INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Improving Boundary Estimation in Audiovisual Speech Activity Detection Using Bayesian Information Criterion

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

Multimodal Signal Processing (MSP) Laboratory - 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

A key preprocessing step in multimodal interfaces is to detect when a user is speaking to the system. While push-to-talk approaches are effective, its use limits the flexibility of the system. Solutions based on *speech activity detection* (SAD) offer more intuitive and user-friendly alternatives. A limitation in current SAD solutions is the drop in performance observed in noisy environments or when the speech mode differs from neutral speech (e.g., whisper speech). Emerging audiovisual solutions provide a principled framework to improve detection of speech boundaries by incorporating lip activity detection. In our previous work, we proposed an unsupervised *visual speech activity detection* (V-SAD) system that combines temporal and dynamic facial features. The key limitation of the system was the precise detection of boundaries between speech and non-speech regions due to anticipatory facial movements and low video resolution (29.97fps). This study builds upon this system by (a) combining speech and facial features creating an unsupervised *audiovisual speech activity detection* (AV-SAD) system, (b) refining the decision boundary with the *Bayesian information criterion* (BIC) algorithm, resulting in improved speech boundary detection. The evaluation considers the challenging case of whisper speech, where the proposed AV-SAD achieves a 10% absolute improvement over a state-of-the-art audio SAD.

Index Terms:Audiovisual SAD,Bayesian information criterion

1. Introduction

A key preprocessing step in speech-based interfaces is the detection of speech segments. Failing to detect segments with relevant information will result in drop of performance in subsequent tasks, such as *automatic speech recognition* (ASR). Since the use of push-to-talk approaches affects the natural interaction with the system, solutions based on *speech activity detection* (SAD) are usually preferred. The speech community has made important advances in *audio-only speech activity detection* (A-SAD) over the past decade. However, current solutions for SAD drop their performance in noisy environments or when the speech mode is not neutral (e.g., whisper speech). It is important to design SAD approaches that can maintain their performance even when dealing with challenging practical applications. An interesting alternative to distortion compensation is to consider visual information capturing lip activity.

In our previous work, we presented an unsupervised visual-only *speech activity detection* (V-SAD) using temporal orofacial features [1]. The system demonstrated robust performance for whisper speech. When we analyzed the results, we noted errors close to speech and non-speech boundaries due to anticipatory facial movements and low video resolution (29.97 fps). This paper addresses this problem by adding two key contributions. First, we combine A-SAD and V-SAD systems to derive a novel unsupervised leveraged AV-SAD. The A-SAD approach corresponds to the "Combo-SAD" system proposed by Sadjadi and Hansen [2], which combines five acoustic metrics using *principal component analysis* (PCA). The first *principal component* (PC), referred to as combo SAD, is used to create two Gaussian distributions using the *expectation maximum* (EM) algorithm, representing speech and silence, respectively. The V-SAD uses a similar framework, combining 25 facial features describing dynamic of the orofacial area. We fuse the outputs of the A-SAD and V-SAD systems, achieving improved performance. Second, we refine the detection around the speech and silence boundaries using the *Bayesian information criterion* (BIC) algorithm. This algorithm detects the point in the signal where changes are observed, producing more accurate boundary between speech and non-speech activity. We evaluate different combinations where BIC is applied to audio-only, visual-only or audiovisual features. We evaluated the performance of the system on neutral and whisper speech. While the ultimate goal is to have a real-time system, the approach is designed and evaluated offline.

This work was funded by NSF (IIS-1217104, IIS-1453781).

2. Background

2.1. Audiovisual Speech Activity Detection

While the area of A-SAD has been active, there are few studies on V-SAD. We describe related work on V-SAD and AV-SAD.

Liu and Wang [3] proposed a supervised V-SAD, where they extracted visual features from the mouth area. They capture dynamic information by augmenting the features with their first order derivative. The feature vector was reduced using PCA, training *Gaussian mixture models* (GMM) for speech and non-speech segments. Petsatodis et al. [4] used the vertical distance of the mouth opening and its derivative. Almajai and Milner [5] used appearance-based features for V-SAD. They extracted 2D *discrete cosine transform* (DCT) coefficients, concatenating the delta information to represent dynamic information. Aubrey et al. [6] investigated other static features including *active appearance model* (AAM). They implemented their system with *hidden Markov model* (HMM). Takeuchi et al. [7] extracted the variance of the optical flow as the visual features. These features, which our study also uses, can represent the orofacial area dynamic due to speech activity. Joosten et al. [8] proposed *support vector machine* (SVM) trained with *spatiotemporal Gabor filters* (SGFs).

