J Multimodal User Interfaces (2015) 9:183–193
DOI 10.1007/s12193-015-0187-2

ORIGINAL PAPER

# Voice activity detection based on facial movement

**Bart Joosten**[1] · **Eric Postma**[1] · **Emiel Krahmer**[1]

Received: 3 October 2014 / Accepted: 19 June 2015 / Published online: 22 July 2015


**Abstract** We present and evaluate a new Visual Voice Activity Detection method based on Spatiotemporal Gabor filters (STem-VVAD). Since Spatiotemporal Gabor filters are dynamic, they offer an attractive method to separate speech from non-speech frames in video, even though they have not been used for this purpose before. We evaluate our method on two datasets, which differ in the ratio of speech to non-speech frames (high versus low), as well as in the head orientation of the speakers (frontal versus profile). We compare models on different regions (applied to the mouth, the head or the entire video frame), and do so both for speaker-dependent, individual models and speaker-independent, generic models. In general, best performances are obtained for speaker-dependent STem-VVAD applied to the mouth region, and combining information from different speeds. In all these cases, the system outperforms two reference systems, relying on frame differencing and static Gabor filters respectively, showing that Spatiotemporal Gabor filters indeed are beneficial for visual voice detection.

**Keywords** Visual voice activity detection · Facial movements · Spatiotemporal Gabor filters

✉ Bart Joosten
b.joosten@tilburguniversity.edu

Eric Postma
e.o.postma@tilburguniversity.edu

Emiel Krahmer
e.j.krahmer@tilburguniversity.edu

[1] Tilburg center for Cognition and Communication, Tilburg University, Tilburg, The Netherlands

## 1 Introduction

Human speech comprises two modalities: the auditory and the visual one. Many researchers have emphasized the close connection between the two (e.g., [26,38]). A speaker cannot produce auditory speech without also displaying visual cues such as lip, head or eyebrow movements, and these may provide additional information to various applications involving speech, ranging from speech recognition to speaker identification. For many of these applications it is important to be able to detect *when* a person is speaking. Voice Activity Detection (VAD) is usually defined as a technique that automatically detects human speech in an auditory signal. Using VAD enables speech processing techniques to focus on the speech parts in the signal, thereby reducing the required processing power. This is, for example, applied in digital speech transmission techniques (e.g., GSM or VoIP), where VAD helps to transmit speech and not silence segments [2,22].

*Auditory voice activity detection* Arguably, the straightforward approach to VAD would be to look into the auditory channel to see when speech starts. This is indeed what various researchers have done, and what is required for situations in which only the auditory signal is available [4,12,31,37]. However, this approach suffers from a number of complications. For instance, when background noise is present it becomes more difficult to differentiate between noise and speech, because they are entwined in one signal. Moreover, when multiple speakers are present, recognizing speech also becomes more difficult (because of overlapping speech). Even though solutions for these problems have been proposed (e.g, [11,19,32]), various researchers have argued that taking the visual signal into account (if available) can help in addressing these issues, e.g. because the presence or absence of lip movements can help in distinguishing

noise from speech [35], and because visual cues can help for speech segmentation. Moreover, importantly, visual cues such as mouth and head movements typically precede the actual onset of speech [40], allowing for an earlier detection of speech events, which in turn may be beneficial for the robustness of speech recognition systems. For this reason, various researchers have concentrated on Visual Voice Activity Detection (VVAD).

*Visual voice activity detection* Previously proposed VVAD methods mostly relied on lip tracking [1,24,36]. While these approaches have been successful, both in detecting voice activity based on visual cues and in combination with auditory VAD approaches, we know that there are more visual cues during speech in the face beyond the movement of the lips [21]. Besides, evidently (extracting features from) lip tracking is challenging when a speaker turns their head sideways. In their overview on *audiovisual automatic speech recognition*, Potamianos et al. [30] point out that robust visual features for speech recognition should be able to handle changing speaker, pose, camera and environment conditions, and they have identified three types of visual features that apply to VVAD as well: (1) appearance based features using pixel information extracted from a region of interest (typically the mouth region), (2) shape based features derived from tracking or extracting the lips, and (3) a combination of the aforementioned types of features. Potamianos et al. [30] note that extensive research comparing these features is still missing.

*Our approach* Since many VVAD studies acknowledge the importance of modeling movement during speech, we choose to explicitly examine movement information at an early stage, an approach called *Early Temporal Integration* [41], by designing a VVAD that incorporates features that represent spatiotemporal information. In this paper, we propose an appearance based approach to VVAD, representing images in terms of movement, without explicitly tracking the lips. Our novel method, which we call STem-VVAD (STem abbreviates Spatiotemporal, but also happens to mean "voice" in Dutch) is based on Spatiotemporal Gabor filters (SGF), a type of filter which is sensitive to movement at a certain direction and speed [29], which have, to the best of the authors knowledge, never been applied to VVAD. Intuitively, lip movements during speech have a specific spatiotemporal signature which may be different from those associated with non-speech (e.g., couching, laughing). In a similar vein, the orientation of movements may show different patterns for speech and non-speech, facilitating VAD.

Spatial two-dimensional (2D) Gabor filters have been frequently used for automatic visual tasks, ranging from texture segmentation [15] to coding of facial expressions (e.g., [23,25]) and automatic speech recognition [20]. The use of 2D Gabor filters in computer vision is inspired by biological findings on the neural responses of cells in the primary visual cortex (e.g., [6,9,16]), as the 2D Gabor function is able to model these responses. This makes them biologically plausible for use in automatic vision systems. Moreover, Lyons and Akamatsu [25] argue that the use of Gabor filters for facial expression recognition is also psychologically plausible, since the properties of the neurons that they are modeled on allow neurons in the higher visual cortex to be able to distinguish between different facial expressions.

Spatiotemporal Gabor filters are the dynamic variants of their spatial counterparts. Whereas spatial Gabor filters respond to visual contours or bars of a certain orientation and thickness, Spatiotemporal Gabor filters respond to moving visual contours or bars. The responses of motion-sensitive cells in primary visual cortex can be modeled by Spatiotemporal Gabor filters and have been shown to be the independent components of natural image sequences [13]. In this paper, we apply Spatiotemperal Gabor filters to Visual VAD, in our STem-VVAD approach.

*Data and evaluation procedure* To examine the extent to which our approach is successful in detecting voice activity, we have conducted a series of experiments on two different datasets, i.e., the CUAVE dataset [28], and our LIVER dataset [17]. The CUAVE dataset contains multiple speakers uttering digits, with frontal as well as profile recordings, whereas our LIVER dataset consists of frontally recorded speakers each with a single speech event, i.e., the uttering of the Dutch word for "liver". In the CUAVE set, the ratio between speech and non-speech is approximately balanced, this in contrast to the LIVER set where the majority of frames is non-speech.

