Improving Acoustic Event Detection using Generalizable Visual Features and Multi-modality Modeling

Po-Sen Huang, Xiaodan Zhuang, Mark Hasegawa-Johnson

Beckman Institute, ECE Department, University of Illinois at Urbana-Champaign, U.S.A.
huang146@illinois.edu, xzhuang2@uiuc.edu, jhasegaw@illinois.edu

Abstract

Acoustic event detection (AED) aims to identify both timestamps and types of multiple events and has been found to be very challenging. The cues for these events often times exist in both audio and vision, but not necessarily in a synchronized fashion. We study improving the detection and classification of the events using cues from both modalities. We propose optical flow based spatial pyramid histograms as a generalizable visual representation that does not require training on labeled video data. Hidden Markov models (HMMs) are used for audio-only modeling, and multi-stream HMMs or coupled HMMs (CHMM) are used for audio-visual joint modeling. To allow the flexibility of audio-visual state asynchrony, we explore effective CHMM training via HMM state-space mapping, parameter tying and different initialization schemes. The proposed methods successfully improve acoustic event classification and detection on a multimedia meeting room dataset containing eleven types of general non-speech events without using extra data resource other than the video stream accompanying the audio observations. Our systems perform favorably compared to previously reported systems leveraging ad-hoc visual cue detectors and localization information obtained from multiple microphones.

Index Terms: acoustic event detection, optical flow, hidden Markov models, multi-stream HMM, coupled hidden Markov models

1. Introduction

Acoustic events help describe human and social activities that occur in many environments. For example, the sound of foot steps or door slamming can be used to detect human activities for surveillance [2] and yawn or chair moving noise reveals audience feedback in a seminar. Detection of nonspeech sounds also helps improve speech recognition performance [3].

Acoustic event detection (AED) aims to identify both timestamps and types of multiple events and has been found to be very challenging. The CLEAR 2007 AED Evaluation and follow-up work [4, 5] highlighted research efforts and challenges in the detection of general acoustic events, in contrast to highlight/key events, such as explosion. In particular, the acoustic footprints of the events are very fuzzy and subject to noise.

Recently, incorporating both audio and visual information for AED has been demonstrated as an effective approach to improve the performance and robustness over the audio-only systems [1, 6, 7]. However, these works either leverage on specific visual object detectors, usually requiring hand-labeled training data, or expect dominance or strong prior of the visual cues in the recorded video, sometimes impossible for real applications.

Leveraging additional visual cues for audio signal analysis has been explored in other applications, such as speech recognition [8] and person identification [9]. In particular, the multi-stream HMM and the couple HMM (CHMM) are two effective models for audio-visual fusion. While audio-visual event detection shares a lot of challenges with audio-visual speech recognition, they differ in multiple ways: First, the visual cues for general acoustic event detection can be much less constrained: there is no consistent visual region, such as the mouth in audio-visual speech processing, in which all the event information is embedded. Second, the synchrony and asynchrony between the two modalities is not governed by a well constrained mechanism, such as human speech articulation. For example, key jingling presents mostly simultaneous audio and visual footprints, but we can observe a person move before or after s/he makes the footstep sound, or a door start moving before making a slamming sound, the asynchrony being more arbitrary than what is observed in audio-visual speech. It is not yet studied whether the audio-visual models in speech processing can be effectively applied in audio-visual event modeling to improve acoustic event detection.

In this work, we study using a generalizable visual representation to improve acoustic event detection, via different audio-visual synchrony and asynchrony modeling. In particular, a combination of optical flow and overlapping spatial pyramid histograms characterizes the visual cues, which can be non-dominant in the recorded video. Compared with more task-specific alternatives [1], the proposed visual features have the merit of requiring minimum labeling efforts: no extra labels required other than the event onset/offset timestamps used for audio-only modeling. We propose applying multi-stream HMMs for synchronized audio-visual event modeling and coupled hidden Markov models [8] for more flexible modeling allowing asynchrony.

Acoustic event detection and classification experiments are performed on meeting room data with eleven general non-speech acoustic events. With the proposed visual representation and multi-modal modeling, the visual cues, often local and subtle in the images, are shown to consistently improve both classification and detection accuracy of the concerned events. All the experiments use the video associated with the audio as the only extra data resource, requiring no additional labeling.

