AN ANALYSIS OF VISUAL SPEECH INFORMATION APPLIED TO VOICE ACTIVITY DETECTION

David Sodoyer, Bertrand Rivet,
Laurent Girin, Jean-Luc Schwartz

Speech Communication Institute
UMR 5009 CNRS, INPG, U3
Grenoble, France

Christian Jutten

Image and Signal Processing Laboratory
UMR 5083 CNRS, INPG, UJF
Grenoble, France

ABSTRACT

We present a new approach to the voice activity detection (VAD) problem for speech signals embedded in non-stationary noise. The method is based on automatic lipreading: the objective is to detect voice activity or non-activity by exploiting the coherence between the speech acoustic signal and the speaker's lip movements. From a comprehensive analysis of lip shape parameters during speech and non-speech events, we show that a single appropriate visual parameter, defined to characterize the lip movements, can be used for the detection of sections of voice activity or more precisely, for the detection of silence sections. Detection scores obtained on spontaneous speech confirm the efficiency of the visual voice activity detector (VVAD).

1. INTRODUCTION

The task of a voice activity detector (VAD) is to assess the presence or the absence of a speech signal in a given acoustic environment. Different methods based on the analysis of the acoustic signal have been proposed in the literature (e.g. [1, 2]). Their weakness is that they generally strongly depend on the acoustic environment, including the nature and the power of possible parasite signals.

Now, speech is a bimodal signal, both acoustic and visual. Since visual speech information is provided by the movements of the speaker's visible articulators, especially lip movements, there exists a specific coherence between acoustic and visual signal features [3]. This coherence has already been exploited in several speech processing applications such as speech enhancement [4, 5], speech sources separation [6, 7], or speech recognition [8].

In this paper, we propose to use visual speech information, namely lip movements, as VAD. Such visual VAD (VVAD) is characterized by a major advantage: Contrary to usual acoustic VADs, it is robust to any acoustic environment (e.g., simultaneous speaker(s), non-stationary background noise, convolutive mixtures, etc.) Thus, the proposed VVAD can be used in any acoustic mixture, including ones with many different speech/audio/noise sources. Previous work on VAD based on visual information can be found in [9]. The authors proposed to model the distribution of the visual information according to two *exclusive* classes: one for speech non-activity (where the visual information is modeled by a single Gaussian) and one for actual speech (where a mixture of only two Gaussian laws is used). Following, for a given visual data, the speech/non-speech decision is taken by likelihood calculation from both distributions. In the present paper, the approach is different since the phenomenology of audio-visual speech is more deeply considered: First, we use a spontaneous speech corpus with natural speech/non-speech sections, and then, we lead a comprehensive analysis of this corpus (Section 2). This analysis shows that the visual spaces related to speech and non-speech sections are actually not exclusive, but on the contrary, strongly overlapping. Thus, in Section 3, we characterize the visual information to be used in the VVAD in dynamical terms. Numerical experiments and detection scores are given in Section 4.

2. LIP-SHAPE ANALYSIS

2.1. Material description

For this study, we designed a dedicated audio-visual corpus of spontaneous speech. Two male French speakers (JL and LG) were set in a spontaneous dialog situation with many speech overlapping and non-speech events. In order to assess the visual VAD, we needed to have the audio signals of the speakers (and possibly other sources) available separately. For this aim, the two speakers were placed and recorded in a different room. They both had a micro-camera (and of course a microphone) focused on the lip region and they could see and hear each other thanks to a monitor screen and headphones providing real-time feedback.

The extracted visual information consists in the time trajectory of basic lip contour geometric parameters, namely

1-4244-0469-X/06/$20.00 ©2006 IEEE

interolabial width A and height B. These parameters were extracted by using the "face processing system" of the ICP [10] which is based on blue make-up, Chroma-Key system and contour tracking algorithms. The parameters are extracted every 20 ms (the video sampling frequency is 50Hz), synchronously with the acoustic signal which is sampled at 16 kHz. Thus, in the following, a signal *frame* is a 20 ms section of acoustic signal associated with a pair of lip parameters (A, B).