For each dataset we assess the voice activity detection capabilities of our STem-VVAD method as well as for two reference VVADs: a VVAD based on frame differencing and a static, "standard" Gabor filter based method. In addition, we determine the contribution of various visual speeds to VVAD performance, to determine if certain speeds of, for instance, lip motion contribute more to VVAD than others. As a third evaluation, three regions in the clips are examined, to determine if zooming in on the mouth region leads to better VVAD performance, or that other dynamic facial characteristics contribute as well to the performance as suggested by [21].

Since human speech is inextricably connected to the idiosyncratic characteristics of its speaker [7] and, moreover, since the location with respect to the camera varies among the subjects, we will evaluate STem-VVAD on a speaker-dependent and a speaker-independent basis. By using these two evaluations we focus on the applicability of SGF in VVAD (speaker dependent) versus the generalizability of our method (speaker independent). In the area of speech recognition, systems tailored towards one specific speaker generally outperform systems that are able to handle multiple speakers. We therefore expect to see better results

with our speaker-dependent scheme than with our speaker-independent scheme. It will be interesting to see how this distinction affects our different VVADs.

## 2 Related work

Previous work on VVAD methods can be distinguished into two classes of models: lip-based approaches and appearance-based approaches. Below, we review examples of each of these classes.

### 2.1 Lip-based approaches

Lip-based approaches employ geometrical models based on the lips. The geometrical models typically consist of a flexible mesh formed by landmarks, or connected fiducial points surrounding the lips, flexible active contours that are automatically fitted to the lip region. In what follows, we describe three examples of lip-based approaches and the features extracted to perform VVAD.

Aubrey et al. [1] employed a geometrical lip model for VVAD that consisted of landmarks. Given a video sequence of a speaking and silent person, the task was to distinguish speech from non-speech. Their landmarks (constituting the lip model) were fitted to the video data of a speaking person by means of an Active Appearance Model (AAM) [5]. For each frame, the two standard geometric features, i.e., the width and height of the mouth, were extracted from the positions of the landmarks and submitted to a Hidden Markov Model.

Using an Active Contour Model [18], also called "snakes", Liu et al. [24] computed the two standard geometric features as well an appearance feature, i.e., the mean pixel values of a rectangular patch aligned with the lip corners and centered at the center of the mouth. For each frame, these three features form the basis of their classification vector, which is extended with dynamic features. To classify a frame as VOICE or SILENT, AdaBoost [10] was used, a technique that incrementally builds a (stronger) classifier by adding a new feature from the classification vector to the previous classifier at each consecutive step of the training process. The snake-based VVAD method was evaluated on a selected YouTube video of a single speaker.

The Sodoyer et al. [36] study relied on segmented lips, which were obtained by painting the lips of recorded speakers in order to be able to extract them from the rest of the face (like in the chroma key technique used in movies). In their study, they employed the chroma key technique to build a 40 min long audiovisual corpus of two speakers, each in a separate room, having a spontaneous conversation. In spontaneous conversation speech events are generally followed up by silence or non-speech audible events such as laughing and coughing. Such events are characterized by specific lip motion (even in silence parts). The aim of the study was to find a relationship between lip movements during speech and non-speech audible events on the one hand and silence on the other. The two standard geometrical features were extracted from the segmented lips of both speakers and used to define a single dynamic feature based on the sum of their absolute partial derivatives.

### 2.2 Appearance-based approaches

Appearance based VVAD approaches go beyond the lips by taking into consideration the surrounding visual information. We describe three examples of appearance-based method, each of which emphasizes another visual feature: color, texture, and optical flow.

Scott et al. [34] propose a VVAD that relies on a comparison of the pixel colors of the mouth region and the skin regions just below the eyes. They defined a *mouth openness* measure, which corresponds to the proportion of non-skin pixels in the mouth region. The regions were extracted with automatic face-detection and facial geometry heuristics. Their manually annotated VVAD dataset consisted of three videos.

Navarathna et al. [27] measured textural patterns in the mouth region using the Discrete Cosine Transform (DCT). Their dataset consisted of frontal and profile faces of the CUAVE dataset [28]. They classified the DCT coefficients by means of a Gaussian Mixture Model using speaker-independent models. This was realized by training and testing on different subsets of groups of speakers.

Tiawongsombat et al. [39] measured the optical flow in the mouth region using the Pyramidal Lucas-Kanade algorithm [3]. They recorded 21 image sequences of 7 speakers to evaluate and 7 individual mouth image sequences to train their method. Classification was done using a two-layered HMM that considers the states *moving* and *stationary* lips at the lower level and *speaking* and *non-speaking* at the higher level simultaneously.

### 2.3 Evaluation of existing approaches

Directly comparing results between the different studies is complex, since they all vary in certain dimensions, e.g., the datasets used differ in size and complexity, different evaluation metrics are employed, and generalizability is often not tested (i.e., evaluations tend to be speaker-dependent). With the exception of the CUAVE dataset, there are no publicly available datasets to enable a comparison across different situations and speakers. However, in general these methods all perform well in comparison to their specific task and in a comparable range. Typically, scores between 70 and 90 % are reported.

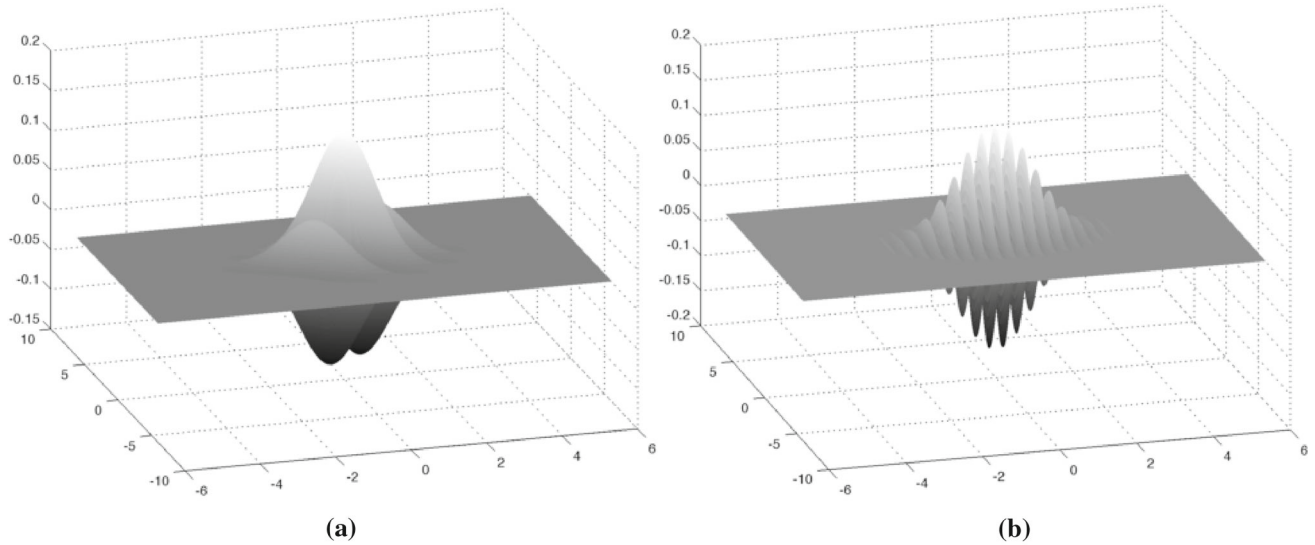


**Fig. 1** Visualisation of 2D Gabor filter. **a** Low spatial frequency, diagonal orientation. **b** High spatial frequency, vertical orientation