The organization of this paper is as follows. Section 2 presents the generalizable visual features adopted in this work, in particular the overlapping spatial pyramid histograms based on optical flow. Section 3 discusses the audio-visual modeling methods, in particular the multi-stream HMM and the coupled HMM. Section 4 presents the experimental results on audio-visual event classification and detection. We conclude the paper in Section 5.

2. Generalizable Visual Features for AED

Previous literatures [1] reported using ad-hoc visual detectors to generate visual features for the purpose of improving event detection. However, training these detectors requires expensive labeling efforts, usually at least bounding boxes of the concerned objects. Moreover, these detectors are task-specific. Alternatively, we explore using visual features that do not require such training and data labeling, and are not task-specific, i.e. generalizable.

In this work, we propose using a combination of optical flow and overlapping spatial pyramid histograms to characterize the visual cues in the acoustic events.

This research is supported by ARO MURI 2009-31, and by NSF08-07329. We would like to thank the authors of [1] for providing the dataset.

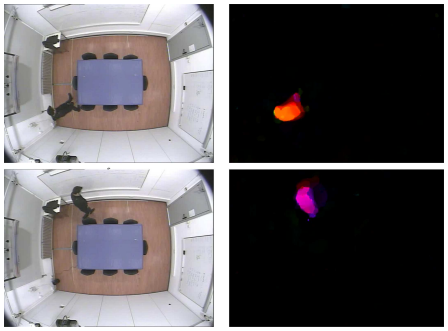

Figure 1: (Left) The image sequence for the event 'foot step' in the overhead camera; (Right) the corresponding optical flow fields for each image, where the flow field is visualized using hue to indicate the direction and intensity for the magnitude.

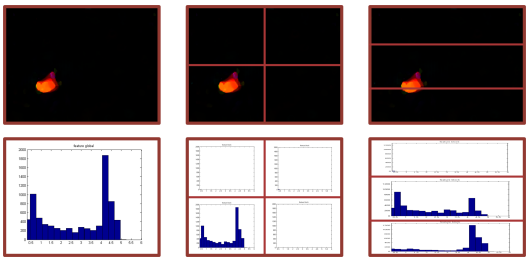

Figure 2: Optical flow based overlapping spatial pyramid histograms for a footstep event: (first row) Spatial pyramid arrangement and optical flow magnitude visualization; (second row) Optical flow magnitude histogram in each corresponding block.

The visual cues of the non-speech audio-visual events are mostly related to motion. We propose using visual features based on optical flow between consecutive frames to capture the movement information. We utilize a highly efficient algorithm on variational methods utilizing a GPU [10] to calculate the optical flow, i.e. the horizontal and vertical movement for each pixel. Fig. 1 illustrates the extracted optical flow for a "foot step" event.

The visual cues of the acoustic events have their spatial correlates: the spatial distribution sometimes, but not always, differs between the different events and the background. Therefore, we define eight overlapping blocks from the whole image, including both the complete image and seven spatially local regions. The histograms of motion vector magnitude within all the blocks are employed as the video features [11]. We refer to this representation as the **overlapping spatial pyramid histograms**. Similar representation was successfully used for kernel estimation in general image scene categorization [12], which shares the property that the visual cues are highly variant and sometimes localized.

An example of the proposed visual representation for a 'foot step' event is illustrated in Fig. 2.

3. Multi-Modality Fusion for AED

We propose using multi-stream HMMs for synchronized audio-visual event modeling, and coupled hidden Markov models [8] for more flexible modeling allowing asynchrony.

Different fusion methods have been explored for the audio and visual modalities. First, feature fusion techniques include plain feature concatenation [13], feature weighting [14] and a data-to-data mapping of either one modality into the space of another or both modalities into a new common space [15]. Second, decision fusion provides a mechanism for capturing reliabilities of each modality by

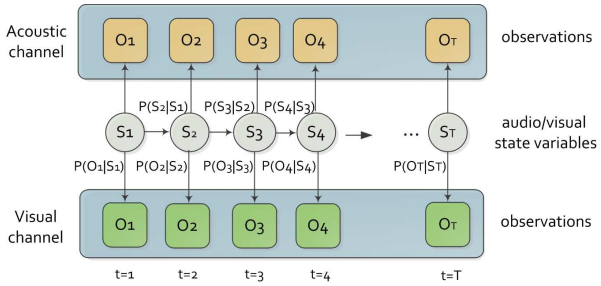

Figure 3: A two stream hidden Markov model encoded as a dynamic Bayesian network

classifier combination. Third, intermediate fusion performs multimodal integration at a level between decision fusion and feature fusion. Intermediate integration strategies have been shown to outperform the early and late integration strategies in various applications [16].