2.2. Analysis

Corpus analysis aims at characterizing possible differences on visual patterns between *silence* (defined as vocal inactivity) and *non-silence* sections for a given speaker. We prefer to use this distinction rather than the distinction between speech and non-speech, because non-speech sections are not bound to be silence, since many kinds of non-speech sounds can be produced by the speaker (e.g., laughs, sighs, growls, moans, etc.) To provide an objective reference for the detection, we first manually identified and labeled acoustic sections of silence and non-silence. Then, we defined a normalized video vector as $\mathbf{v}(t) = [A(t)/\mu_A,\ \alpha\, B(t)/\mu_B]^T$ where α is the coefficient of linear regression between $A(t)$ and $B(t)$, μ_A and μ_B the mean values of $A(t)$ and $B(t)$ calculated on the complete corpus for each speaker (^{T} denotes the transpose operator). Fig. 1 represents the distribution of the components $v_1(t)$ and $v_2(t)$ of $\mathbf{v}(t)$ for the non-silence frames (Fig. 1(left)) and the silence frames (Fig. 1(right)) for speaker JL. Fig. 1 (left) is classical for lip-shapes during speech. Fig. 1(right) shows that a large subset of silence frames are grouped around the origin, which correspond to closed lips. Besides, another important subset is located in a non-closed region within the general set of speech shapes provided in Fig. 1 (left). On the other hand, closed lip-shapes are present in both distributions and they cannot be systematically associated with a silence (or non-silence) frame. In summary, there is no direct relationship between silence and closed lips, or speech and open lips: static lip-shape is not sufficient to characterize silence or non-silence.

Now, further investigations revealed that silence frames can be better characterized by the lip-shape movements. Indeed, in silence sections, the lip-shape variations are small. On the contrary, during speech sections these variations are generally quite stronger. Given these observations, we propose to identify the silence sections with one or several visual parameters of dynamical nature.

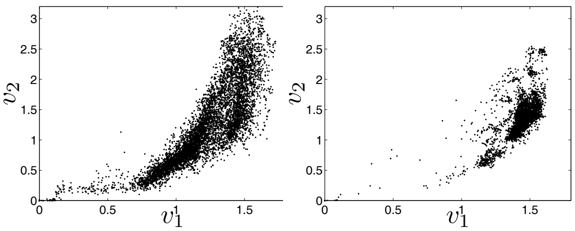

Fig. 1. Distribution of the visual parameter $\mathbf{v}(t)$ for non-silence frames (left) and silence frames (right). Note that 10% and 36% of the points are at the origin for the left and right figures respectively.

3. VISUAL DETECTION OF SILENCE

In pilot experiments, we tested the efficiency of several dynamical parameters. These tests have shown that the sum of the absolute values of the gradient of $v(t)$ components

$$\pi(t) = \left|\frac{\partial v_1(t)}{\partial t}\right| + \left|\frac{\partial v_2(t)}{\partial t}\right| \tag{1}$$

gave the best performances for the detection task described in the following. Thus we retained this dynamical parameter. In spontaneous speech, there may exist some short lip movements in silence, because of, e.g., smiles, funny faces or changes of the "lip rest position". To avoid the association of corresponding $\pi(t)$ values with a speech frame in this case, we propose to integrate these values in time. For this, we filter $\pi(t)$ such as: $\rho(t) = h(t) * \pi(t)$ where $h(t)$ is a truncated first-order low-pass filter defined by: $h(t) = \frac{1}{\tau}\sum_{n=0}^{T-1}\exp(\frac{-n}{\tau})\delta(t - n)$, with T fixed to 100 frames. Now, the principle of the VVAD is to compare, for each frame t, the visual parameter $\rho(t)$ with a threshold λ that remains to be determined. Thus, if $\rho(t) < \lambda$ the frame t is considered as silence, else it is considered as non-silence. We plotted on Fig. 2 the distribution of $\rho(t)$ for the entire corpus, for speaker JL and for three values of the time constant τ: 0.1 (instantaneous case), 5 and 100. We can see that $\rho(t)$ is grossly distributed among three classes for spontaneous speech: from left to right, the first one corresponds to closed lips, the second one corresponds to slow movements and the third one corresponds to fast movements. The goal is to tune the time-integration such as the silence and non-silence distributions of $\rho(t)$ are as much separate as possible, so that the estimation of λ is easy and the visual decision fits well with the manual audio labeling. The different figures show that the choice of the integration duration must be considered carefully: as seen before, a short time constant (or no integration at all, see Fig. 2(left)) makes the detection too sensitive to local perturbations (short lip movements during silence and short stable lip shapes during speech activity.) In contrast, a too long time constant (e.g. on Fig. 2(right)) mixes silence and non-silence sections into a common window, leading to quite unaccurate detection. Eventually, the histogram plotted on Fig. 2(middle) shows that a suitable time constant can largely improve the separation of silence and non-silence sections.

