelli[3,4], Marco Crocco[2], and Vittorio Murino[1,2]

[1] Dipartimento di Informatica, University of Verona,Italy
[2] Istituto Italiano di Tecnologia, Italy
[3] University of Glasgow, UK
[4] Idiap Research Institute, Switzerland
marco.cristani@univr.it anna.pesarin@univr.it
Alessandro.Vinciarelli@glasgow.ac.uk marco.crocco@iit.it
vittorio.murino@iit.it

Abstract. This paper proposes an approach for Voice Activity Detection (VAD) based on the automatic measurement of gesturing. The main motivation of the work is that gestures have been shown to be tightly correlated with speech, hence they can be considered a reliable evidence that a person is talking. The use of gestures rather than speech for performing VAD can be helpful in many situation (e.g., surveillance and monitoring in public spaces) where speech cannot be obtained for technical, legal or ethical issues. The results show that the gesturing measurement approach proposed in this work achieves, on a frame-by-frame basis, an accuracy of 71 percent in distinguishing between speech and non-speech.

1 Introduction

It is common experience to observe that people accompany speech with *gestures*, the "[...] *range of visible bodily actions that are, more or less, generally regarded as part of a person's willing expression* [...]" [10]. Far from being independent phenomena, speech and gestures are so tightly intertwined that every important investigation of language has taken gestures into account, from *De Oratore* by Cicero (1^{st} Century BC) to the latest studies in cognitive sciences [9, 11, 14] showing that the two modalities are "[...] *components of a single overall plan* [...]" [10].

Hence, this paper proposes the detection of gesturing as a means to perform Voice Activity Detection (VAD), i.e. to automatically recognize whether a person is speaking or not. The main rationale is that audio, the most natural and reliable channel when it comes to VAD, might be unavailable for technical, legal, or privacy related issues. A condition that applies in particular to surveillance scenarios where people are monitored in public spaces and are not necessarily aware of being recorded.

Several approaches have exploited the relationship between speech and other cues to accomplish different technological tasks. The synchronization between

pitch and gestures has been used to make artificial agents more realistic [13]. Multimodal speaker diarization techniques (detection of *who speaks when*) based on the joint modeling of speech, facial and bodily cues (e.g., mouth movement, fidgeting, body pose, etc.) have been proposed in [1, 6, 8, 15–18]. To the best of our knowledge, the only work where diarization has been tried with solely visual cues is in [7], where the experiments measure the performance decrease when the audio is absent.

This paper aims at performing VAD with solely visual cues, but it considers a scenario more challenging than the one proposed in [7] for diarization. While the experiments of the latter work are performed in a smart meeting room setting (multiple cameras capturing each person individually at close distance), the results of this paper have been obtained in a surveillance scenario where there is only one camera positioned 7 meters above the scene (see Figure 2 for an example). In particular, the experiments focus on people involved in standing conversations, with a tracker that follows each individual. (see in [5] how groups of interacting people are detected). The VAD approach is based on a local video descriptor that extracts the body optical flow, encoding its energy and "complexity" using an entropy-like measure. This allows one to discriminate between body oscillations or noise introduced by the tracker, where the optical flow is low and homogeneous, and genuine gestures, where the movement of head, arms and trunk produces a local flow field which is diverse in both intensity and direction. The descriptor extracted for each participant produces a signal that can be used for VAD.

The proposed approach is interesting under three main respects. The first is that the relationship between speech and gestures has been widely documented and studied, but relatively few quantitative investigations of the phenomenon have been made. The second is that approaches like the one proposed here might help to infer information about privacy protected data (speech in this case) from publicly accessible data (gestures in this case). This is important to establish whether the simple absence of a certain channel is sufficient to protect the privacy of people and how much. The third is that inferring missing data from available ones can make techniques dealing with challenging scenarios more effective and reliable.