Studies have also proposed audiovisual fusion for SAD. Takeuchi et al. [7] combined the V-SAD and A-SAD decision boundaries using logical "AND" and "OR" operators. Almajai and Milner [5] simply concatenated the acoustic and visual features. Petsatodis et al. [4] also considered AV-SAD using a rule-based approach. If the face is detected, the A-VAD was only activated when lip activity was detected. Otherwise, the decision relied only on A-VAD. Our study combines multiple facial features in a principled manner, capturing dynamic pat-

http://dx.doi.org/10.21437/Interspeech.2016-406

terns associated with speech activity in the orofacial area. It combines combo SAD and BIC algorithms, obtaining accurate boundaries between speech and non-speech regions.

2.2. Whisper Speech

While most studies have evaluated the benefits of audiovisual solutions in the presence of noisy speech [5, 9, 10], we consider the challenging problem of detecting speech activity in neutral and whisper speech. Whisper speech is a common speech mode used to communicate confidential information, speak in quiet places, and cope with temporary or permanent speech disorders (e.g., amygdalitis, cold and heavily smoker conditions). Whisper speech is a production mode characterized by lack of periodic excitation, affecting temporal and spectral properties of speech [11]. These differences significantly affect the performance of speech based interfaces [12, 13].

Recent studies have demonstrated the benefits of using audiovisual solutions for this problem [14, 15]. Tran et al. [16] studied the differences in acoustic and visual features between neutral and whisper speech. The study revealed that visual features are more invariant against changes produced by whisper speech, suggesting that they are good features to consider. To the best of the authors knowledge, this is the first study on AV-SAD in whisper speech.

2.3. Supervised versus Unsupervised SAD

Studies on V-SAD have considered supervised [5–8, 17, 18] and unsupervised [19, 20] methods. Even though supervised learning works well when SAD is trained and tested with similar speech conditions, potential mismatches may affect the performance of the system. For examples, classifiers trained with neutral speech may not work for whisper speech. In contrast, unsupervised learning does not rely on predefined thresholds, which are automatically learned from the distribution of the data. Therefore, we expect that they generalize better to new conditions, as long as speech and non-speech regions present differences in the feature space. We believe that this is an important property for SAD system, so our approach relies on an unsupervised learning approach.

3. Data and Feature Extraction

3.1. MSP-AVW corpus

This study uses the *audiovisual whisper* (MSP-AVW) corpus [16]. The MSP-AVW corpus was recorded from 40 American native speakers, including 20 females and 20 males. The audio was collected with a close-talking microphone at a sampling rate of 48 kHz, and the video was collected with two SONY *high definition* (HD) cameras set with a 1440×1080 resolution at 29.97 frames per second. Subjects read slides or talked about topics prompted in a monitor. The corpus includes three datasets: isolated digits, read sentences and spontaneous speech. This study only uses the set with read sentences, because the duration of isolated digits is very short (less than 0.5s), and spontaneous speech has not been processed (transcriptions, manual segmentation of speech regions). The recording of read sentences considered 129 TIMIT sentences. From these sentences, 30 sentences were read by all speakers in both whisper and neutral modes. In addition, each subject read 60 additional randomly selected sentences, where 30 were read in neutral mode, and 30 were read in whisper mode. In total, each subject read 120 sentences. The details of the corpus are given in Tran et al. [16].

3.2. Visual Features

For V-SAD, we aim to extract facial features from the orofacial area conveying the temporal patterns caused by speech articulation. After detecting the *region of interest* (ROI) around the lips, we estimate geometrical and optical flow features for V-SAD. The mouth width and height are extracted from facial landmark locations forming our geometrical features. These values are normalized per sentence using z-normalization. We further multiply ($\approx$ mouth area) and add ($\approx$ half mouth perimeter) the normalized width and height of the mouth, creating a 4D feature vector (width, height, width $\times$ height, width + height). We estimate the optical flow from the ROI extracting its variances across vertical and horizontal axes. These values are also normalized per sentence. Furthermore, we sum the normalized horizontal (OP_h) and vertical (OP_v) optical flow variance, creating a 3D feature vector (OP_h, OP_v, OP_h + OP_v). We use linear interpolation for these geometric and optical flow features for the frames where we are not able to extract the ROI. This approach forms a 7-D feature frame. We describe the details of this process in Tao et al. [1].