In the next sections, we present our own appearance based method (STem-VVAD), which is inspired by the biological example of early spatial-temporal integration in the brain. In addition, to get a better understanding of the problem, and in view of the complex, difficult to compare pattern of results in related work, here we systematically compare analyses of the mouth area with full facial analyses as well as analyses of the entire frame, and we look at different speeds of movement, both in isolation and combined into one feature vector. We evaluate the method on two different datasets (including CUAVE [28]), and look at both speaker-dependent and speaker-independent models.

## 3 The Spatiotemporal visual voice activity detection method (STem-VVAD)

The Spatiotemporal visual voice activity detection (STem-VVAD) method is based on two stages: (1) the preprocessing stage consisting of Spatiotemporal Gabor filters to determine the energy values at certain speeds, and (2) the aggregation and classification stage employing summation and a classifier to summarize and map the aggregated energy values onto the binary classes SPEECH and NON- SPEECH.

### 3.1 Preprocessing sage

The preprocessing stage transforms video sequences with Spatiotemporal Gabor Filters (SGFs) into a so-called energy representation [14,29,41]. The Spatiotemporal Gabor filters may be considered to be dynamic templates, i.e., oriented bars or gratings of a certain thickness that move with a certain speed and in a certain direction. The actual shapes of Gabor filters are illustrated in Fig. 1. The examples shown differ in two parameters: orientation and spatial frequency. The transformation of a video sequence by means of SGFs proceeds by means of convolution, in which each SGF (dynamic template) is compared with the contents of the video sequence at all pixel locations and at all frames. The presence of a moving elongated object in the video that matches the SGF in terms of orientation, thickness, speed and direction, results in a large "energy value" at the location and time of the elongated object. A better match results in a larger energy value. Each SGF results in one energy value for each pixel per frame of the video. Hence, the result of convolving a video sequence with a single filter, yields an energy representation that can be interpreted as an "energy video sequence" in which the pixel values represent energies. Large energy values indicate the presence of the filter's template at the spatial and temporal location of the value.

In order to capture all possible orientations, thicknesses (spatial frequencies), speeds and directions, a Spatiotemporal Gabor filter bank is used which consists of filters whose parameters (orientation, spatial frequency, speed and direction) are evenly distributed over the relevant part of the parameter space. Each of these filters generates an "energy movie" and hence convolving a video sequence with a filter bank gives rise to an enormous expansion of the data. Given a video of $F$ frames and $N$ pixels per frame, convolution with a filter bank of $G$ filters results in $G \times F \times N$ energy values. The number of filters, $G$, is determined by the range and number of parameter values selected. In the STem-VVAD method the direction of movement is always perpendicular to the orientation. Hence, the number of filters is defined as $G = k \times d \times s$, where $k$ is the number of spatial frequencies, $d$ the number of orientations and $s$ the number of speeds.

### 3.2 Aggregation and classification stage

The applied filter bank of $G$ filters (that vary in three dimensions of parameter space, i.e., spatial frequency, speed and orientation) result in $G$ energy videos obtained from the convolution in the preprocessing stage. Representing the energy value for Gabor feature $g$, frame $f$, and pixel $n$ by $E_g(f, n)$, the aggregated features $A_g(f)$ are computed by summating the energy values for feature $g$ for each frame, which results in, $A_g(f) = \sum_{n=1}^{N} E_g(f, n)$. The aggregation generates one $G$-dimensional vector $A(f)$ per frame, the elements of which signal the presence of a filter-like visual pattern in the video frame under consideration. Since the $G$ filters represent different combinations of spatial frequencies, speeds and orientations, the summated energy values signal the presence of moving contours with these frequencies, speeds and orientations.

## 4 Experimental evaluation of the STem-VVAD method

As stated in the introduction, the experimental evaluation of the STem-VVAD method consist of three parts. First, its performance is evaluated on two video datasets. Second, it is compared to two reference VVADs: (1) to determine the contribution of using a sophisticated spatiotemporal filtering method, the STem-VVAD method's performance is compared to the simplest method of change detection called frame differencing, and (2) to assess the contribution of dynamic information, a comparison is made with a version of the method in which the speed is set to zero, thereby effectively creating static Gabor filters. Third, the VVAD performances obtained for three spatial regions or visual regions of analysis are compared. These regions are: the entire frame, the face, and the mouth.

### 4.1 Datasets

As stated in the introduction, the two datasets used to evaluate the VVAD method are the publicly available CUAVE dataset[1] [28] and our own *LIVER* dataset.[2] [17] Both datasets were recorded for different purposes and have different characteristics.

#### *CUAVE*

The CUAVE dataset is an audio-visual speech corpus of more than 7000 utterances. It was created to facilitate multimodal speech recognition research and consists of video recorded speakers uttering digits. The dataset contains both individual speaker recordings and speaker-pair recordings. We used the individual speaker recordings only. The set contains 36 different speaker video recordings (19 male and 17 female) in MPEG-2, 5000 kbps, 44 KHz stereo, 720 × 480 pixels, at 29.97 fps. All speech parts are annotated at millisecond precision. The speakers vary in appearance, skin tones, accents, glasses, facial hair and therefore represent a diverse sample. Speakers were recorded under four conditions of which we used the following two: stationary frontal view and stationary profile view. In both cases speakers were successively pronouncing the digits. In these clips, the frontal face videos have an average length of 52 s (sd = 14 s) compared to 24 s (sd = 6 s) for the profile videos.

#### *LIVER*

Our LIVER dataset was constructed in the context of a surprise elicitation experiment [17]. This experiment yielded a dataset of 54 video sequences of 28 participants (7 male and 21 female) uttering the Dutch word for liver ("lever") in a neutral and in a surprised situation resulting in two recordings per person. The participants all sit in front of the camera but are allowed to move their heads and upper body freely. The videos are in WMV format, 7000 kbps, 48 KHz stereo, 29.97 fps, at 640 by 480 pixels and were automatically annotated for speech using a VAD based solely on the audio channel. By means of visual inspection we checked the correctness of annotations. The recordings are cropped at approximately 4 s (i.e. around 120 frames) and start when the participants are about to speak. Contrary to in the CUAVE database, where speakers produce speech about half of the time, speakers in the LIVER dataset produce just one word in a 4 s interval, resulting in a dataset that is unbalanced for speech and non-speech frames (1053–6524, respectively).