The rest of the paper is organised as follows. In Sect. 2, the VAD methods is described, detailing the entropy-like measure utilised. Section 3 illustrates the experimental trial on a publicly available video dataset and, finally, conclusive remarks and future activities are summarised in Section 4.

2 Gesturing Activity Measurement

This section describes the technique adopted to measure the gesturing activity in videos like those shown in Figure 2. Once a group of interacting individuals has been detected (see [5] for the technique applied), each person is tracked individually and a square *Region of Interest* (ROI) is defined around her. The size of the ROI is set automatically to include all gestures of the individual. Areas

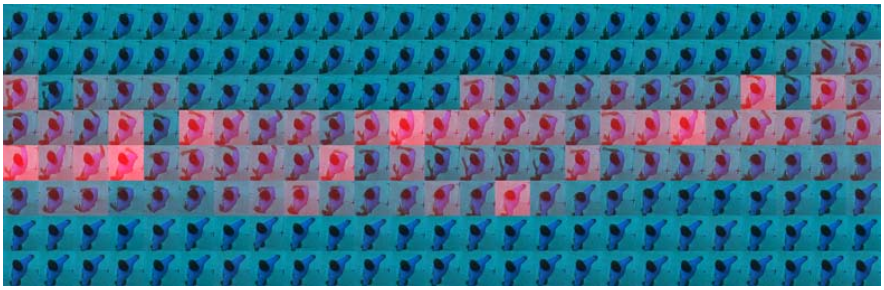

Fig. 1. Qualitative analysis of our descriptor: in the sequence above, an high tonality of red means great gesture activity.

where multiple ROIs overlap have been ignored to avoid possible confusions between neighboring people.

The measurement technique is applied to each ROI individually and it is expected to accomplish two goals: the first is to discriminate between gestures and postural oscillations typically observed when people stand. The second is to normalize the tracking errors that cause abrupt and spurious shifts of the ROI. The body parts most commonly involved in gesturing are hands, arms, head, and trunk. Their individual movements tend to be very different during gesturing and the measurement values associated to a given ROI try to capture such an aspect:

$$v(t) = \max_{\text{int}}(\{f(t)\}) \times S_{\text{int}}(\{f(t)\}) \times S_{\text{ori}}(\{f(t)\}) \tag{1}$$

where $\{f(t)\}$ is the set of motion flow vectors associated to each pixel of the ROI at time t, $S_{\text{int}}(\{f(t)\})$ is the entropy of the motion flow intensities, and $S_{\text{ori}}(\{f(t)\})$ is the entropy of the orientation values, both calculated over $\{f(t)\}$[5]. The maximum over the flow intensities values $\max_{\text{int}}(\{f(t)\})$ encodes the "energy" associated to the movement, while the two entropic terms serve to highlight those motion flow values which exhibit higher variability in intensity and orientation. In this way, postural oscillations and shifts due to unprecise tracking receive a low score because they cause a global, homogeneous set of intensities and orientations, corresponding to low entropy values. Alternative expressions of $v(t)$ have been considered that use mean and median rather than maximum, or do not include one of the entropy terms. In all cases, the resulting performance is lower than the one obtained with the expression above. A graphical idea of the measurement is given in Figure 1 where colours shift towards red when gesturing activity is higher.

[5] The optical flow has been obtained with the package available at the following URL: `http://server.cs.ucf.edu/~vision/source.html`.

Fig. 2. Some frames of the video sequences used.

3 Experiments

The goal of the experiments is twofold: first, to provide a quantitative measure of the correlation between gestures and speech; second, to measure the effectiveness of the function $v(t)$ (see Section 2) in a VAD task. Both tasks have been accomplished over *TalkingHeads*, a new dataset publicly available upon request[6] (see some frames in Figure 2).

The dataset contains four conversations lasting, on average, 6 minutes. The data was recorded in a 3.5×2.5 meters wide outdoor area, during a cloudy day in summer. The total number of subjects is 15 (1 female and 14 males), with 4 different participants per conversation (only one subject participated in two conversations). The subjects include 4 academics, 5 undergraduate students, 2 MSc students, 3 postdoctoral researchers, and 1 PhD student. The ages range between 20 and 40 years and the subjects were unaware of the actual goals of the experiments.