The key task in this study is detecting speech activity, so temporal information is important. Starting from the 7D feature vector, we compute several statistics over temporal windows. To balance the tradeoff between time resolution (i.e., short windows) and robust estimation of the statistics (i.e., long windows), we set the window size to 9 frames (about 0.3s), estimating the following statistics:

Temporal variance: we compute the variance for each dimension of the 7-D feature vector. Speech activity produces movement leading to changes in amplitude of the selected features. While the appearance of frames for non-speech activities (e.g., smile) may be similar to the ones for speech, their temporal variance will be smaller.

Zero crossing rate (ZCR): we compute ZCR for each dimension of the 7-D feature vector. If the lips open and close quickly, the ZCR of visual features will be higher. We expect higher ZCR values during speech activity than during non-speech activity. Equation 1 computes the ZCR of a signal s_t, where T is the window length, and $\mathbb{1}$ is an indicator equals to one when s_t and s_{t-1} have different signs.

$$ZCR = \frac{1}{T-1}\sum_{t=1}^{T-1}\mathbb{1}\left\{s_t, s_{t-1}\right\} \tag{1}$$

Speech periodic characteristic (SPC): we compute SPC, defined by Equation 3, for each dimension of the 7-D feature vector. $R(\cdot)$ is the auto-correlation, t is the time index, and T is the window length. $Y(t)$ gives the equation of the line passing through $R(0)$ and $R(T-1)$. Since $R(0)$ is the peak of the auto-correlation, this line has negative slope. With these definitions, $Y(t) - R(t)$ measures the distance from the line to the autocorrelation value. This value is smaller for periodic signals. Speech activity produces periodic movements in the orofacial area. The auto-correlation will show several peaks during speech activity. In contrast, we expect to observe only the first peak during non-speech activity. Consequently, the SPC value is expected to be lower during speech activity.

$$Y(t) = \frac{R(T-1) - R(0)}{T-1}t + R(0) \tag{2}$$

$$SPC = \sum_{t=0}^{T-1}(Y(t) - R(t)) \tag{3}$$

First order derivative: we compute this metric to only geometric features. We did not apply this statistic to optical flow variance (i.e., acceleration) since this information may not be informative of speech.

We concatenate these statistics forming a 25-D visual feature. Since the overall optimal flow variance provides useful

dynamic information about orofacial area, we added this feature creating a 26-D feature vector.

3.3. Acoustic Feature

We implement the unsupervised state-of-the-art A-SAD proposed by Sadjadi and Hansen [2] (Sec. 4.1). The system considers five acoustic features capturing harmonicity, clarity, prediction gain, periodicity and perceptual spectral flux (see details in [2]). The acoustic features achieve high SAD accuracy for neutral speech. However, performance significantly drops for whisper speech, due to the underlying acoustics differences.

4. Approach

Figure 1 shows the flowchart of the proposed AV-SAD approach. First, we estimate decision boundaries for A-SAD and V-SAD using the combo SAD approach (Sec. 4.2). The decision boundaries are fused using "AND" operator (Sec. 4.3). We apply the BIC algorithm to refine SAD boundaries (Sec. 4.3).

4.1. Combo-SAD

Sadjadi and Hansen [2] proposed the combo SAD framework consisting of two steps. The first step combines N features describing speech activity into a single metric using PCA. First, the N features are individually normalized using z-normalization, preventing one feature with higher values to dominate other features. We estimate the PCA of the N-D feature vector, keeping only the first principal component, which we refer to as "combo" feature. The second step derives speech and non-speech regions using the *expectation maximization* (EM) algorithm. The method uses the EM algorithm to fit two Gaussian distributions in the "combo" feature. The mode with higher mean represents speech, and the mode with the lower mean represents non-speech. The threshold between classes is automatically learned by the EM algorithm providing an appealing unsupervised approach for SAD.