### 4.2 Implementation details

For the preprocessing stage of the STem-VVAD method, we used the SGF implementation of Petkov and Subramanian [29].[3] We created a filter bank of $G = 6 \times 8 \times 2$ filters sensitive to 6 different speeds ($v = \{0.5, 1, 1.5, 2, 2.5, 3\}$ pixels per frame), 8 orientations ($\theta = \{0, 0.25\pi, 0.50\pi, 0.75\pi, \ldots, 1.75\pi\}$ radians) covering the range of speeds and orientations in our datasets, and two spatial frequencies, defined by the parameter $\lambda_0^{-1}$, where $\lambda_0^{-1} = \{1/2, 1/4\}$. The dimensionality of the resulting STem-VVAD feature vector for frame $f$, $A(f)$, is equal to $G_{\text{STem-VVAD}} = 6 \times 8 \times 2 = 96$. A separate version with the same parameters, but with $v = 0$

[1] http://www.clemson.edu/ces/speech/cuave.html.

[2] The dataset was created by our colleague prof. Swerts, and is available upon request.

[3] http://www.cs.rug.nl/~imaging/spatiotemporal_Gabor_function/GaborApp.html.

was used for comparison. In this version, the dimensionality of feature vector $A(f)$ is equal to $G_{\text{zero-speed}} = 2 \times 8 = 16$. This is the same dimensionality as the STem-VVADs where we take only one speed into consideration. We implemented frame differencing by taking the absolute differences of the pixel intensities of two consecutive frames and computing their sum, average and standard deviation, yielding three values per frame.

The video sequences in the datasets were convolved with the SGFs. The resulting energy values were aggregated as specified in Sect. 3.2. For the three regions of analysis, i.e., frame, face, and mouth, the aggregation was performed over the entire frame, the rectangle enclosing the face, and the rectangle enclosing the lower half of the face, respectively. The lower half of the face was defined as the half of the bounding box enclosing the face region. The face region was detected automatically using the OpenCV implementation of the Viola-Jones face detector with Local Binary Pattern features. Since we used face detection in each frame instead of face tracking, we had to deal with false positives and frames in which the detector failed to find a face. By manually ascertaining that the face in the first frame of each video sequence was correctly detected by the face detector, we could automatically remove false positives in subsequent frames by stipulating that a bounding box' size and location should not differ more than a fixed number of pixels, 50 pixels in our setup, from the face detected in the previous frame. We used a simple heuristic to account for the missing detections by interpolating between the previous and upcoming detected face's bounding boxes. Visual inspection of the detected face regions throughout the video sequences confirmed that this procedure worked for almost all videos. Eight video sequences in total (i.e., two in the CUAVE frontal condition, one in the CUAVE profile condition, and five in the LIVER dataset) yielded too little face detections and were excluded from the experiments. This amounts to 5 % of the total data, which suggests that any biases introduced by face detection failures are minimal.

A support vector machine was used to classify each frame as SPEECH or NON- SPEECH using feature vectors of the aggregated values as input. Feature vectors were classified with a linear Support Vector Machine, for which we used the LIBLINEAR SVM library [8].

### 4.3 Evaluation procedure

The generalization performance is an estimate of how well the VVAD performs on unseen videos. To estimate the generalization performance we used two validation procedures: 10 fold cross validation for the speaker-dependent evaluation and Leaving One Speaker Out (LOSO) cross validation for the speaker-independent evaluation. The LOSO cross validation measures the performance on speakers not included in the training set. The resulting generalization performances obtained for (1) frame differencing, (2) the zero-speed version, (3) separate speed versions, and (4) the full-fledged STem-VVAD, are reported in terms of F1-scores. The F1-score, which originates from Information Retrieval, is the harmonic mean of precision and recall [33]. The use of F1-scores is motivated by the unequal distributions of our two datasets (i.e., the CUAVE dataset is approximately balanced, while the liver dataset contains more non-speech frames than speech frames). In contrast to accuracy, the F1-score is insensitive to the unbalance of the two classes. In our tables and figures in the next section we also report the F1-score of the chance classifier, i.e., the classifier that randomly picks between the classes SPEECH and NON- SPEECH. The final F1-score at chance level is the average F1-score between all folds for the specific evaluation procedure.

## 5 Results

Our results are divided over two sections, i.e., speaker-dependent results, and speaker-independent results. In each section we start by presenting the results of the frontal-view speakers in both the CUAVE and the LIVER dataset, followed by the results of the profile-view speakers, obtained only on the CUAVE dataset.

#### Speaker-dependent results

The upper part of Table 1 summarizes the overall results obtained on the frontal faces of the CUAVE dataset. Inspection of this table reveals that, as expected, the best results (for all three detector types, FD, zero-speed and STem-VVAD) are obtained for the mouth region. Looking closer at the results for the mouth region, we can see that, importantly, the STem-VVADs outperform the two reference methods (FD and zero-speed). Of the six nonzero speeds examined, the STem-VVAD with 0.5 pixels per frame performs best, with an F1-score of 0.7, which is almost 0.15 above the reference methods. Performance of the single-speed STem-VVADs decreases slightly with increasing speed. The best result is obtained for the full-fledged STem-VVAD in which all speeds are combined: an F1-score of 0.78. This result is comprised of a precision of 0.76 and a recall of 0.79.

Figure 2 visualizes the distributions over speakers of the results for the mouth region with box-whisker-plots as a function of VVAD. Each plot visualizes the distribution of the mean F1-scores per speaker. The horizontal line in the middle of each box represents the median of the data, while the top and bottom horizontal lines of the box represents the upper and lower quartile of the data, respectively. The upper whisker depicts the largest data value which is smaller than the upper quartile plus 1.5× inter-quartile-range (i.e.,

**Table 1** Average speaker-dependent F1-scores obtained on all three datasets

| Dataset | Region | References | | STem-VVADs | | | | | | All |
|---|---|---|---|---|---|---|---|---|---|---|
| | | FD | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 | 3 | |
| CUAVE frontal | Frame | 0.5 | 0.5 | 0.67 | 0.64 | 0.6 | 0.58 | 0.58 | 0.57 | **0.72** |
| | Head | 0.51 | 0.53 | 0.67 | 0.66 | 0.64 | 0.63 | 0.62 | 0.62 | **0.75** |
| | Mouth | 0.56 | 0.55 | 0.7 | 0.68 | 0.67 | 0.66 | 0.65 | 0.65 | **0.78** |
| LIVER | Frame | 0.34 | 0.55 | 0.51 | 0.43 | 0.41 | 0.42 | 0.44 | 0.44 | **0.7** |
| | Head | 0.4 | 0.56 | 0.63 | 0.56 | 0.51 | 0.54 | 0.55 | 0.53 | **0.8** |
| | Mouth | 0.4 | 0.57 | 0.68 | 0.58 | 0.57 | 0.6 | 0.62 | 0.6 | **0.86** |
| CUAVE profile | Frame | 0.48 | 0.53 | 0.63 | 0.61 | 0.58 | 0.56 | 0.55 | 0.54 | **0.71** |
| | Head | 0.52 | 0.59 | 0.66 | 0.66 | 0.64 | 0.62 | 0.61 | 0.61 | **0.78** |
| | Mouth | 0.54 | 0.63 | 0.7 | 0.69 | 0.68 | 0.65 | 0.65 | 0.64 | **0.8** |