Data were captured at 25 frames per second with a camera positioned 7 meters above the floor and facing downward. The subjects were asked to wear differently colored shirts, in order to make the tracking/localization easier. Tracking has been performed by simple template association. The audio was recorded at 44100 Hz with 4 wireless headset microphones, each transmitting to its own receiver.

Each audio recording has been segmented into speech and non-speech segments using a robust VAD algorithm based on pitch [12]. This latter was extracted at regular time steps of 10 ms with Praat [3], a package including the pitch extraction technique described in [2]. The motivation behind this choice is not only that silence segments are characterized by frequencies way higher than those observed in speech, but also that the pitch tends to be correlated with the "*beat*" gesture typically accompanying syllables where the intonation is stressed [4, 20]. Then, in order to synchronize audio and video data, audio was resampled at 25Hz, averaging the pitch values occurring in each time period. The averaged pitch values constituted the samples of the audio signal that will be analyzed in the following.

[6] `http://profs.sci.univr.it/~cristanm/datasets/TalkingHeads/`

3.1 Pitch-Gesturing Correlation Analysis

This section shows how the correlation between the pitch (as extracted with Praat), and the gesturing activity (as measured with the approach proposed in Section 2) has been measured.

After the application of the techniques described in the previous sections, each video results into two signals per person, showing the value of pitch and $v(t)$ at regular time steps of 40 ms. Plots (a) and (b) of Figure 3 provide an example of such signals. The simple visual inspection shows that the two signals tend to change according to one another. However, $v(t)$ appears to be more noisy of the pitch because of the sensibility of the optical flow. Hence, both signals have been smothed with an average filter applied to 8 s long windows. The result are plots (c) and (d) of Figure 3, showing the smoothed version of pitch and $v(t)$, respectively. The smoothed audio and video signals, normalized with respect to their maximum value, are shown in Fig. 4.

Table 1 reports the Pearson correlation coefficients between $v(t)$ and pitch. Off-diagonal values account for correlations between signals extracted from different individuals. In this way, it is possible to better assess how strong is the correlation between speech and gestures for a given individual.

	A sub.1	A sub.2	A sub.3	A sub.4
V sub.1	**0.7310**	0.1338	0.2490	0.0670
V sub.2	0.1900	**0.6454**	0.4460	0.0254
V sub.3	0.1867	0.1966	**0.4838**	-0.0356
V sub.4	-0.2592	0.0472	0.0389	**0.4204**

Table 1. Quantitative measures: correlation coefficients matrix for Seq. 1 . The matrix rows and columns corresponds respectively to the four subsampled video signals (Vsub) and the four subsampled audio signals (Asub) (the non-significant coefficients (p-value$\geq$ 0.05) are underlined in red.

We performed a similar analysis on the other conversations, with the same parameters, obtaining in total four correlation matrices. Mediating over all the entries in the main diagonal (they were all statistically significant), we obtained a mean correlation score of 0.53, while considering the statistically significant off-diagonals entries we get 0.19. This suggests that $v(t)$ might be a reliable indicator of voice activity. Hence, in the following section, we show how the video signal can be employed to perform VAD.

3.2 Voice Activity Detection

The VAD task proposed in this section consists of labeling each frame as *speech* or $non - speech$. As an approximation, each person is treated independently of the others even though the exchange of turns (the opportunity of speaking)