We use combo SAD framework to derive a V-SAD system. Starting from the 26-D feature vector for facial features (Sec. 3.2), we implement the combo SAD framework to derive a 1-D "combo" feature. With the exception of SPC features, speech activity will tend to create higher values for the 26 facial features (e.g., the mean of OP_h, OP_v should be higher during speech). Therefore, we expect that the mode with higher values represent the speech class. We use this framework to derive an A-SAD system using the 5-D speech features (Sec. 3.3). We employ a median filter to suppress spikes in the signal where we use a 5-point median filter for the A-SAD system, and a 7-point median filter for the V-SAD system.

4.2. Logical Fusion

The decision boundaries for A-SAD and V-SAD are combined using with the logical "AND" operation. We choose the "AND" operator, because it requires both modalities to agree, creating a decision fusion stricter than with operators such as "OR??, reducing miss detections. Because the frame rate of the visual features is lower than the one for audio features, we up-sample the visual features to 100 fps before estimating the BIC algorithm.

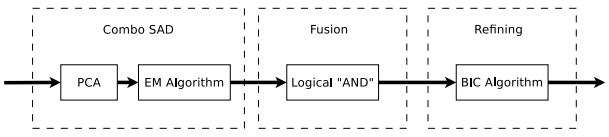

Figure 1: Proposed approach. Decision boundaries for A-SAD and V-SAD are estimated with the combo SAD approach. After fusing the decision boundaries, the BIC algorithm refines AV-SAD boundaries.

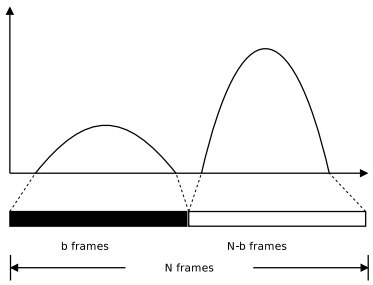

Figure 2: BIC algorithm used in SAD. If there is a change in the signal within a window, we fit a bimodal model. The first b frames fit one Gaussian distribution, and the remaining frames fit a second Gaussian distribution.

4.3. Improving speech boundaries with BIC algorithm

The detection boundary given by the EM algorithm may not be accurate near the transitions between speech and non-speech regions, since the features are similar. Zhou and Hansen [21] proposed the use of BIC [22] for audio stream segmentation. This scheme is suitable to improve the decision boundary precision.

The BIC is a criterion used to select a model among potential candidate models. In the context of SAD, the criterion evaluates whether the data near a transition is better modeled by a single distribution or by bimodal distributions. We can apply BIC to detect changes in the signal by successively splitting a window centered at the boundaries into two partitions. For a given split, the competing hypotheses are (1) $BIC(M_1)$ – the data come from a single Gaussian distribution, and (2) $BIC(M_2)$ – the first b frames belong to one Gaussian distribution and the remaining frames belong to another Gaussian distribution (Fig. 2). These hypotheses are compared by computing the difference of the BIC values for the bimodal and unimodal models. The Δ BIC represents the difference between BIC values (Equation 4). N is the total number of frames in the window, b is the last frame from the first first Gaussian distribution. $\hat{\Sigma}, \hat{\Sigma_1}, \hat{\Sigma_2}$ are the covariances for the N frames, the first b frames, and the last $N-b$ frames, respectively. The feature dimension is d and the $|\cdot|$ represents the determinant.

$$\begin{aligned}\Delta BIC(b) &= BIC(M_2) - BIC(M_1) \\ &= \frac{1}{2}(N\log|\hat{\Sigma}| - b\log|\hat{\Sigma_1}| - (N-b)\log|\hat{\Sigma_2}|) \\ &- \frac{1}{2}(d+\frac{1}{2}d(d+1))\log N \end{aligned} \quad (4)$$

ΔBIC is positive when a bimodal distribution is better than single distributions to describe the data. We can determine the best boundary, by splitting different partitions within the window, picking the one with the highest delta value.

To compute the delta BIC value, we define a window around the potential boundaries (Sec. 4.2). Figure 3 illustrates the search process. The window size is one second centered at the boundary. We evaluate all potential partitions within this window by setting $b = \{1, \ldots, N\}$. Because this method relies on the estimation of covariances, we define heading and tailing windows of 0.2s length which are added to the window to estimate the covariance. In cases where two potential boundaries are too close, we define the window boundaries to the middle point between the two potential boundaries. We compute the Δ BIC for each point in this window, selecting the partition with the highest value. In cases where the Δ BIC values are negative for all the partitions (i.e., BIC favors a single distribution), we do not modify the boundary given by Combo-SAD approach.

