

# An Unsupervised Visual-only Voice Activity Detection Approach Using Temporal Orofacial Features

Fei Tao, John H.L. Hansen, Carlos Busso

Multimodal Signal Processing (MSP) Laboratory - Center for Robust Speech Systems (CRSS)  
Department of Electrical Engineering, The University of Texas at Dallas, Richardson TX 75080, USA

fxt120230@utdallas.edu, john.hansen@utdallas.edu, busso@utdallas.edu

## Abstract

Detecting the presence or absence of speech is an important step toward building robust speech-based interfaces. While previous studies have made progress on *voice activity detection* (VAD), the performance of these systems significantly degrades when subjects employ challenging speech modes that deviate from normal acoustic patterns (e.g., whisper speech), or in noisy/adverse conditions. An appealing approach under these conditions is *visual voice activity detection* (VVAD), which detects speech using features characterizing the orofacial activity. This study proposes an unsupervised approach that relies only on visual features, and, therefore, is insensitive to vocal style or time-varying background noise. This study proposes an unsupervised approach that relies on visual features. We estimate optical flow variance and geometrical features around lips, extracting the short-time zero crossing rates, short-time variances, and delta features over a small temporal window. These variables are fused using *principal component analysis* (PCA) to obtain a “combo” feature, which displays a bimodal distributions (speech versus silence). A threshold is automatically determine using the expectation-maximization (EM) algorithm. The approach can be easily transformed into a supervised VVAD, if needed. We evaluate the system in neutral and whisper speech. While speech based VADs generally fail to detect speech activity in whisper speech, given its important acoustic differences, the proposed VVAD achieves near 80% accuracy in both neutral and whisper speech, highlighting the benefits of the system.

**Index Terms:** Visual voice activity detection, whisper speech

## 1. Introduction

Speech-based interfaces rely on *voice activity detection* (VAD) to effectively process only the signals containing speech. Robust VAD techniques exploit the statistical patterns observed in acoustic features which separate speech from other audio signals (harmonic, prosodic, and spectral properties) [1]. In many cases, environmental conditions or the nature of the applications introduce challenging conditions for existing VAD solutions. For example, when noise consists of background speech (e.g., babble noise) segmenting the target speech is very challenging. The performance also decreases when the speech mode deviates from neutral speech. An important case considered in this study is whisper speech which presents significant acoustic differences from neutral speech. For example, whisper speech lacks voiced excitation, and the formants are shifted [2–5]. Thus, the performance of speech-based VAD decreases. In these cases, an appealing alternative is using *visual voice activity detection* (VVAD) systems, which use facial features [6–10]. In theory, these facial features are insensitive to time-varying background noise or speech mode conditions.

VVAD systems offer suitable solutions to several interesting applications. For example, they can play an important role in silent speech interfaces [11] for speech-impaired individuals. They can also support *audiovisual automatic speech recognition* (AV-ASR) [12] systems. In the context of smart rooms [13], VVAD can facilitate the detection of the active speaker during group discussions. These approaches are also useful in the context of surveillance, security, and defense.

Previous studies have considered VVAD, proposing supervised [6–9, 14, 15] or unsupervised [10, 16] approaches. From an application perspective, unsupervised methods offer more flexible solutions across problems, since they can adapt to mismatches between test conditions. Inspired by the speech-based Combo-SAD system proposed by Sadjadi and Hansen [1], we present an unsupervised solution for VVAD which provides state-of-the-art performance. We extract optical flow variation and geometric features from the orofacial area. We capture the temporal facial fluctuations that characterize speech by estimating their short-time zero crossing rates, short-time variances, and delta features over a small temporal window. We fuse these features using *principal component analysis* (PCA), forming a “combo” feature, which displays a bimodal distribution, one for silence and the other for speech. The classes are automatically determined by using the *expectation-maximization* (EM) algorithm over the “combo” feature. We achieve competitive performance that does not decrease in the presence of whisper speech. Finally, we demonstrate that the system can be easily transformed into a supervised method, if needed.

