On the Improvement of Multimodal Voice Activity Detection

Matt Burlick\textsuperscript{1}, Dimitrios Dimitriadis\textsuperscript{2}, Eric Zavesky\textsuperscript{2}

\textsuperscript{1}Computer Vision Laboratory, Stevens Institute of Technology, Hoboken, NJ, USA. \textsuperscript{2}AT&T Labs - Research, Florham Park, NJ, USA.

mburlick@stevens.edu, ddim@research.att.com, ezavesky@research.att.com

Abstract

As mobile devices, intelligent displays, and home entertainment systems permeate digital markets, the desire for users to interact through spoken and visual modalities similarly grows. Previous interactive systems limit voice activity detection (VAD) to the acoustic domain alone, but the incorporation of visual features has shown great improvement in performance accuracy. When employing both acoustic and visual (AV) information the central recurring question becomes “how does one efficiently fuse modalities”. This work investigates the effects of different features (from multiple modalities), different classifiers, and different fusion techniques for the task of AV-VAD on data with varying acoustic noise. Furthermore we present a novel multi-tier classifier that combines traditional approaches, feature fusion and decision fusion, with independent modality classifiers and combined intermediary decisions with raw features as inputs to a second stage classifier. Our augmented multi-tier classification system concatenates the output of a set of base classifiers with the original fused features for a final classifier. Experiments over various noise conditions show average relative improvements of 5.0% and 4.1% on the CUAVE\textsuperscript{1} dataset and 2.6% and 11.1% on the MOBIO\textsuperscript{2} dataset over majority voters and LDA respectively. 

\textbf{Index Terms:} Multi-Modal Fusion, Computer Vision, Voice Activity Detection

1. Introduction

Voice activity detection allows interactive systems with speech recognition capabilities to determine when a person starts or stops talking (speech endpointing). Problems arise when the acoustic signal-to-noise ratio (SNR) decreases and information carried by an acoustic signal used to detect voice activity is lost, rapidly degrading performance in audio only VAD (A-VAD)\textsuperscript{3} [3]. Throughout the body of work in audio-only VAD (A-VAD) and video-only VAD (V-VAD) many different features have been proposed and evaluated (predominantly mel-frequency cepstral coefficients, zero-crossings and signal energy for A-VAD and lip parameterization, Discrete Cosine Transform coefficients and Local Binary Patterns for V-VAD), some of which have also been tested for their robustness against the existence of noise in their respective modality.

Inspired by the way humans use multiple senses for information validation, recent work has incorporated multiple modalities for VAD\textsuperscript{4} [4, 5, 6, 7, 8]. These modalities are typically audio and video, which when used together, form an audio-visual voice activity detection (AV-VAD) system. 

In speech-based applications and beyond, there has also been a variety of methods proposed for the fusing of multiple modalities, often tuned for one scenario. These fusion techniques include both simple and straightforward approaches, like feature concatenation and majority voting, and more sophisticated approaches, like dynamic feature weighting schemes. As to the more sophisticated methods, authors in [5] blend the responses based on acoustic SNR whereas authors in [7] use feature uncertainty. However, both methods either assume prior knowledge of the SNR or model the feature uncertainty.

In this paper we explore the performance of several acoustic and visual features for the task of voice activity detection as corrupted by additive noise signals, Section 2. We also examine the effectiveness of several basic fusion techniques with both early and late fusion approaches. The former includes approaches that first fuse features and then use a single classifier to determine the presence of speech. The latter first applies several classifiers (one per feature) whose outputs are then combined to make a final decision. We propose a method that uses both of these approaches and incorporates many features from both modalities. In contrast to prior works, our approach incorporates a large number of different features, whose independent classifiers can conceptually be thought of as weak classifier inputs to our multi-tier system. Authors in [6] recently used such a multi-tiered approach, but only a few features were considered (face certainty, lip dimensions, and a decoder-based audio feature) and only Bayesian reasoning fuses the responses. In contrast, by taking into account raw features and the output of the first tier of classifiers for the training of a second classifier, our approach learns when to “trust” outputs generated by the first-tier; effectively simulating a dynamically switching decision plane based on input features. These classifiers are described in detail in Section 3 and evaluated on two fully hand-
Finally, note that in this work we are exploring instantaneous AV-VAD. That is to say, we are only examining the current frame of audio/video when classifying it as speech or non-speech, without introducing contributions from the following or preceding frames. This is in contrast to other work that typically uses temporal smoothing via multiple frames, Hidden Markov Models (HMMs), or other multi-sample methods, which essentially include temporal information for subsequent higher-level interaction decisions. The classifiers evaluated in this work may act as a potential front-end to these aforementioned systems. Additionally, an instantaneous method can be considered a fallback for a continuous solution when there is signal degradation as a result of dropped video or audio packets due to low-latency but high-bandwidth requirements (i.e. remote/cloud processing).