Since the two silence and non-silence classes cannot be completely separated, it is impossible to find a "perfect" threshold λ allowing to correctly detect all silence frames

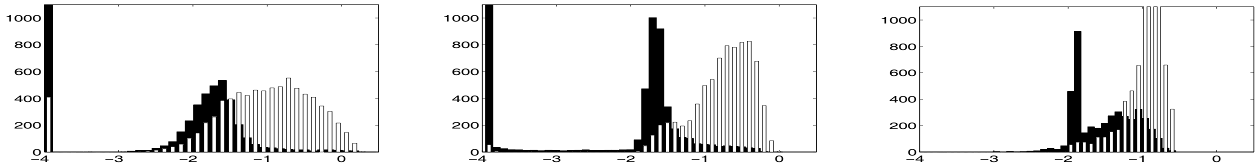

Fig. 2. Histograms of the dynamical visual parameter $\rho(t)$ (on log-scale) for three integration durations: instantaneous (left), suitable value of $\tau = 5$ frames (middle) and too large value of $\tau = 100$ (right). The black and white histograms respectively represent the $\rho(t)$ values associated to silence and non-silence sections respectively.

without making false silence detections. In applications using VAD (e.g., speech enhancement: speech inactivity sections are used to estimate background noise), false detections can dramatically reduce the performances. In this case, it is better for our VVAD to favor the actual detection of silence (i.e. reducing the false detections) rather than reducing the missed detection. This can be done arbitrarily by choosing to detect only the silence sections with long durations. From this last point, we finally retain as silence sections all sections composed of at least N consecutive individual silence frames.

4. QUANTITATIVE ASSESSMENT

4.1. Experimental procedure

We tested the proposed VVAD on about 13200 20ms-frames of the corpus (about 4.4 min of spontaneous speech). We applied the VVAD for different time constants: $\tau = 0.1$, suitable time constant ($\tau = 5$) and too large time constant ($\tau = 100$). For each configuration, we made the threshold λ vary uniformly between the minimum and the maximum of $\rho(t)$. Then, we compared the results of automatic silence frame detection using the VVAD with the reference label provided by the manual identification. This procedure was repeated twice: A first time for the detection of all acoustic silence frames independently and a second time for the detection of acoustic silence composed of at least N consecutive silence frames.

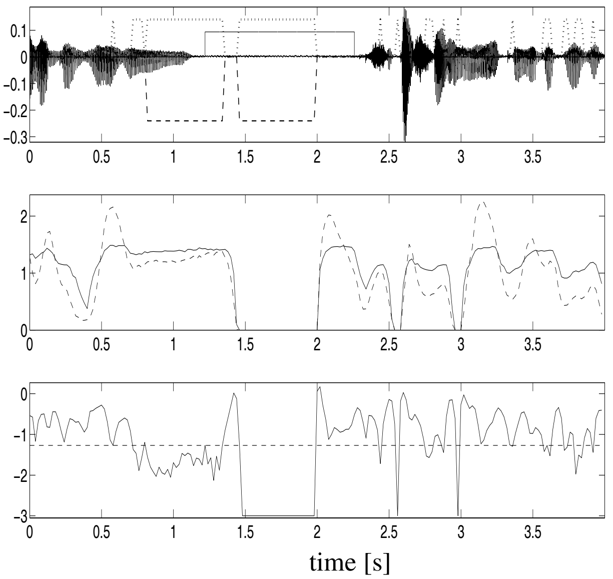

Fig. 3. Silence detection. Top: Acoustic speech signal with silence reference (solid line), frames detected as silence (dotted line) and frames eventually retained as silence when $N = 20$ consecutive silence frames (dashed line). Middle: Static visual parameters v_1 (solid line) and v_2 (dashed line). Bottom: Dynamical visual parameter $\rho(t)$ integrated with $\tau = 5$ (solid line, truncated at -3 for the $-\infty$ value) and the threshold λ (dashed line).

4.2. Results

First of all, Fig. 3 illustrates the different possible relations between visual and acoustic data: movement of the lips in non-silence (e.g., from 0s to 0.5s) and in silence (e.g., just before 1.5s, or between 2s and 2.3s), non-movement of the lips in silence with opened lips (e.g., from 1.2s to 1.4s) and closed lips (from 1.5s to 1.9s), and non-movement in non-silence (from 0.9s to 1.1s). The VVAD, adequately tuned (with $\tau = 5$ and post-processing with $N = 20$), performs quite well. It fails to avoid some false detections (between 0.8s and 1.1s) or to detect some silence sections (between 2 and 2.3s). But it discards several possible false detections in the speech section between 2.3s and 4s, in spite of both closed lips sections and small movements in some regions.