## 2. Relation to Prior Work

Recent advances in AV-ASR have inspired the community to consider VVAD as an appealing approach to segment speech. Previous studies have considered supervised and unsupervised approaches. This section describes past relevant studies.

Most VVAD approaches rely on supervised frameworks where data is needed to train classifiers. Navarathna et al. [6] extracted DCT features around the mouth region. After concatenating the coefficients for seven consecutive frames, they projected the vector into a 40-dimensional space using *linear discriminant analysis* (LDA), forming a dynamic feature vector. Next, they used *Gaussian mixture models* (GMM) as the classifier. In addition to frontal views, they demonstrated that it is possible to detect speech activity using profile views of the subjects. Aubery et al. [8] used an *active appearance model* (AAM) and retinal filter to detect speech activity using *hidden Markov models* (HMMs), and later explored the use of optical flow [7]. They set thresholds on the likelihood of the models to define speech/silence regions. Joosten et al. [14] investigated the performance of VVAD at different speeds of facial movements. The system forms a feature vector by combining



Figure 1: The MSP-AVW corpus. The figure displays a subject and the setting used to record the database.

spatiotemporal Gabor filters and frame differencing methods. They relied on *support vector machine* (SVM) as their classifier. Takeuchi et al. [9] extracted the variance of optical flow as visual features, and proposed a multimodal VAD system that combines acoustic and visual features.

Supervised models for VVAD trained with specific data are less flexible than unsupervised models. However, very few studies have proposed unsupervised VVAD methods. Sodoyer et al. [10] proposed an unsupervised method to detect lip activity. They estimated the height and width of the mouth. Next, they normalized their values, and estimated their derivatives. Finally, they added the values, forming a single feature describing the dynamic of the lip motion. A threshold was used to define lip activity boundaries. That study demonstrated that dynamic features (i.e., derivative) are more effective than the actual values describing the lip configuration. Notice that their work does not distinguish between lip motion associated with speech or non-speech events (e.g. smack, smile or laugh).

We propose a flexible unsupervised VVAD system that combines different temporal features capturing orofacial fluctuations caused by speech articulation. We develop the approach as an alternative to speech-based VAD. We evaluate the approach with whisper speech, achieving competitive performance for speech detection under neutral and whisper speech.

### 3. MSP-AVW Corpus

The study relies on the *audiovisual whisper* (AVW) corpus [17], which consists of whisper and normal speech recordings from 40 subjects (20 male and 20 female). The database has three parts: isolated digits (1-9, “zero” and “oh” – 220 samples per speaker), read sentences (TIMIT – 120 sentences per speaker), and spontaneous speech (answering to prompted questions – 10 questions per speaker). The speakers used whisper and neutral speech modes to collect each of these recordings. The data was recorded in a 13ft × 13ft ASHA certified sound booth, illuminated by two professional LED light panels (see Fig. 1(b)). We collected the audio with a SHURE close-talk microphone (48KHz) placed such that it does not occlude the participants’ face (see Fig. 1(a)). We recorded a frontal and profile views of the participants using two high definition SONY cameras (1440×1080, 29.97fps). For most of the recordings, we include two chroma key green screens to facilitate video processing steps. We describe the corpus in details in Tran et al. [17].

After segmenting the recordings, we annotate the phonetic boundaries. We use forced alignment for the neutral portion of the data to estimate word and phoneme boundaries (SAILAlign [18]). At the present time, we do not have the transcriptions for

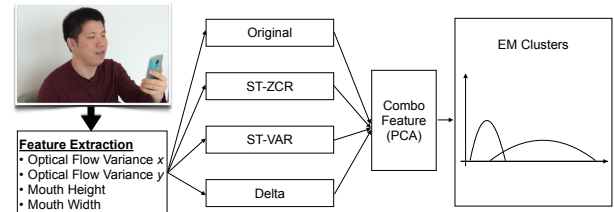


Figure 2: Flow chart of the proposed VVAD system.

the spontaneous speech, so this study relies on read sentences. Given the acoustic differences in whisper speech, we manually annotate the word boundary in the whisper speech. This study uses read sentences from 39 subjects, corresponding to approximately 10 hours of data (the facial extraction algorithm failed for one subject).