2. Features and Feature Extraction

Two keystones in AV systems are its low-level features and how they are fused. The following sections introduce those used in prior works and the proposed final combination.

2.1. Acoustic Features

Historically, the most commonly used acoustic features for the task of A-VAD, AV-VAD, and speech recognition include zero-crossings, signal energy, and Mel-frequency cepstral coefficients (MFCCs). Of these, the most commonly used feature for VAD tasks is MFCCs [5, 7, 8]. To model the signal dynamics, several works, including [5], use the first and second derivatives of the MFCCs in addition to the static feature.

In this work we examined MFCCs and audio segment signal energy, in addition to their first derivatives. The MFCCs are extracted using the Intel AVCSR tracker [9] and computed by segmenting the acoustic data into 20ms audio frames with 10ms overlap. We only look at the MFCCs of the audio frame closest (temporally) to the current video frame. For example, given a video frame rate of 30fps, several frames are available for every video frame.

2.2. Visual Features

For visual features, prior work has looked at both parametric and appearance-based features. Parametric features proposed in [6, 10, 11] extract the mouth region and use the width and the height of the lips. Work in [7] uses the parameters of an active appearance model fitted to the lower face region. For appearance features, proposed features include mouth region intensity (grayscale) histograms [12], discrete cosine transforms (DCTs) [5], the mean intensity of the mouth region [10], local binary patterns (LBPs) [13], and horizontal and vertical mean and variance vectors based on the optical flow [8].

This work explores four visual features: the mouth height and width (lip dimensions), the mouth region intensity distribution, and the mouth region DCTs and LBPs, as well as their first time derivatives. The lip dimensions form a 2D feature vector using the estimated height and width. From the mouth region we compute a 32-bin grayscale histogram, a 14D DCT feature (as performed in [5]), and a 256D LBP feature (as performed in [14]). We also use the first derivatives of these features to encode the feature dynamics.

Table 1 provides a list of the features we explored and their length.

<table>
<thead>
<tr>
<th>Feature Name</th>
<th>Modality</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>Audio</td>
<td>13</td>
</tr>
<tr>
<td>Delta MFCC</td>
<td>Audio</td>
<td>15</td>
</tr>
<tr>
<td>Energy</td>
<td>Audio</td>
<td>1</td>
</tr>
<tr>
<td>Delta Energy</td>
<td>Audio</td>
<td>1</td>
</tr>
<tr>
<td>Lip Dimensions</td>
<td>Video</td>
<td>2</td>
</tr>
<tr>
<td>Delta Lip Dimensions</td>
<td>Video</td>
<td>2</td>
</tr>
<tr>
<td>Mouth Intensity Distribution</td>
<td>Video</td>
<td>32</td>
</tr>
<tr>
<td>Delta Mouth Intensity Distribution</td>
<td>Video</td>
<td>32</td>
</tr>
<tr>
<td>Mouth DCT</td>
<td>Video</td>
<td>14</td>
</tr>
<tr>
<td>Delta Mouth DCT</td>
<td>Video</td>
<td>14</td>
</tr>
<tr>
<td>Mouth LBP</td>
<td>Video</td>
<td>256</td>
</tr>
<tr>
<td>Delta Mouth LBP</td>
<td>Video</td>
<td>256</td>
</tr>
</tbody>
</table>

Table 1: Features evaluated for audio-visual voice activity detection, their modalities and vector lengths.

2.3. Feature Synchronization

To guarantee synchronization between the audio and video streams, an audio buffer is flushed after processing the previous video frame and populated with audio data until the next video frame. At this point the audio is downsampled to 16kHz and acoustic and visual features are extracted on the audio buffer and video frame, respectively as described in Sections 2.1 and 2.2. After this preprocessing, acoustic and visual features are available for every video frame.