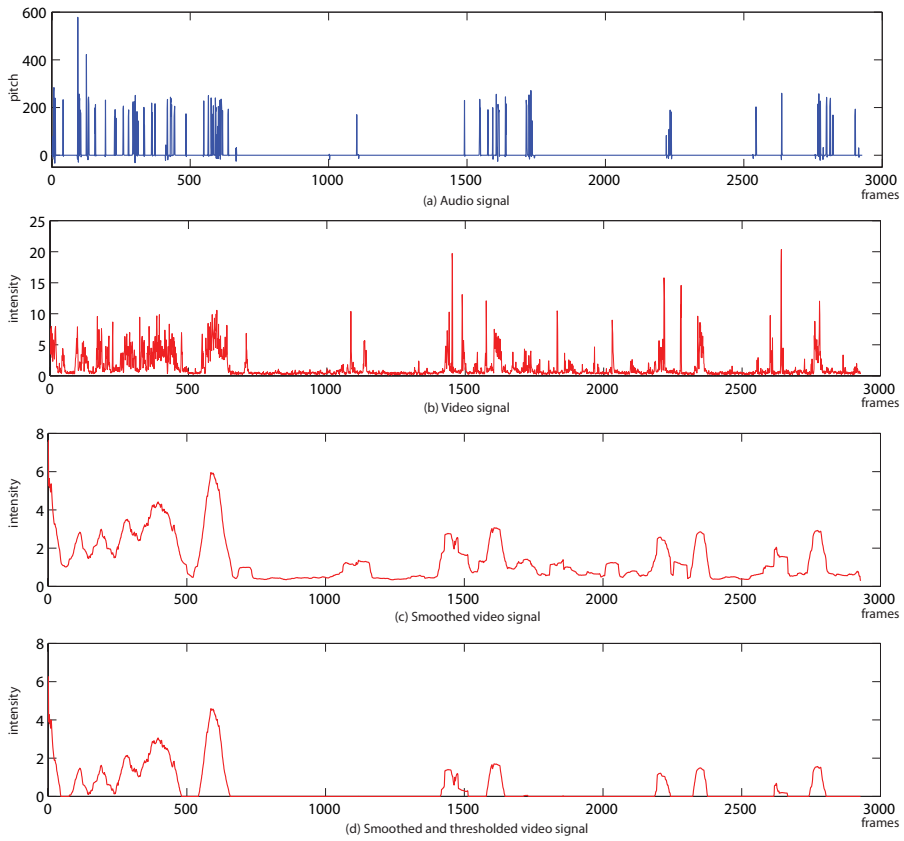

Fig. 3. Examples of signals employed in the analysis. (a) Audio input signal. (b) Video signal produced by our descriptor of a subject involved in the Seq.1. (c) The video signal was smoothed for evaluating the crossmodal correlation (Sec. 3.1). (d) The video signal was thresholded for the audio classification (Sec. 3.2).

tends to follow regularities that might be helpful in improving the performance. The original pitch signal, which has non-zero entries only when the subjects talk, is used as groundtruth.

As a video signal to be used to infer speech, we considered the smoothed signal described above for the correlation analysis. In this way, high frequency components of the original signal have been filtered. The discrimination between speech and non-speech samples has been performed with a thresholding technique. Essentially, as suggested by Fig. 3 and Fig. 4, the video signal has a continuous component caused by small values of optical flow that are always present in the analysis. For this reason, we subtracted the mean to the signal, and we keep the intensities above zero, setting them at 1's. Smoothing and sub-

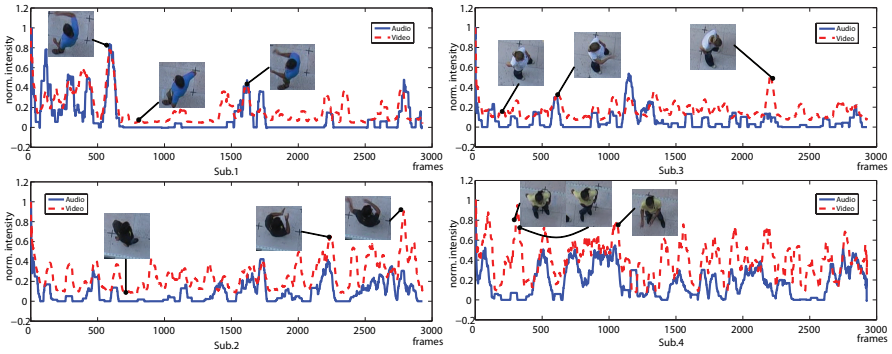

Fig. 4. Visual analysis of the audio and video smoothed data: each plot depicts the smoothed audio (solid blue) and the smoothed video (dashed red) signals for each participant to the dialog. The thumbnails give the feeling of the gesturing activity carried out in a particular instant.