## 4. Proposed Visual Voice Activity Detection

The proposed approach is inspired by the speech-based *Combospeech activity detection* (SAD) system proposed by Sadjadi and Hansen [1], which combines five different features capturing various acoustic properties (periodicity, harmonicity, spectral flux, clarity, prediction gain). We build our approach by considering temporal orofacial features (see Fig. 2). We estimate a set of facial features capturing fluctuations caused by speech production. These features are combined using PCA, creating a “combo” feature. We use the EM algorithm to create speech and silence clusters over the “combo” feature. This section describes the approach.

### 4.1. Feature Set

Figure 3 shows the flowchart of the feature extraction process. We follow the methodology presented in our previous work [19]. We use the CSIRO face analysis SDK to extract 66 facial landmarks from the frontal videos. The toolkit uses a deformable model to fit landmarks from each frame to a template manually created for each subject [20]. To verify the correct detection of landmarks, we detect the face using the Viola-Jones detector algorithm [21]. We implement a generative model to independently obtain 9 facial landmarks [22, 23]. We compare the results from both approaches by estimating the coordinates of the lips corner, eyes corners and nose points. We also compare the mouth width and face size (Fig. 3). If we detect errors, we mark these frames. If a video has less than 10% missing features, we interpolate these values. Otherwise, we discard the entire video from the analysis.

After landmarks are detected, we define the *region of interest* (ROI) around the mouth. We estimate the width ( $W$ ) and height ( $H$ ) of the mouth. These geometrical features are derived from the lips landmarks. We also estimate the mouth area ( $area = WH$ ) and the overall geometrical distance ( $H + W$ ). Likewise, we estimate optical flow features to capture the lip

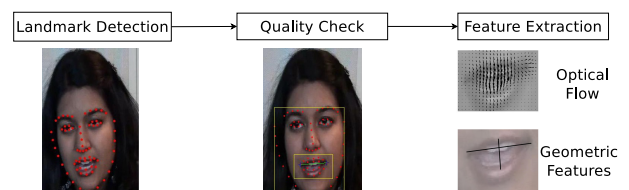


Figure 3: The feature extraction procedure.

Table 1: Temporal facial features. “X” denotes the statistic/functional estimated per feature.  $OF_x$  and  $OF_y$  are the horizontal and vertical optical flow variances.  $OF_{xy}$  is the overall optical flow.  $H$  and  $W$  are the height and width of the mouth, and  $H+W$  is the overall distance.  $area$  is the mouth area (*Original*: original feature set; *ST-ZCR*: short-time zero crossing rate; *ST-Var*: short-time variance; *Delta*: first order difference).

Feature Set							
Set	$OF_x$	$OF_y$	$OF_{xy}$	$H$	$W$	$H+W$	$area$
Original			X				
ST-ZCR	X	X	X	X	X	X	X
ST-Var	X	X	X	X	X	X	X
Delta				X	X	X	X
Final Feature (19 Dimensions)							
ST-ZCR (7D) + ST-Var (7D) + Delta (4D) + OFxy (1D)							

motion dynamics. We downsample the frame size by a factor of four to reduce overall computation. Once we estimate the optical flow for a given frame, we compute its variance across the horizontal ( $OF_x$ ) and vertical ( $OF_y$ ) directions over the mouth region. We obtain the overall optical flow variance ( $OF_{xy} = OF_x + OF_y$ ) representing the lip motion intensity. This 7D vector is our original feature set [ $OF_x$ ,  $OF_y$ ,  $H$ ,  $W$ ,  $OF_{xy}$ ,  $H+W$ ,  $area$ ].