Multi-stream HMMs and coupled HMMs are used as two intermediate fusion methods . The synchrony and asynchrony between the modalities are modeled by the hidden state transitions. Though such models have been successfully applied in audio-visual speech recognition [8], they have not been applied in improving general non-speech acoustic event detection.

3.1. Multi-stream Hidden Markov Models

In a two-stream HMM, the state-dependent emission of the audiovisual observation $o_{av,t}$ is governed by $P(o_{av,t}|S_t) = P(o_{a,t}|S_t)^{\lambda_{a,S_t,t}} P(o_{v,t}|S_t)^{\lambda_{v,S_t,t}}$ for all HMM states S_t, where $\lambda_{s,S_t,t}$ denotes the nonnegative stream weights, which models the stream reliabilities as a function of modality s, HMM state S_t and time t.

Multistream HMMs assume the state synchrony between audio cues and visual cues. Because of the simple topology, it's relatively easy to obtain robust estimation of the parameters.

Fig. 3 illustrates a two-stream HMM, where the transitions probabilities are referred to as $P(S_t|S_{t-1})$. State observation distributions are referred to as $P(o_{av,t}|S_t)$. S_t is a multinomial random variable representing the state of the CHMM system variable at time t. Note, both the streams progress in a synchronous fashion.

3.2. Coupled Hidden Markov Models

The assumption of audio-visual state synchrony is not always satisfied. For example, in an object dropping event, the acoustic sound is not always in existence when the object is in motion, but when the object stops dropping. Similarly, a door slamming sound occurs in the end of the door movement. Though the asynchrony between modalities can be alleviated by a larger local time window for each frame, a more flexible statistical model allowing asynchrony between the hidden state sequences for the two modalities is desired.

In this work, we propose using coupled HMM to model modality asynchrony in audio-visual events. We select the transition-only Coupled Hidden Markov Model (CHMM), in which different modalities are coupled through state transitions. The CHMM is capable of capturing both the synchronous and asynchronous inter-modal dependencies between two information channels. CHMM proves to be an effective method in audio-visual speech recognition [8].

CHMM can be viewed as parallel rolled-out HMM chains coupled through cross-time and cross-chain conditional state transition probabilities. An n-chain CHMM has n hidden nodes in a time slice, each connected to itself and its nearest neighbors in the next time slice. In our task, we use a 2-chain CHMM for audio-visual modeling, as shown in Fig. 4, where circular nodes in each slice are the

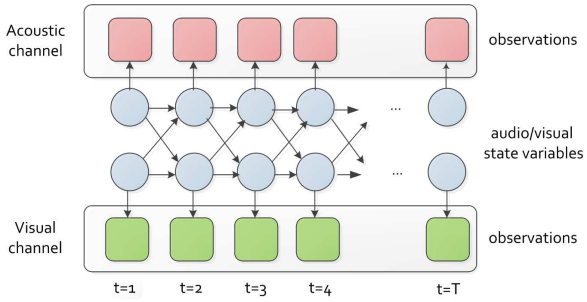

Figure 4: Audio-visual fusion using CHMM

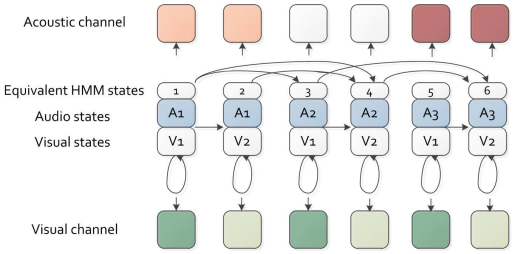

Figure 5: Converting a CHMM to an equivalent HMM by state-space mapping and parameter tying

multinomial state variables, square nodes in each slice represent the observation variable, and the directed links represent conditional dependence between nodes.

The state of the CHMM system in each time slice is jointly determined by the two multinomial state variable, each depending on its two parent states in the previous time slice. The configuration permits unsynchronized progression of the two chains while keeping the Markov property that a future state variable is conditionally independent of the past given the present state variables. Note that CHMM can be seen as a generalized multi-stream HMM.