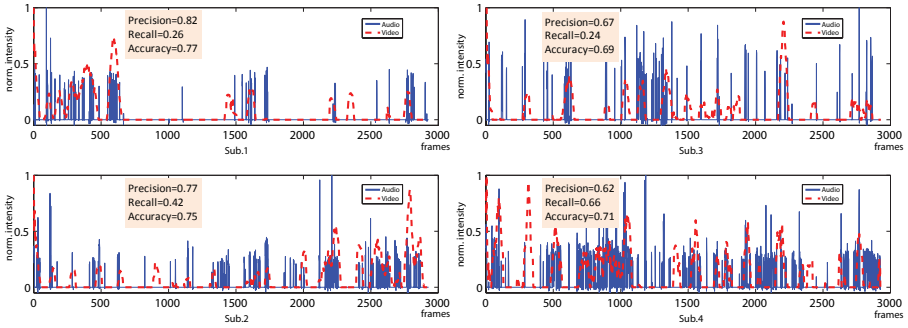

Fig. 5. Audio classification by video analysis. Each plot portrays the audio (solid blue) and the video (dashed red) signals for each participant to the dialog. For the sake of clarity, we report the (normalized) continuous signals, and not their binary versions (that we used). Precision, recall and accuracy scores related to each individual are also indicated.

traction of the mean represent a thresholding operation that does not need the tuning of any parameter.

At this point, we can compare the two signals, and the detailed analysis of Seq. 1 is shown in Fig. 5.

For the sake of clarity, we report in the figure the (normalized) continuous signals, and not their binary versions which were actually used. As visible, many of the speech sampled are correctly captured by the video signal. The figure also

reports the precision, recall and accuracy values. In this sequence, the classifier tends to have low recall and high precision (assuming the speech as positive values). Considering all the subjects employed, we reach an average accuracy of 71%, average precision of 67%, and average recall of 40%.

4 Conclusions

This work has proposed a gesturing-based approach for performing VAD, the automatic detection of people that speak. The reason for using gestures in VAD, typically performed using speech recordings, is that the use of microphones is difficult or illegal in many scenarios of potential interest, including surveillance of public spaces, monitoring of potentially dangerous plants, etc. The core idea behind the approach is that cognitive sciences have demonstrated that speech and gestures, far from being independent expression modalities, are two faces of the same phenomenon. Therefore, gestures can be considered a reliable evidence of speech taking place at the same time.

The preliminary results presented in this paper provide a quantitative confirmation of the finding above and, most importantly, show that the detection of gesturing activity helps to predict whether a person is speaking or not with an accuracy of 71 percent (on a frame-by-frame basis). While not being conclusive about the possibility of reconstructing the actual turns and of performing diarization, the results are certainly promising in the direction of reconstructing conversational dynamics in absence of audio. This appears particularly important as turn-organization has been widely shown to be fundamental in inferring socially important information such as roles, dominance, personality, etc [19].

Besides, this work shows that it is possible to infer information about missing data (speech in this case) from available evidence (videos in this case). In a surveillance setup like the one of the experiments, this opens two conflicting perspectives: on one hand, surveillance approaches can be significantly improved by predicting phenomena considered so far non-accessible with the sensors at disposition. On the other hand, privacy protection measures applied so far (i.e., legal limitation on the use of microphones in public spaces) might become obsolete and uneffective. In this respect, experiments of the type presented in this work might change the notion of privacy and of its protection.