The detection results are presented on Fig. 4 as Receiver Operating Characteristics (ROC). These curves represent the percentage of silence detection (defined as the ratio between the number of detected silence frames and the actual number of silence frames) as a function of the percentage of false silence detection (defined as the ratio between the number of non-silence frames detected as silence frames and the actual number of silence frames). We can see on this fig-

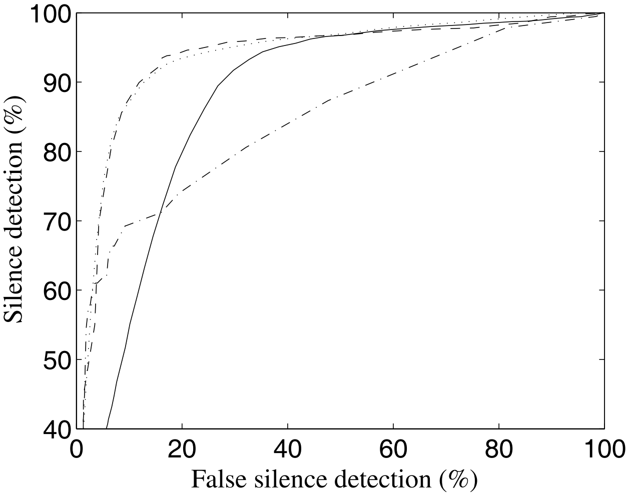

Fig. 4. ROC silence detection curves for: First, two integration durations of the visual parameter $\rho(t)$: instantaneous (solid line), and suitable integration ($\tau = 5$, dotted line). Second,with N consecutive silence frames, in dashed line $N = 20$ frames (400ms) and in dash-dot line $N = 200$ frames (4s)

ure the benefit of low-pass filtering of the parameter $\rho(t)$: by lessening the influence of short static lip-movements in speech and also short lip-movements in silence, it allows to significantly decrease the false silence detection ratio for a given silence detection ratio, compared to the case where no integration is performed (e.g., the point 20%-80% without integration becomes 5%-80% with a correct integration). Futhermore, Fig. 4 shows the effect of post-processing for unfiltered version of $\rho(t)$. The ROC curves show that the a too large duration ($N = 200$ frames corresponding to 4s) leads to dramatically decrease the silence detection ratio. On the contrary, using a reasonable ($N = 20$ frames corresponding to 400ms) allows to efficiently decrease the false silence detection ratio without decreasing the silence detection ratio (in comparison to the case with no post-processing, i.e. $N = 1$ frame). The gain due to post-processing is similar to the gain due to low-pass filtering.

Altogether, we obtain silence detection scores that are exploitable in real speech processing applications like enhancement, separation or recognition in noise. The compromise such as 5%-80% can guarantee a sufficently low false detection rate for a correct exploitation of the information.

5. CONCLUSION

Direct VAD from raw lip parameters cannot lead to satisfactory performances because of the intricate relationship between visual and acoustic speech information. However, we showed in this paper that considering a single appropriate dynamical parameter together with temporal integration at both the feature level (low-pass filtering of this parameter) and the decision level (testing N successive frames) can lead to efficient visual voice *non-activity* detection. Moreover, these performances are completely independent of the acoustic environment. Other part of this work concerns the use of the VVAD for speech source separation, as presented in another paper submitted to the present conference.

6. REFERENCES

[1] R. Le Bouquin-Jeannès and G. Faucon, "Study of a voice activity detector and its influence on a noise reduction system," *Speech Comm.*, vol. 16, pp. 245–254, 1995.

[2] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Lett.*, vol. 6, no. 1, pp. 1–3, January 1999.

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

[4] L. Girin, J.-L. Schwartz, and G. Feng, "Audio-visual enhancement of speech in noise," *JASA*, vol. 109, no. 6, pp. 3007–3020, June 2001.

[5] S. Deligne, G. Potamianos, and C. Neti, "Audio-visual speech enhancement with AVCDCN (AudioVisual Codebook Dependent Cepstral Normalization)," in *Proc. ICSLP*, 2002, pp. 1449–1452.

[6] D. Sodoyer, L. Girin, C. Jutten, and J.-L. Schwartz, "Developing an audio-visual speech source separation algorithm," *Speech Comm.*, vol. 44, no. 1–4, pp. 113–125, October 2004.

[7] B. Rivet, L. Girin, and C. Jutten, "Solving the indeterminations of blind source separation of convolutive speech mixtures," in *Proc. ICASSP*, Philadelphia, USA, March 2005.

[8] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior, "Recent advances in the automatic recognition of audio-visual speech," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1306–1326, 2003.

[9] P. Liu and Z. Wang, "Voice activity detection using visual information," in *Proc. ICASSP*, Montreal, 2004.

[10] T. Lallouache, "Un poste visage-parole. Acquisition et traitement des contours labiaux," in *Proc. Journées d'Etude sur la Parole (JEP) (French)*, Montréal, 1990.