The left part of the table shows the results for the frame differencing (FD) and the zero-speed (0) version VVADs and the right part of the table lists the F1-scores for the STem-VVAD method. The columns labeled 0.5–3 contain the scores of the associated speeds, the rightmost column labeled *All*, lists the result for the full-fledged STem-VVAD in which all speeds are included. The three rows for each dataset show the results for the three regions of analysis: frame, face, and mouth. The best scores are printed in bold-face. Chance level F1-scores for the three datasets are 0.47, 0.23 and 0.49 respectively. All scores are significantly different from chance level scores as determined by a two-sample Kolmogorov-Smirnov test at the 1 % significance level

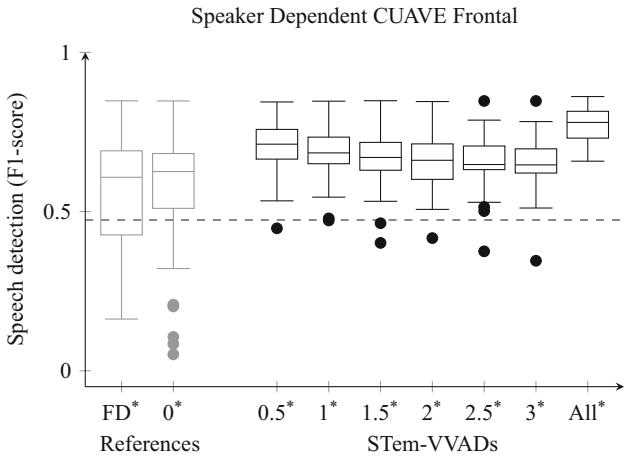


**Fig. 2** Boxplots of speaker-dependent F1-scores obtained on the CUAVE frontal dataset. The *boxes* correspond to the **Mouth** results in the upper part of Table 1. The *left* part of the figure shows the distribution for the frame differencing (FD) and the zero-speed (0) version VVADs and the *right* part of the Figure displays box plots of F1-scores for the STem-VVAD method. The *boxes* labeled *0.5–3* represent the F1-scores of the associated speeds, the rightmost *box* labeled *All*, shows the F1-scores for the full-fledged STem-VVAD in which all speeds are included. The *dashed line* indicates performance at chance level

absolute difference between upper and lower quartile). The reverse holds for the lower whisker, i.e., the smallest data value larger than the lower quartile minus 1.5× inter-quartile-range. Any data larger or smaller than the upper and lower whisker respectively is considered an outlier and is depicted by a dot. The spread of the STem-VVADs is considerably smaller than those of the reference methods, implicating a more robust detection performance for the STem-VVADs. The positions of the box plots' medians are in line with the mean values reported on the last line of the upper part of Table 1, showing a gradual descent for increasing speeds and a best performance when combining all speeds.

The results of our VVADs on the LIVER dataset evaluated with ten-fold CV are summarized in the middle part of Table 1. The overall pattern of results is similar to those obtained on the CUAVE dataset. The performances improve with smaller regions, with the best performance obtained for the mouth region. For the mouth region, the single-speed STem-VVADs outperform the reference methods (best single-speed performance is obtained for speed 0.5 (0.68). Again, the full-fledged STem-VVAD yields the best overall performance on all three regions of analysis (0.86 on the mouth region). When we zoom in on this result, we see that the recall here is higher, i.e., 0.93, than the precision, which is 0.8.

The corresponding box-whisker plots for the mouth region in Fig. 3 show a similar pattern of results as obtained for the CUAVE dataset. The most striking result is the superior performance obtained for the STem-VVAD.

The lower part of Table 1 shows the speaker-dependent results obtained on the subset of profile faces in the CUAVE dataset. A comparison with the results obtained for the frontal faces in the upper part of Table 1, reveals that the STem-VVAD method can deal with profile faces very well. The mouth-region results are displayed in Fig. 4.

### Speaker-independent results

The upper part of Table 2 gives the results for the CUAVE database with the Leave One Speaker Out validation method,

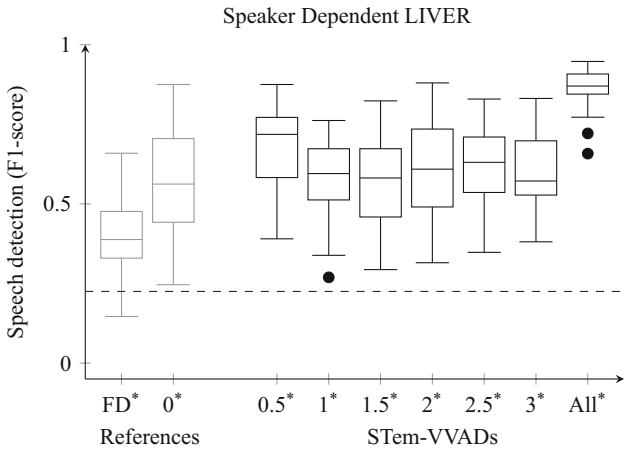


**Fig. 3** Boxplots of speaker-dependent F1-scores obtained on the LIVER dataset. The *boxes* correspond to the *Mouth* results in the middle part of Table 1. For explanation see Fig. 2

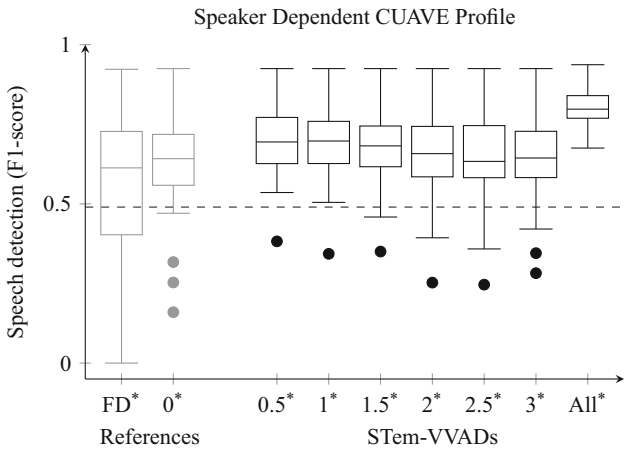


**Fig. 4** Boxplots of speaker-dependent F1-scores obtained on the CUAVE profile dataset. The *boxes* correspond to the *Mouth* results in the lower part of Table 1. For explanation see Fig. 2

which tests the generalizability of our VVAD methods across speakers. Inspection of this table reveals a similar pattern of results as in the upper part of Table 1, although with a lower overall performance. In particular, results for the mouth region are generally better overall than those for the head and the mouth region. Moreover, the best performing individual method is the STem-VVAD with speed 0.5 pixels per frame, although the difference with the FD reference VVAD is much less pronounced than in the ten-fold cross validation results in the upper part of Table 1. Interestingly to remark here is the performance of the FD reference method (0.53 %) for the entire frame compared to all the other detectors applied to the same region, since it is the best performing VVAD. Moreover, this VVAD also has a higher score than it's equivalent applied to the head region. In general the FD's performances here are only slightly below the best performing VVADs, i.e., the 0.5 pixels per frame and the combined speeds, whereas the zero-speed's performance here is considerably less.