Future work can take two major directions: the first is to move from VAD to full diarization. This requires the application of probabilistic sequential models taking into account temporal constraints between neighboring frames and a larger amount of data. The second is to try automatic conversation analysis based on gestures and to verify whether (and to what extent) it is possible to perform tasks like role recognition, conflict detection, etc., typically performed using turn-organization and other conversational cues.

References

1. Anguera, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., Vinyals, O.: Speaker diarization : A review of recent research. IEEE Transactions on Audio,

Speech, and Language Processing (to appear) (2011)
2. Boersma, P.: Accurate short term analysis of the fundamental frequency and the harmonics to noise ratio of a sampled sound. IEEE Transactions on Image Processing 17, 97–110 (1993)
3. Boersma, P.: Praat, a system for doing phonetics by computer. Glot International 5(9/10), 341–345 (2001)
4. Cassell, J., Steedman, M., Badler, N., Pelachaud, C., Stone, M., Douville, B., Prevost, S., Achorn, B.: Modeling the interaction between speech and gesture. In: Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society. pp. 153–158 (1994)
5. Cristani, M., Bazzani, L., Paggetti, G., A.Fossati, Bue, A.D., Menegaz, G., Murino, V.: Social interaction discovery by statistical analysis of f-formations. In: Proceedings of the British Machine Vision Conference (2011)
6. Fisher, J.W., Freeman, W.T., Darrell, T., Viola, P.: Learning joint statistical models for audio-visual fusion and segregation. Advanced in Neural Inf. Process. Syst. 13, 772–778 (2001)
7. Hung, H., Ba, S.O.: Speech/non-speech detection in meetings from automatically extracted low resolution visual features. In: ICASSP. pp. 830–833 (2010)
8. Hung, H., Huang, Y., Yeo, C., Gatica-Perez, D.: Associating audio-visual activity cues in a dominance estimation framework. In: First IEEE Workshop on CVPR for Human Communicative Behavior Analysis (2008)
9. Kendon, A.: Gesticulation and speech: Two aspects of the process of utterance. The Relationship of verbal and nonverbal communication pp. 207–227 (1980)
10. Kendon, A.: Language and gesture: unity or duality?, pp. 47–63. Cambridge University Press (2000)
11. Kendon, A.: Gesture: Visible Action as Utterance. Cambridge University Press, Cambridge (2004)
12. Khondaker, A., Ghulam, M.: Improved noise reduction with pitch enabled voice activity detection. In: ISIVC2008 (2008)
13. Kopp, S., Wachsmuth, I.: Synthesizing multimodal utterances for conversational agents. Computer Animation and Virtual Worlds 15(1), 39–52 (2004)
14. McNeill, D.: Hand and mind: What gestures reveal about thought. Chicago University Press, Chicago (1992)
15. Noulas, A., Englebienne, G., Krose, B.J.A.: Multimodal speaker diarization. IEEE Transactions on Pattern Analysis and Machine Intelligence 99 (2011)
16. Rao, R., Chen, T.: Cross-modal prediction in audio-visual communication. In: IEEE international Conference on Acoustics, Speech, and Signal Processing. ICASSP-96. vol. 4, pp. 2056–2059 (1996)
17. Siracusa, M.R., John W, F.: Dynamic dependency tests : Analysis and applications to multi-modal data association (2007)
18. Vajaria, H., Islam, T., Sarkar, S., Sankar, R., Kasturi, R.: Audio segmentation and speaker localization in meeting videos. In: 18th International Conference on Pattern Recognition ICPR 2006. vol. 2, pp. 1150 –1153 (2006)
19. Vinciarelli, A., Pantic, M., Heylen, D., Pelachaud, C., Poggi, I., D'Errico, F., Schröder, M.: Bridging the gap between social animal and unsocial machine: A survey of social signal processing. IEEE Transactions on Affective Computing (to appear) (2011)
20. Wells, G., Petty, R.: The effects of over head movements on persuasion. Basic and Applied Social Psychology 1(3), 219–230 (1980)