Following a transformation strategy based on state-space mapping and parameter tying [8], we can convert a CHMM to an equivalent HMM, whose hidden states each corresponds to the state of the system described by the CHMM. The number of hidden states in the equivalent HMM equals the number of possible combinations of states from both modalities. Fig. 5 illustrates a 2-chain CHMM with $Q_a = 3$ and $Q_v = 2$, where Q_a and Q_v are the numbers of audio and visual states respectively. For example, state 3 in the equivalent HMM corresponds to the CHMM state defined by audio state $q_a = 2$ and visual state $q_v = 1$. The modality-dependent observation probabilities corresponding to the same observation distribution in the original CHMM are tied and coded using the same tag. For example, the output densities modeling the visual stream in state 1, 3, 5 are tied and tagged as "V_1", because they correspond to $P(O_1|q_v = 1)$ in the CHMM.

In this work, we use a left-to-right non-skip HMM for each of the two modalities in the CHMM. The allowed state transitions in the equivalent HMM are derived from state space mapping. For example, in the state diagram in Fig. 5, given state 1 ($q_a = 1, q_v = 1$) at present, in next time slice, q_a can either transit to $q_a = 2$ or stay in $q_a = 1$, and q_v can either transit to $q_v = 2$ or stay in $q_v = 1$. Hence, state 1 can either stay in itself or transit to CHMM state 2 ($q_a = 1, q_v = 2$) or state 3 ($q_a = 2, q_v = 1$) ,or state 4 ($q_a = 2, q_v = 2$).

For robust estimation of the CHMMs, we perform the CHMM training in two stages. In the first stage, the observation distributions for both modalities are initialized using simpler models. The initial simpler models can be a two-stream audio-visual HMM, which requires strict state synchrony between audio and visual modalities; or one audio-only HMM and one video-only HMM, which impose no explicit state correspondence between the two modalities. In the second stage, the audio and visual observation distributions from the multi-stream HMM or two single-modality HMMs are used to construct the CHMM-equivalent HMM. Additional parameter estimation iterations using the Balm-Welch algorithm are performed with this HMM.

4. Experiments

4.1. Dataset and Setup

We use the audio-visual dataset collected by the Universitat Politecnica de Catalunya [1]. The database contains multimodal recordings of acoustic events (AEs) in a meeting room environment. The target events in this dataset include: Knock door/table (kn), Door slam (ds), Steps (st), Chair moving (cm), Spoon/cup jingle (cl), Paper work - listing, warping (pw), Key jingle (kj), Keyboard typing (kt), Phone ringing/Music (pr), Applause (ap) and Cough (co). There are approximately 90 instances per event class for the whole dataset of six sessions (S01-S06). Among S01-S04, we use three sessions for training, and one for testing. All reported measures are averaged from four-fold cross validation. Additional two sessions (S05, S06) are used as the development set. We use the observations from a far field microphone and an overhead camera.

To make the task more realistic we add different levels of Gaussian white noise to the clean recorded audio, to illustrate the performance of the different approaches at different noise levels. Perceptual Linear Prediction coding (PLP) coefficients are used as the audio features. In particular, PLP coefficients, including 12 coefficients and the 0^{th} cepstral coefficient, are extracted from 30ms Hamming windows with a temporal step of 20ms. The delta and acceleration coefficients are computed and appended to the static PLP coefficients. Cepstral mean normalization is performed on each recorded session.

The visual features are obtained according to Section 2 using 20 bins for each histogram of optical flow magnitude. The concatenation of histograms from all blocks is projected into 40 dimensions using Principle Component Analysis, retaining 98% of the total energy. These visual features are interpolated to match the 20ms frame period of the audio features.

In this work, each multistream HMM or CHMM has 4 audio and 4 video states with stream weights tuned on the development data using coarse-to-fine grid search. For simplicity, the stream weights are time-invariant. The different methods are evaluated using classification accuracy and detection accuracy AED-ACC [1, 4]. A set of audio-only HMMs are used for comparison, given their effectiveness [17].

4.2. CHMM Training Schemes

Initialization of the observation distributions in the CHMM is important, because of the high degree of freedom in the CHMM topology. As discussed in Section 3, we explore two different initialization schemes for CHMM, referred to as **CHMM**$_m$ and **CHMM**$_s$, in which the observation distributions of the CHMMs are initialized using multistream HMMs, or pairs of audio-only and video-only HMMs respectively.