Again, we zoomed in on the results for the mouth region and visualized them using a box-whisker-plot, as depicted in Fig. 5. Compared to Fig. 2 the boxes generated from the LOSO experiment are less compressed, corresponding to a wider spread of the individual results, it does however, show roughly the same pattern of performance as the previous plot when comparing them individually.

The middle part of Table 2 shows the speaker-independent results of our VVADs applied to the LIVER dataset, using a Leave One Speaker Out CV. The speaker-independent results are clearly inferior to the speaker-dependent results listed in the middle part of Table 1. Interestingly, simple frame differencing often outperforms single-speed STem-VVADs. The full-fledged STem-VVAD shows the best performance at all three regions of analysis with the best result (0.55) obtained for the mouth region. The box plots in Fig. 6 illustrate the corresponding results for the mouth region.

**Table 2** Speaker-independent F1-scores obtained on all three datasets

| Dataset | Region | References | | STem-VVADs | | | | | | All |
|---|---|---|---|---|---|---|---|---|---|---|
| | | FD | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 | 3 | |
| CUAVE frontal | Frame | **0.53** | 0.38 | 0.51 | *0.45* | *0.44* | *0.45* | *0.45* | *0.42* | 0.5 |
| | Head | 0.51 | 0.39 | 0.51 | 0.5 | 0.49 | 0.5 | 0.52 | 0.51 | **0.53** |
| | Mouth | 0.53 | 0.38 | 0.55 | 0.54 | 0.54 | 0.53 | 0.51 | 0.5 | **0.58** |
| LIVER | Frame | 0.38 | 0.34 | 0.36 | *0.29* | *0.27* | 0.3 | 0.31 | 0.29 | **0.46** |
| | Head | 0.37 | 0.31 | 0.43 | 0.38 | 0.3 | 0.3 | 0.32 | 0.34 | **0.44** |
| | Mouth | 0.4 | *0.22* | 0.4 | 0.38 | 0.35 | 0.35 | 0.33 | 0.32 | **0.55** |
| CUAVE profile | Frame | ***0.49*** | 0.42 | *0.41* | *0.42* | *0.41* | *0.42* | *0.4* | 0.39 | *0.42* |
| | Head | *0.5* | *0.49* | *0.49* | *0.51* | *0.51* | *0.51* | *0.5* | *0.49* | ***0.53*** |
| | Mouth | *0.51* | 0.53 | *0.52* | *0.55* | 0.56 | *0.55* | *0.54* | *0.54* | **0.56** |

For explanation, see table 1. Chance level F1-scores are 0.48, 0.24 and 0.49 respectively. Values in bold-face are the best scores. Italic values indicate F1-scores which or not significantly different from the chance level F1-scores as determined by a two-sample Kolmogorov-Smirnov test at the 1 % significance level

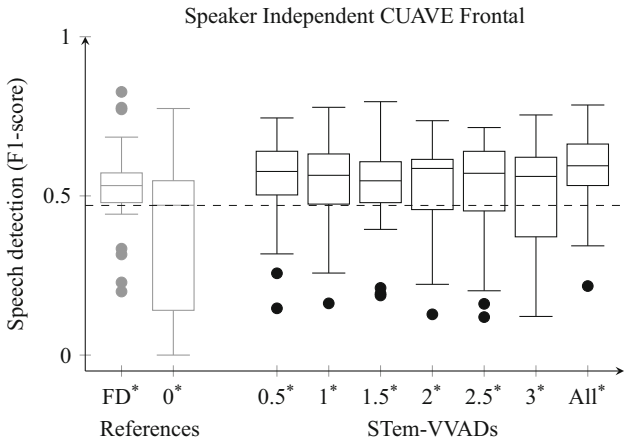


**Fig. 5** Boxplots of speaker-independent F1-scores obtained on the CUAVE frontal dataset. The *boxes* correspond to the *Mouth* results in the upper part of Table 2. For explanation see Fig. 2

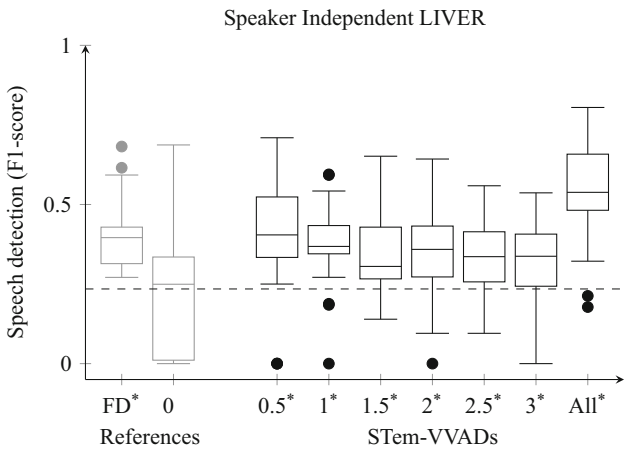


**Fig. 6** Boxplots of speaker-independent F1-scores obtained on the LIVER dataset. The *boxes* correspond to the *Mouth* results in the middle part of Table 2. For explanation see Fig. 2

The lower part of Table 2 lists our VVAD results obtained on the profile faces of CUAVE dataset. Compared to the lower part of Table 1 the full-fledged STem-VVADs here do not show a clear prevailing performance. Although the performance tends to improve when zooming in from frame to head to mouth, at each level the results for all VVADs are very similar. This little difference in results is visualized by Fig. 7 which contain the results for the mouth area.