Dynamic features provides better performance than actual values describing the lip configuration [10]. Therefore, we estimate temporal features from the 7D original feature vector: *short-time zero crossing rate* (ST-ZCR), *short-time variance* (ST-Var), and *absolute first order difference* (Delta). The first two temporal features (ST-ZCR and ST-Var) are derived from short windows of consecutive frames. There are two important constraints on the window size considered in this study: it should be short enough to have good resolution in the estimation of the silence-speech boundaries; and, it should be long enough to capture temporal speech fluctuations. We set the window length to nine video frames (300ms) to balance this tradeoff. Since we assign the resulting values of the functionals to the central frame of the window, the boundary resolution is only 150ms. Table 1 lists the temporal statistics derived from the 7D features. Next sections describe these temporal features.

#### 4.1.1. Short-Time Zero Crossing Rate (ST-ZCR)

We use ST-ZCR to capture the fluctuation in facial features generated by voice activation. ST-ZCR is obtained by calculating the *zero crossing rate* (ZCR) of each of the seven facial features during the short window (see Table 1). Since some of the variables are always positive (e.g., optical flow variances), we normalize the seven features by subtracting their mean, before estimated ZCR. This temporal functional captures the periodic fluctuations observed in facial features while speaking.

#### 4.1.2. Short-Time Variance (ST-Var)

We also estimate ST-Var over the facial features to capture temporal variations. We follow a similar methodology used for the ST-ZCR calculation, where the values are estimated over nine frames, and assigned to the central frame. Notice that optical flow features convey spatial variances. By estimating the ST-Var over the short window, we estimate the temporal variance. The ST-Var and ST-ZCR functionals provide complementary information. ST-ZCR functional captures the frequency associated with the signal fluctuation. The ST-Var functional captures the intensity associated with the signal fluctuation. We estimate ST-Var for all facial features as shown in Table 1.

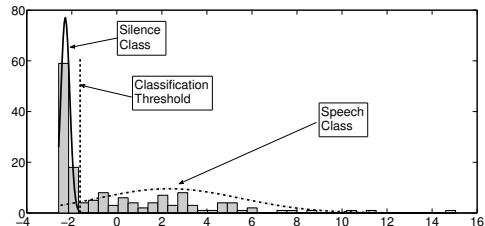


Figure 4: Distribution of the 1D “combo” feature for one video. It shows two Gaussian distributions where silence (first mode) is clearly different from speech (second mode). The vertical dash line gives the classification threshold estimated with EM.

#### 4.1.3. Absolute First Order Difference

A popular approach in speech processing to incorporate temporal information is the use of delta features. We calculate the absolute first order differences of the height, width, overall geometrical distance, and mouth area (Table 1). We expect that changes in these geometric features will provide more information about speech fluctuations than their actual values. Notice that we do not estimate delta features of the optical flow variable. These features describe the spatial variance in the mouth area. Delta features over optical flow will provide acceleration which has an unclear relationship with speech fluctuation. Furthermore, preliminary results demonstrate that including delta features for optical flow variables do not improve the overall performance of our VVAD.

## 4.2. Fusion of Temporal Features

As listed in Table 1, we use seven ST-ZCR features, seven ST-Var features and four delta features. From the original features, we do not consider geometric features. We only consider  $OF_{xy}$ , since it provides the overall lip motion intensity. These features define a 19D vector describing different aspects of speech fluctuation in the orofacial area, as discussed before.

We are interested in an unsupervised framework for VVAD. Following the approach proposed by Sadjadi and Hansen [1], we fuse the feature vector by estimating PCA, where the first principal component is used as a 1D “combo” feature. Since the 19D temporal features have different scales, we apply Z-normalization for each video using Equation 1, where  $\bar{f}$  and  $\sigma$  are the mean and standard deviation of the features.