The CHMMs parameters (the Gaussian means, covariance, mixtures weights, and the state transition probabilities) are further estimated with a few iterations using the Balm-Welch algorithm. We found in our pilot experiments that allowing estimation of all the CHMM parameters above is better than estimating any subset of parameters above and using the initialized parameters for the rest.

4.3. Results

Table 1 and Table 2 present the classification and detection results using the proposed visual representation coupled with different audio-visual modeling methods as well as the audio-only and video-only models. In both detection and classification, the multistream HMM system consistently improves from the audio-only system as well as

the video-only system for all SNR conditions studied in this work. Further, CHMM-based systems (CHMM$_s$ and CHMM$_m$) outperform the multistream HMM system in event detection for all SNR conditions.

We also performed event detection using original clean audio, the same condition studied in [1]. The proposed visual features and audio-visual modeling perform favorably, compared to the best systems reported in [1]. These reference systems [1] leverage a person tracker, a laptop detector, a face detector, a door activity estimator to capture the visual cues and optional localization information obtained from multiple microphones (denoted as "AV" and "AVL" in Table 2 respectively).

Fig. 6 shows the confusion matrix of event classification using the audio-only HMM, audio-visual multistream HMM, CHMM$_m$ and CHMM$_s$ systems. Using the proposed generalizable visual features with the multistream HMM or the CHMM boosts classification accuracy for most event classes compared to the audio-only system. The more flexible CHMM-based systems (CHMM$_s$ and CHMM$_m$) further improve classification of some events, such as kn: knock (door, table) and co: cough from the multistream HMM system.

To verify that the audio-visual state asynchrony allowed by the CHMM systems is utilized, we examine the state sequences found by the Viterbi decoding. The percentages of observation frames claimed by the CHMM states defined by an asynchronous pair of audio and video states are 65.944% for CHMM$_s$, and 65.842% for CHMM$_m$ respectively. Note that the multistream HMM system assigns all frames to states that are defined by synchronous audio and visual states.

Classification Accuracy (%) mean±standard error					
SNR	Audio-only	Video-only	Multistream	CHMM$_m$	CHMM$_s$
10dB	28.05±4.40	61.57±3.18	64.35±4.35	67.22±3.76	65.76±4.36
20dB	51.54±5.21	61.57±3.18	72.33±6.15	76.40±5.87	76.92±5.09
30dB	77.45±6.96	61.57±3.18	89.07±4.13	89.12±3.51	87.10±4.36

Table 1: Classification accuracy with different audio SNR. ("Multistream": the bimodal system using multistream HMMs. "CHMM$_m$": the CHMM-based system initialized using multistream HMMs. "CHMM$_s$": the CHMM-based system initialized using audio-only and video-only HMMs.)

Detection Accuracy (%) mean±standard error					
SNR	Audio	Video	Multistream	CHMM$_m$	CHMM$_s$
10dB	26.73±6.99	45.22±2.22	45.45±3.04	50.47±2.97	48.35±2.33
20dB	47.96±6.03	45.22±2.22	63.74±3.78	65.89±3.98	66.28±3.95
30dB	69.35±5.26	45.22±2.22	78.55±4.13	79.50±2.71	79.54±2.27
clean	87.54±2.99	45.22±2.22	90.57±2.07	91.85±2.11	90.79±2.97
clean	"AV" [1]	85	"AVL" [1]	86	

Table 2: Detection accuracy with different audio SNRs. ("AV": [1] system using video features from multiple ad-hoc detectors. "AVL": "AV" system plus localization information obtained via multiple microphones [1].

5. Conclusion

In this work, we study using generalizable visual features to improve acoustic event detection via audio-visual intermediate integration. We propose using optical flow based spatial pyramid histograms to represent the highly variant visual cues for the acoustic events. This representation is demonstrated to significantly improve event classification and detection using systems based on multistream HMMs or coupled HMMs. Compared to the multistream HMMs, the coupled HMMs further boost the performance by allowing state asynchrony between the audio and visual modalities. Our systems with the proposed generalizable visual features and audio-visual modeling perform favorably compared to previously reported systems leveraging ad-hoc visual cue detectors and localization information obtained from multiple microphones [1].