## 6 General discussion and conclusion

In this paper, we studied whether it is possible to detect voice activity based on facial movements, which has various potential applications when auditory voice detection is difficult (e.g., when there is background noise or when there are multiple speakers). Obviously, movement is an essential ingredient

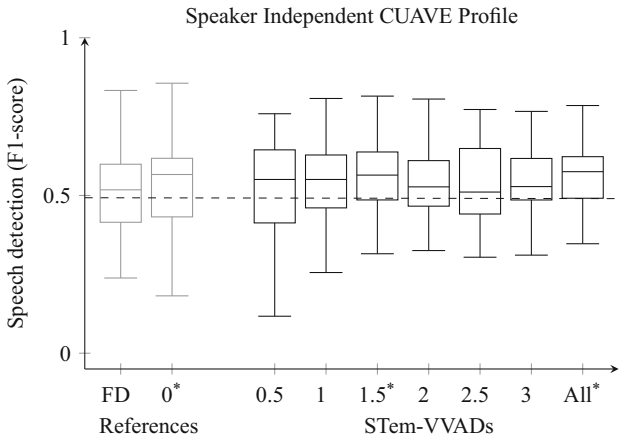


**Fig. 7** Boxplots of speaker-independent F1-scores obtained on the CUAVE profile dataset. The *boxes* correspond to the *Mouth* results in the lower part of Table 2. For explanation see Fig. 2

of visual voice activity detection (VVAD), and hence we studied whether Spatiotemporal Gabor filters could be used successfully for this task. Our set-up was as follows: we compared the performance of Spatiotemporal Gabor filters in our Stem-VVAD approach with two reference methods, namely a straightforward frame differencing method and a static Gabor filter method (i.e., zero-speed STem-VVAD), allowing us to capture the added value of both Spatial and Temporal information. We compared results on two different datasets (representing two extremes in the speech to silence ratio, which is low in the LIVER and high in the CUAVE dataset). We looked at both frontal and profile recorded faces, and compared performance at three levels of granularity (entire frame, entire face, mouth only). Finally, we evaluated the performance of the VVADs with both speaker-dependent models (where each speaker is used both for training and testing) and speaker-independent models (where we train and test on separate speakers).

The results present a clear picture. In almost all comparisons, the STem-VVAD (combining all speeds) yields the best performance, outperforming both the two baseline systems (and the chance performance level), sometimes by a wide margin.

Our STem-VVAD does not suffer from unbalanced training and test data. The results obtained from the LIVER dataset appear to be slightly better than those obtained on the CUAVE dataset for both individual and generic models. This suggests that the information extracted from this single-speech event data is informative enough to distinguish between speech and non-speech, even though the model is trained with an abundance of non-speech frames. As we pointed out above, the LIVER dataset was originally collected to study verbal and non-verbal expressions of surprise. It is interesting to point out that apparently the facial movements associated with speech differ from the ones associated

with surprise, since our STem-VVAD approach picks up on the former but not the latter.

Given the similar results obtained on the frontal and profile conditions of the CUAVE dataset we argue that our STem-VVAD is robust to turning faces (most notably in the speaker-dependent version). STem-VVAD does not rely on advanced lip models, which makes it potentially well suited for automatic speech detection in conference systems, where speakers tend to move their heads freely.

VVAD performance increases when focusing on the mouth; for all three techniques (FD, zero-speed, STem-VVAD), better results are usually obtained when taking only the head into account rather than considering the entire frame, and better results still when zooming in on just the mouth. Even though it has been argued that information from the upper part of the face (e.g., eyebrows) can be a useful cue for VVAD, this turned out not to help for the techniques we studied, perhaps because when considering a larger region of interest the chance of picking up speech irrelevant movements increase, and the movement cues that could be informative are more likely to be lost in the noise.

In addition, the speaker-dependent models (10-fold) perform (substantially) better than the generic models (LOSO), even though all three methods usually perform better than chance. This is perhaps not surprising because the speaker-dependent models capture some of the idiosyncratic properties of each speaker, which is not case for the generic models.

Perhaps more importantly for our current purposes, we find that adding temporal information, as we do in the Spatiotemporal Gabor filters, does pay off for VVAD. Zooming in on the mouth (where VVAD works best in our set-up) the best performing STem-VVAD, which combines different speeds, outperforms both reference VVADs, in both datasets, both frontal and profile, and in both individual models as well as for the generic LIVER models. Although the full-fledged mouth results for the generic CUAVE models are better than the reference methods, the differences are negligible.

Looking at the experimental data for the mouth region we can see that our STem-VVAD approach with all speeds could be a valuable addition to traditional auditory VAD systems, especially in the speaker-dependent case were a system is trained on an individual speaker basis. Achieving average F1-scores of 0.78, 0.86 and 0.8, respectively for the three datasets, a reasonable performance by itself. In the speaker-independent case the average F1-scores obtained for the mouth region of our full fledged STem-VVAD appear to be inaccurate enough for useful VVAD applications.

Various possibilities for future research exist. For example, we found that performance on the entire face was generally worse compared to performance on just the mouth region, even though others have claimed that other parts of the face may contain useful information as well. It would be interesting to see whether, say, trying to detect movement of just the eyebrows (possibly in combination with mouth movements) does lead to an improvement in VVAD. Another possibility would be to further integrate temporal information. Instead of looking at frames in isolation and trying to classify them, a window of frames or a dynamical model (e.g. a Hidden Markov Model) could improve results.

Our current method does not generalize very well, looking at the considerable differences between the speaker-dependent and the speaker-independent results. Apparently, idiosyncratic speech characteristics are prevailing over general speech patterns, considering the high F1-scores in the speaker dependent case. Another possibility could be the non-linearity of the feature space, to which we applied a linear SVM. In [41] the authors used Spatiotemporal Gabor filters to classify facial expressions. Although they report that using a non-linear SVM instead of a linear SVM yielded no significant performance increase, they state that their considerably large feature space (i.e., more than $2.2M$ per video sequence) generated by the non-linear Spatiotemporal Gabor filter responses might have made their problem linearly separable. In our case the dimensionality of the feature space was never greater than 96. Not being able to generalize very well is a disadvantage for practical application where you would want to use these techniques out-of-the box, for new speakers. It is conceivable that better results for the generic model can be obtained when more data from more different speakers become available. In addition, in future work we plan to experiment with techniques that have the potential to make our STem-VVAD method generalize better to unseen speakers. For instance by scaling the mouth's bounding box to a fixed size, or by taking the complete (normalized) SGF transformed mouth area (after dimensionality reduction) as input to a classifier.

In general, we can conclude that SGFs offer a promising account for visual voice activity detection. In particular, we have shown that adding temporal information to the widely used spatial Gabor filters yields substantially better results, than can be obtained with Frame Differencing or "standard" Gabor filters, since SGFs make better use of the inherent visual dynamics of speech production.