$$f_{norm} = \frac{(f - \bar{f})}{\sigma} \quad (1)$$

The 1D combo feature has a bimodal distribution, where silence frames are clearly separated from speech frames. Figure 4 shows an example estimated from one of the videos. We assume that the 1D combo feature follows a *Gaussian mixture model* (GMM) with two univariate mixtures. We use the EM algorithm to estimate the means and variances of these distributions, which determine the threshold to automatically classify voice and silence frames. For each video, we run 50 times the EM algorithm on the “combo” feature to cluster the two classes. We select the result with the highest log-likelihood. Orofacial fluctuations caused by speech production increase the 1D value of the “combo” feature. Therefore, we assume that the cluster with higher mean represents speech frames, and the cluster with lower mean represents silence frames. We apply a median filter over the resulting decision values to remove spike noise given by the EM algorithm. Since we use a nine-frame window to calculate temporal statistics, we apply a median filter of order 19. Figure 5 shows the boundaries after this post-processing step.

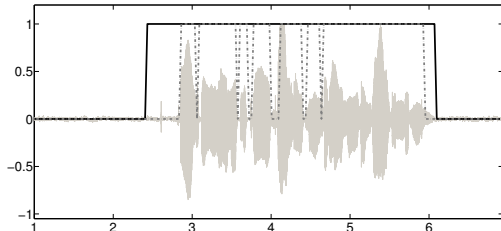


Figure 5: Silence/speech decisions boundary. The dashed line gives the ground truth, and the solid line gives the decision boundaries generated by the system.

Table 2: Performance of speech-based VAD [1]. (*NSen*: normal sentences, *WSen*: whisper sentences).

Set	Precision [%]	Recall [%]	F-score [%]	Accuracy [%]
Nsen	98.0	91.3	94.5	93.9
Wsen	72.3	78.7	75.3	74.8

## 5. Experimental Evaluations

We evaluate our VVAD on read sentences in whisper (*WSen*) and neutral (*NSen*) modes, reporting precision, recall, F-score, and accuracy (speech is relevant class for precision and recall).

For comparison, we use the state-of-the-art speech-based VAD system proposed by Sadjadi and Hansen [1] to evaluate the read sentences. Table 2 reports the results, which show that the F-score for the speech based VAD system is 94.5% in neutral sentences, but decreases to 75.3% for whisper speech. Lee et al. [24] proposed a supervised approach for speech-based VAD on neutral and whispering speech. Under matched scenarios (training and testing data are under same speech mode), they achieved around 93% accuracy. However, the performance drops under mismatched scenarios. When the supervised models are trained with neutral speech and tested with whispering speech, the accuracy is only 83% [24]. These results show the need for visual-based VAD.

Table 3 reports performance of the proposed VVAD approach. The F-score rates for whisper and neutral sentences is above 81%. For an unsupervised approach, this performance is very competitive. For comparison, Sodoyer et al. [10] reported 90% correct silence detection rate with 12% false detection rate over a small corpus (4.4 minutes). Joosten et al. [14] reported 65% precision rate and 70% recall rate for speech detection in isolated words sentences. Our unsupervised approach achieves over 80% accuracy on a challenging corpus (subjects in our corpus were allowed to move their body and head, some of them wore eye glasses, hat, and ear rings).

An interesting result is the similar performance for whisper and neutral speech. We have analyzed the differences in facial features produced under these two speech modes [17]. While the differences are not as evident as the patterns observed in the acoustic features, we noticed statistical significant differences in certain facial features (e.g., lip spreading). Speakers change

Table 3: Results of the proposed VVAD system (*NSen*: normal sentences, *WSen*: whisper sentences).

Set	Precision [%]	Recall [%]	F-score [%]	Accuracy [%]
Nsen	73.8	90.7	81.4	80.0
Wsen	73.5	90.3	81.1	79.4

Table 4: Results of the supervised VVAD over testing set (19 subjects) with linear kernel SVM. We report the results of the unsupervised VVAD for these partitions (*P*: precision, *R*: recall, *F*: F-Score, *A*: accuracy, *NSen*: normal sentences, *WSen*: whisper sentences)

Set	Supervised VVAD				Unsupervised VVAD			
	P [%]	R [%]	F [%]	A [%]	P [%]	R [%]	F [%]	A [%]
Nsen	84.3	89.1	86.6	86.9	73.1	90.5	80.8	79.1
Wsen	84.2	88.7	86.4	86.7	73.7	90.1	81.1	79.2

their articulatory production strategy while whispering. Since our approach did not use supervised methods, the performance is robust against characteristic patterns of whisper speech.