Table 1: Performance of SAD for single modality. (*NSen*: normal sentences, *WSen*: whisper sentences. *Pre* is precision, *Rec* is recall, *Acc* is accuracy and *F* means F-score).

Modality	Set	Acc [%]	Pre [%]	Rec [%]	F [%]
A-SAD	Nsen	94.05	97.15	89.85	93.35
	Wsen	67.96	61.02	88.65	72.28
V-SAD	Nsen	78.06	75.11	89.45	80.40
	Wsen	78.20	72.69	89.10	80.06
AV-SAD	Nsen	89.47	97.90	79.93	88.00
	Wsen	81.28	81.73	79.21	80.45

This study uses acoustic (5-D) and visual (26-D) features. We separately evaluate the BIC algorithm in both sets. We also concatenate them to estimate BIC with audiovisual data.

5. Experimental Evaluation

We evaluate the proposed approach on the MSP-AVW corpus. Ground truth is manually labeled based on the audio. We use the following standard metrics for speech detection: accuracy, precision, recall and F-score. The first evaluation considers only the combo SAD framework without the boundary improvement with the BIC algorithm. Table 1 gives the results for A-SAD and V-SAD for neutral and whisper speech. The results show that the combo A-SAD approach is sensitive to speech mode. The performance drops more than 20% (absolute) in accuracy and F-score. While the performance for V-SAD is not as high as the one for A-SAD for neutral speech, the results show that the performance is not affected by speech mode. The table also lists the results for the AV-SAD framework after fusing the decision boundaries. The fusion of the decision boundaries improve the results by taking advantages of the strengths of the A-SAD system for neutral speech and V-SAD for whisper speech.

For long sentences, accuracy, precision, recall and F-score are not very sensitive to improvements in correct boundary detection when most of the region is recognized as speech. This study particularly focuses on improvements in boundary detection in SAD by using the BIC algorithm. Therefore, we also need a metric that captures accuracy around the decision boundaries. We define a *median local boundary mismatch* (MLBM) metric, inspired by the average mismatch metric described by Huang and Hansen [23]. We compute the mismatch between the detected boundary and ground truth in local regions by estimating the absolute value of frames between them. Then, we estimate the median values across all transitions (median is less sensitive to outliers). For a non-speech to speech transition detected by our proposed AV-SAD approach, we consider the closest non-speech to speech transition in the labels to estimate this metric. We follow the same approach for speech to non-speech transitions. MLBM measures the correct boundary detection, ignoring miss-detections, which is reflected on other metrics.

Table 2 lists the results after correcting the decision boundaries of the AV-SAD with the BIC algorithm. The sampling rate is 100 fps across conditions. The first two rows consider the AV-

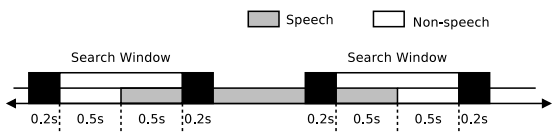

Figure 3: Defining the window for BIC. The gray (speech) and white (non-speech) regions correspond to the decision boundary after fusing A-SAD and V-SAD (Sec. 4.2). We add heading and tailing segments (black) to improve the estimation of the covariance matrices.

Table 2: Performance of the AV-SAD system after correcting the boundaries with the BIC algorithm. (*No BIC*: the result after fusion without BIC correction, *A-BIC*: BIC correction based on acoustic feature, *V-BIC*: BIC correction based on visual feature, *AV-BIC*: BIC correction based on audiovisual features).

AV-SAD Plus	Set	Acc [%]	Pre [%]	Rec [%]	F [%]	MLBM [fps]
No BIC	Nsen	89.47	97.90	79.93	88.00	35
	Wsen	81.28	81.73	79.21	80.45	64
A-BIC	Nsen	91.11	97.47	83.77	90.10	25
	Wsen	82.91	84.47	79.48	81.90	56
V-BIC	Nsen	88.53	92.22	83.18	87.47	42
	Wsen	78.67	76.63	80.54	78.53	71
AV-BIC	Nsen	91.25	97.49	84.05	90.27	25
	Wsen	82.87	83.76	80.37	82.03	53