## References

1. Aubrey A, Rivet B, Hicks Y, Girin L, Chambers J, Jutten C (2007) Two novel visual voice activity detectors based on appearance

models and retinal filltering. In: Proceedings of the 15th European Signal Processing Conference, EUSIPCO-2007. pp 2409–2413
2. Beritelli F, Casale S, Cavallaero A (1998) A robust voice activity detector for wireless communications using soft computing. Sel Areas Commun IEEE J 16(9):1818–1829
3. Bouguet JY (2000) Pyramidal implementation of the lucas kanade feature tracker. Intel Corporation, Microprocessor Research Labs
4. Chang JH, Kim NS, Mitra SK (2006) Voice activity detection based on multiple statistical models. Signal Process IEEE Trans 54(6):1965–1976
5. Cootes TF, Edwards GJ, Taylor CJ (2001) Active appearance models. Pattern Anal Mach Intell IEEE Trans 23(6):681–685
6. Daugman JG (1985) Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. J Optical Soc Am A 2(7):1160–1169
7. Dellwo V, Leemann A, Kolly MJ (2012) Speaker idiosyncratic rhythmic features in the speech signal. In: Proceedings of Interspeech, Portland (USA)
8. Fan RE, Chang KW, Hsieh CJ, Wang XR, Lin CJ (2008) Liblinear: A library for large linear classification. J Mach Learn Res 9:1871–1874
9. Field DJ (1987) Relations between the statistics of natural images and the response properties of cortical cells. J Opt Soc Am A 4(12):2379–2394
10. Freund Y, Schapire RE (1995) A desicion-theoretic generalization of on-line learning and an application to boosting. In: Computational learning theory. Springer, pp 23–37
11. Furui S (1997) Recent Advances in Speaker Recognition. In: AVBPA '97: Proceedings of the First International Conference on Audio- and Video-Based Biometric Person Authentication. Springer
12. Ghosh PK, Tsiartas A, Narayanan S (2011) Robust voice activity detection using long-term signal variability. Audio Speech Lang Process IEEE Trans 19(3):600–613
13. van Hateren JH, Ruderman DL (1998) Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex. Proc Royal Soc London Ser B Biol Sci 265(1412):2315–2320
14. Heeger DJ (1987) Model for the extraction of image flow. J Optical Soc Am A 4(8):1455–1471
15. Jain AK, Farrokhnia F (1990) Unsupervised texture segmentation using gabor filters. In: Systems, Man and Cybernetics, 1990. Conference Proceedings., IEEE International Conference on, IEEE. pp 14–19
16. Jones JP, Palmer LA (1987) An evaluation of the two-dimensional gabor filter model of simple receptive fields in cat striate cortex. J Neurophysiol 58(6):1233–1258
17. Joosten B, Postma E, Krahmer E, Swerts M, Kim J (2012) Automated Measurement of Spontaneous Surprise. In: Grieco F, Krips OE, Loijens LWS, Noldus L, Zimmerman PH (eds) Spink AJ. Proceedings of Measuring Behavior, Utrecht. pp 385–389
18. Kass M, Witkin A, Terzopoulos D (1988) Snakes: Active contour models. Int J Comput Vision 1(4):321–331
19. Kinnunen T, Li H (2010) An overview of text-independent speaker recognition: From features to supervectors. Speech Commun 52(1):12–40
20. Kleinschmidt M, Gelbart D (2002) Improving word accuracy with gabor feature extraction. In: International Conference on Spoken Language Processing. Denver, CO. pp 25–28
21. Krahmer E, Swerts M (2005) Audiovisual prosody and feeling of knowing. J Memory Lang 53(1):81–94
22. Lee H, Kwon T, Cho DH (2005) An enhanced uplink scheduling algorithm based on voice activity for VoIP services in IEEE 802.16d/e system. IEEE Commun Lett 9(8):691–693
23. Littlewort G, Whitehill J, Wu T, Fasel I, Frank M, Movellan J, Bartlett M (2011) The computer expression recognition toolbox (cert). In: Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition. pp 298–305
24. Liu Q, Wang W, Jackson P (2011) A visual voice activity detection method with adaboosting. Sensor Signal Processing for Defence (SSPD 2011). pp 1–5
25. Lyons M, Akamatsu S, Kamachi M, Gyoba J (1998) Coding facial expressions with gabor wavelets. In: Proceedings of Third IEEE International Conference on Automatic Face and Gesture Recognition. pp 200–205
26. McGurk H, MacDonald J (1976) Hearing lips and seeing voices. Nature 264:746–748
27. Navarathna R, Dean D, Sridharan S, Fookes C, Lucey P (2011) Visual voice activity detection using frontal versus profile views. In: Digital Image Computing Techniques and Applications (DICTA), 2011 International Conference on, IEEE. pp 134–139
28. Patterson EK, Gurbuz S, Tufekci Z, Gowdy JN (2002) Moving-talker, speaker-independent feature study, and baseline results using the CUAVE multimodal speech corpus. EURASIP J Appl Signal Process 2002:1189–1201
29. Petkov N, Subramanian E (2007) Motion detection, noise reduction, texture suppression, and contour enhancement by spatiotemporal Gabor filters with surround inhibition. Biol Cybern 97(5–6):423–439
30. Potamianos G, Neti C, Luettin J, Matthews I (2012) Audio-visual automatic speech recognition. In: Bailly G, Perrier P, Vatikiotis-Bateson E (eds) Audiovisual Speech Processing. Cambridge University Press, Cambridge, pp 193–247
31. Ramírez J, Segura JC, Benítez C, de la Torre Á, Rubio A (2004) Efficient voice activity detection algorithms using long-term speech information. Speech Commun 42(3–4):271–287
32. Reynolds D (2002) An overview of automatic speaker recognition. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp 4072–4075
33. van Rijsbergen C (1979) Information Retrieval. Butterworth
34. Scott D, Jung C, Bins J, Said A, Kalker A (2009) Video based vad using adaptive color information. In: Proceedings of 11th IEEE Internation Symposium on Multimedia (ISM '09). pp 80–87
35. Sodoyer D, Rivet B, Girin L, Schwartz JL, Jutten C (2006) An analysis of visual speech information applied to voice activity detection. Proceedings of 2006 IEEE International Conference on Acoustics. Speech and Signal Processing, IEEE. pp 1184–1196
36. Sodoyer D, Rivet B, Girin L, Savariaux C, Schwartz JL, Jutten C (2009) A study of lip movements during spontaneous dialog and its application to voice activity detection. J Acoust Soc Am 125:1184
37. Sohn J, Kim NS, Sung W (1999) A statistical model-based voice activity detection. Signal Process Lett IEEE 6(1):1–3
38. Stekelenburg JJ, Vroomen J (2012) Electrophysiological evidence for a multisensory speech-specific mode of perception. Neuropsychologia 50(7):1425–1431
39. Tiawongsombat P, Jeong MH, Yun JS, You BJ, Oh SR (2012) Robust visual speakingness detection using bi-level HMM. Pattern Recogn 45(2):783–793
40. van Wassenhove V, Grant KW, Poeppel D (2005) Visual speech speeds up the neural processing of auditory speech. Proc Natl Acad Sci United States Am 102(4):1181–1186
41. Wu T, Bartlett M, Movellan JR (2010) Facial expression recognition using Gabor motion energy filters. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on. pp 42–47