The differences in recall and precision rates suggest that setting the threshold using a supervised approach may improve the overall F-score rate. We evaluate this hypothesis by training a linear kernel *support vector machine* (SVM) with the 19D feature vector used to form the PCA-based “combo” feature (Sec. 4.2). We split the data into two speaker-independent, gender balanced partitions, where data from 20 subjects is used for training, and data from 19 subjects for testing. This classifier is trained with only the neutral sentences of the training set, using the speech/silence labels from forced alignment. Table 4 shows the results for neutral and whisper sentences on the testing set (19 subjects). The table also lists the results for the unsupervised VVAD for this set. The supervised version of the approach achieves an F-score of 86%, increasing 5% (absolute) over the unsupervised version of the approach. The recall rates increase approximately 10% as a result of setting the class boundaries with SVM, supporting our hypothesis. While the supervised version of the approach provides better overall performance, it is important to note that (1) it requires labeled data, (2) when the mismatch between test condition increases, the hyperplane defined by the SVM may not be optimum, depending on the training speakers. The supervised version of the proposed approach do not have these limitations.

## 6. Conclusions and Discussion

This study has proposed a new unsupervised visual-only VAD approach that can be easily adapted into a supervised VVAD, if needed. We used temporal and dynamic features to capture orofacial fluctuations caused by speech production. The approach utilizes PCA to fuse multiple features, and the EM algorithm to implement unsupervised classification of speech and silence. In spite of potential articulatory differences between neutral and whisper speech, the approach achieves similar performances under both conditions. The proposed unsupervised system is robust to both speech modes.

We expect that most errors are close to the speech-silence boundaries, given the asynchronies between speech and lip articulation (the boundaries are annotated based on speech). Also the use of temporal windows, which provides important information for this task, reduces the boundary resolution to 150ms. If the ultimate goal is to correctly determine speech segments, a simple solution would be to slightly extend the speech boundaries provided by the VVAD. An important advantage of the proposed approach is the flexibility to incorporate other facial cues or information from other modalities (e.g., speech). For future work, we will explore audiovisual solutions for VAD.

## 7. Acknowledgements

This work was funded by NSF (IIS-1217104) and Samsung Research America, Dallas Technology Laboratories.