SAD without BIC correction. These results are the same values reported in Table 1, where we add the MLBM results. The A-BIC condition is when the BIC algorithm is only implemented with the 5-D acoustic features. The BIC correction based on the acoustic feature improves the speech detection performance by about 1%-2% absolute difference. When we consider the MLBM scores, A-BIC achieves 28.5% relative improvement for neutral speech, and 12.5% relative improvements for whisper speech. The V-BIC condition is when the BIC algorithm is only implemented with the 26-D visual features. The decision boundaries do not improve in this case. This result is explained by two reasons (1) the ground truth of the labels was annotated based only on audio, ignoring anticipatory facial activity, and (2) the actual resolution for video features before up-sampling the rate is only 29.97 fps reducing the information to estimate the BIC algorithm. Finally, the AV-BIC condition consider both set of features for BIC correction. For whisper speech, the results are slightly better than the results for the A-BIC condition.

When we compare the F-score of the A-SAD for whisper speech (72.28%) with the best performance obtained by the proposed AV-SAD (82.03% using AV-BIC), we conclude that we achieve around 10% absolute improvement. This result highlights the benefits of audiovisual speech activity detection for multimodal interfaces. We notice that under neutral condition, the performance of the audiovisual model slightly drops compared to the results for A-SAD. It indicates that we need to explore more sophisticated fusion schemes that make our system robust again noise and speech mode, as our proposed system, but do not drop the performance on ideal conditions.

6. Conclusions

This study proposed an unsupervised approach for SAD. We created decision boundaries for acoustic and visual features using the Combo-SAD framework, which relies on PCA and the EM algorithm. We fused the decision boundaries provided by A-SAD and V-SAD using a logical operator. The fused decision boundary is corrected with the BIC algorithm. We evaluated alternatives in how the BIC algorithm was implemented. Most improvements from BIC are obtained when trained with only acoustic features. Under this condition, the accuracy of the decision boundary improves 28.5% (relative) for neutral speech, and 12.5% (relative) for whisper speech.

Future work includes using the framework in actual multimodal interfaces. We are currently evaluating the performance of the approach in noisy recordings to anticipate some of the issues associated with collecting data in less controlled environments. We are also working on audiovisual SAD solutions for portable devices. We expect that future multimodal interfaces will benefit from advances in audiovisual SAD.

7. References

[1] F. Tao, J. Hansen, and C. Busso, "An unsupervised visual-only voice activity detection approach using temporal oro-facial features," in *Interspeech 2015*, Dresden, Germany, September 2015, pp. 2302–2306.

[2] 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.

[3] P. Liu and Z. Wang, "Voice activity detection using visual information," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2004)*, vol. 1, Montreal, Quebec, Canada, May 2004, pp. 609–612.

[4] T. Petsatodis, A. Pnevmatikakis, and C. Boukis, "Voice activity detection using audio-visual information," in *International Conference on Digital Signal Processing (ICDSP 2009)*, Santorini, Greece, July 2009, pp. 1–5.

[5] 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.

[6] 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 filltering," in *European Signal Processing Conference (EUSIPCO 2007)*, Poznań, Poland, September 2007.

[7] 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.

[8] 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.

[9] C. Neti, G. Potamianos, J. Luettin, 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.

[10] G. Papandreou, A. Katsamanis, V. Pitsikalis, and P. Maragos, "Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 3, pp. 423–435, March 2009.

[11] C. Zhang and J. Hansen, "Analysis and classification of speech mode: Whisper through shouted," in *Interspeech 2007 - Eurospeech*, Antwerp, Belgium, August 2007, pp. 2289–2292.

[12] T. Ito, K. Takeda, and F. Itakura, "Analysis and recognition of whispered speech," *Speech Communication*, vol. 45, no. 2, pp. 139–152, February 2005.

[13] 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.

[14] X. Fan, C. Busso, and J. Hansen, "Audio-visual isolated digit recognition for whispered speech," in *European Signal Processing Conference (EUSIPCO-2011)*, Barcelona, Spain, August-September 2011, pp. 1500–1503.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] B. Zhou and J. Hansen, "Efficient audio stream segmentation via the combined T^2 statistic and bayesian information criterion," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 4, pp. 467–474, July 2005.

[22] S. Chen and P. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the Bayesian information criterion," in *Proceedings DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, VA, USA, February 1998, pp. 127–132.

[23] R. Huang and J. Hansen, "Advances in unsupervised audio classification and segmentation for the Broadcast news and NGSW corpora," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 907–919, May 2006.