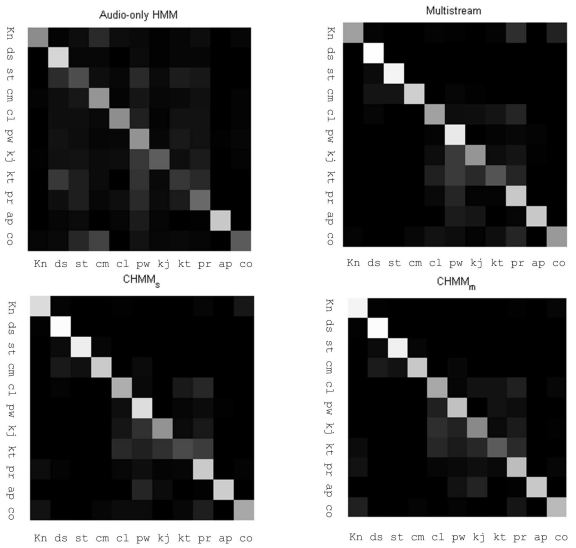

Figure 6: Confusion Matrix for Event Classification (averaged over SNRs 10dB, 20dB, 30dB) based on audio-only HMM, audio-visual multistream HMM, CHMM$_m$ and CHMM$_s$ respectively.

6. References

[1] C. Canton-Ferrer, T. Butko, C. Segura, X. Giro, C. Nadeu, J. Hernando, and J. Casas, "Audiovisual event detection towards scene understanding," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2009, pp. 81–88.

[2] C. Clavel, T. Ehrette, and G. Richard, "Events detection for an audio-based surveillance system," in *IEEE International Conference on Multimedia and Expo*, 2005.

[3] F. Beaufays, D. Boies, M. Weintraub, and Q. Zhu, "Using speech/non-speech detection to bias recognition search on noisy data," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2003.

[4] A. Temko, "CLEAR 2007 AED evaluation plan," http://isl.ira.uka.de/clear07, 2007.

[5] X. Zhuang, X. Zhou, M. A. Hasegawa-Johnson, and T. S. Huang, "Real-world acoustic event detection," *Pattern Recognition Letters*, vol. 31, no. 12, pp. 1543 – 1551, 2010.

[6] M. R. Naphade, A. Garg, and T. Huang, "Duration-dependent input-output Markov models for audio-visual event detection," in *International Conference on Multimedia*, 2001.

[7] Z. Xiong, R. Radhakrishnan, A. Divakaran, and T. Huang, "Audio-visual event recognition with application in sports video," in *Intelligent Multimedia Processing with Soft Computing*, ser. Studies in Fuzziness and Soft Computing, 2005, vol. 168, pp. 129–149.

[8] S. M. Chu and T. S. Huang, "Audio-visual speech modeling using coupled hidden Markov models," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2002.

[9] C. Sanderson and K. K. Paliwal, "Identity verification using speech and face information," *Digital Signal Processing*, vol. 14, no. 5, pp. 449 – 480, 2004.

[10] C. Zach, T. Pock, and H. Bischof, "A duality based approach for realtime tv-l1 optical flow," in *Pattern Recognition (Proc. DAGM)*, Heidelberg, Germany, 2007, pp. 214–223.

[11] N. Ikizler, R. Cinbis, and P. Duygulu, "Human action recognition with line and flow histograms," in *19th International Conference on Pattern Recognition*, 2008.

[12] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *CVPR*, 2006.

[13] A. Adjoudani and C. Benoit, "On the integration of auditory and visual parameters in an HMM-based ASR," *In D. G. Stork and M. E. Hennecke (Eds.), Speechreading by Humans and Machines. Berlin: Springer-Verlag*, pp. 461–471, 1996.

[14] P. Teissier, J. Robert-Ribes, and J. L. Schwartz, "Comparing models for audiovisual fusion in a noisy-vowel recognition task," *IEEE Trans. Speech Audio Processing*, vol. vol. 7, pp. 629–642, 1999.

[15] K. Chaudhuri, S. M. Kakade, and K. Livescu, "Multi-view clustering via canonical correlation analysis," in *In Proc. of ICML09*, 2009.

[16] S. Nakamura, "Statistical multimodal integration for audio-visual speech processing," *IEEE Transactions on Neural Networks*, vol. 13, no. 4, 2002.

[17] X. Zhuang, X. Zhou, T. S. Huang, and M. Hasegawa-Johnson, "Feature analysis and selection for acoustic event detection," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2008.