## 8. References

- [1] S. Sadjadi and J. H. L. Hansen, "Unsupervised speech activity detection using voicing measures and perceptual spectral flux," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 197–200, March 2013.
- [2] X. Fan and J. Hansen, "Acoustic analysis for speaker identification of whispered speech," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2010)*, Dallas, TX, USA, March 2010, pp. 5046–5049.
- [3] X. Fan and J. H. L. Hansen, "Speaker identification within whispered speech audio streams," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1408–1421, July 2011.
- [4] —, "Acoustic analysis and feature transformation from neutral to whisper for speaker identification within whispered speech audio streams," *Speech Communication*, vol. 55, pp. 119–134, January 2013.
- [5] C. Zhang and J. H. L. Hansen, "Whisper-island detection based on unsupervised segmentation with entropy-based speech feature processing," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 883–894, May 2011.
- [6] R. Navarathna, D. Dean, S. Sridharan, C. Fookes, and P. Lucey, "Visual voice activity detection using frontal versus profile views," in *International Conference on Digital Image Computing Techniques and Applications (DICTA 2011)*, Noosa, Queensland, Australia, December 2011, pp. 134–139.
- [7] A. Aubrey, Y. Hicks, and J. Chambers, "Visual voice activity detection with optical flow," *IET Image Processing*, vol. 4, no. 6, pp. 463–472, December 2009.
- [8] A. Aubrey, B. Rivet, Y. Hicks, L. Girin, J. Chambers, and C. Jutten, "Two novel visual voice activity detectors based on appearance models and retinal filtering," in *European Signal Processing Conference (EUSIPCO 2007)*, Poznań, Poland, September 2007.
- [9] S. Takeuchi, T. Hashiba, S. Tamura, and S. Hayamizu, "Voice activity detection based on fusion of audio and visual information," in *International Conference on Audio-Visual Speech Processing (AVSP 2009)*, Norwich, United Kingdom, September 2009, pp. 151–154.
- [10] D. Sodoyer, B. Rivet, L. Girin, J.-L. Schwartz, and C. Jutten, "An analysis of visual speech information applied to voice activity detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2006)*, vol. 1, Toulouse, France, May 2006, pp. 601–604.
- [11] B. Denby, T. Schultz, K. Honda, T. Hueber, J. Gilbert, and J. Brumberg, "Silent speech interfaces," *Speech Communication*, vol. 52, no. 4, pp. 270–287, April 2010.
- [12] C. Neti, G. Potamianos, J. Luetttin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari, and J. Zhou, "Audio-visual speech recognition," Workshop 2000 Final Report, Technical Report 764, October 2000.
- [13] C. Busso, S. Hernanz, C. Chu, S. Kwon, S. Lee, P. Georgiou, I. Cohen, and S. Narayanan, "Smart Room: Participant and speaker localization and identification," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005)*, vol. 2, Philadelphia, PA, USA, March 2005, pp. 1117–1120.
- [14] B. Joosten, E. Postma, and E. Krahmer, "Visual voice activity detection at different speeds," in *International Conference on Auditory-Visual Speech Processing (AVSP 2013)*, Annecy, France, August–September 2013.
- [15] I. Almajai and B. Milner, "Using audio-visual features for robust voice activity detection in clean and noisy speech," in *European Signal Processing Conference (EUSIPCO 2008)*, Switzerland, Lausanne, August 2008.
- [16] R. Ahmad, S. Raza, and H. Malik, "Unsupervised multimodal VAD using sequential hierarchy," in *IEEE Symposium on Computational Intelligence and Data Mining (CIDM 2013)*, Singapore, April 2013, pp. 174–177.
- [17] T. Tran, S. Mariooryad, and C. Busso, "Audiovisual corpus to analyze whisper speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013)*, Vancouver, BC, Canada, May 2013, pp. 8101–8105.
- [18] A. Katsamanis, M. P. Black, P. Georgiou, L. Goldstein, and S. Narayanan, "SailAlign: Robust long speech-text alignment," in *Workshop on New Tools and Methods for Very-Large Scale Phonetics Research*, Philadelphia, PA, USA, January 2011.
- [19] F. Tao and C. Busso, "Lipreading approach for isolated digits recognition under whisper and neutral speech," in *Interspeech 2014*, Singapore, September 2014, pp. 1154–1158.
- [20] J. Saragih, S. Lucey, and J. Cohn, "Deformable model fitting by regularized landmark mean-shift," *International Journal of Computer Vision*, vol. 91, no. 2, pp. 200–215, January 2011.
- [21] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001)*, vol. 1, Kauai, HI, USA, December 2001, pp. 511–518.
- [22] M. Everingham, J. Sivic, and A. Zisserman, "'Hello! my name is ... buff' – automatic naming of characters in TV video," in *Proceedings of the British Machine Vision Conference (BMVC 2006)*, Edinburgh, Scotland, September 2006, pp. 899–908.
- [23] J. Sivic, M. Everingham, and A. Zisserman, "'Who are you?' – learning person specific classifiers from video," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, Miami Beach, FL, USA, June 2009, pp. 1145–1152.
- [24] P. Lee, D. Wee, H. Toh, B. Lim, N. Chen, and B. Ma, "A whispered mandarin corpus for speech technology applications," in *Interspeech 2014*, Singapore, September 2014, pp. 1598–1602.