Two sets of visual features (appearance-based and parameter-based) are derived from detected mouth regions. Appearance-based features are extracted after mouth regions are detected in one of three ways. First, faces are detected by a pre-trained Haar cascade [15]. For each face found, we attempt to find the mouth using another pre-trained Haar cascade with its search region constrained to the lower half of the face. If no mouth is detected, a second method loosely fits a region around mouth points detected from a pre-trained Active Shape Model (ASM) [16]. Finally, if no ASM model can be fit for the face, we heuristically select a region relative to the detected face region. From the mouth region we compute the intensity histogram, the DCTs, and the LBPs. Parameter-based features for lips are extracted from detected mouth regions. After fitting an ASM to the mouth alone, pre-defined points on this model are used to compute the width and height of the lips.

Acoustic features are extracted for each video frame, based on an accumulated buffer. This buffer is then used to obtain the acoustic features described in Section 2.1.

3. Classifiers

In investigating potential classifiers for voice activity detection we focused on discovering classifier models that were both robust to noise and generic enough to accommodate features of different types and dimensionality. We first looked at the performance of different classification approaches based on different feature types. We call these base classifiers. We then looked at different traditional early and late fusion methods and proposed our own augmented multi-tier classifier that combines concepts from both early and late fusion techniques.

686
We also selected the most discriminating features using the Fisher Discriminant Ratio (FDR) \cite{18} and selected projected features using Linear Discriminant Analysis (LDA) \cite{19}. To determine the number of features to keep, features were greedily removed (ranked by the FDR values or eigenvalues) and chose the number of features that provided the best training rate. For late fusion we evaluated at majority voting and weighted majority voting. For the latter, the weights are calculated according to the training success rate, as determined by validation on a sub-partition of training data.

### 3.2.1. Multi-Tiered Classifier

We created a two-tier classifier approach to maximize late fusion performance. Here, the output of base classifiers (the random forests, one built per feature vector) are treated as input to a second classifier (also random forests). By training a second classifier on the output of the base classifiers, the system learns a non-linear way to best fuse those outputs.

### 3.2.2. Augmented Multi-Tiered Classifier

In analyzing the performance of various classifiers, we noticed that while the multi-tiered approach worked well, it was often outperformed by the simple fused feature approach and/or by weighted majority voting. Through experimentation we found that augmenting the multi-tier classifier to also include the raw fused features, as seen in Fig. 2, further improved classification results. This new classifier leveraged the advantages of both the two-tiered approach as well as the fused feature approach and allowed a final classifier to learn dynamic fusion strategies according to raw feature input. The final evaluation (discussed in 4.1) did demonstrate gains in classification rates for the majority of acoustic environments across both datasets.

### 4. Evaluation

To evaluate the performance of both single- and fused-classifiers, we used the publicly available MOBIO \cite{2} and CUAVE \cite{1} datasets. Both datasets consist of users speaking to a camera sitting directly in front of them. The CUAVE dataset restricts itself to spoken digits whereas the MOBIO dataset does not, allowing for varying expressiveness. While the CUAVE datasets came with ground-truth, the MOBIO datasets was analyzed with a forced alignment method \(^{1}\) to label sections of each video as speech or non-speech at the audio frame resolution (10ms). Only the normal and moving poses of the CUAVE videos were included in experiments. The CUAVE videos were also re-encoded to remove detected audio jitter, resulting in a minor loss of video quality. Fig. 3 shows example video frames from these databases.

\(^{1}\)We used an acoustic model based on MFCC features to perform fully automated segmentation of the audio signals.
Correct Frame Classification Rates (%) for MOBIO

<table>
<thead>
<tr>
<th>Feature</th>
<th>-20dB</th>
<th>-10dB</th>
<th>0dB</th>
<th>10dB</th>
<th>20dB</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (random)</td>
<td>50.0</td>
<td>50.0</td>
<td>50.0</td>
<td>50.0</td>
<td>50.0</td>
<td>50.0</td>
</tr>
<tr>
<td>MFCC</td>
<td>52.7</td>
<td>61.1</td>
<td>65.8</td>
<td>75.9</td>
<td>85.4</td>
<td>68.2</td>
</tr>
<tr>
<td>∆ MFCC</td>
<td>50.0</td>
<td>53.1</td>
<td>63.1</td>
<td>69.7</td>
<td>74.3</td>
<td>62.5</td>
</tr>
<tr>
<td>Energy</td>
<td>50.0</td>
<td>50.0</td>
<td>50.0</td>
<td>57.6</td>
<td>68.8</td>
<td>55.3</td>
</tr>
</tbody>
</table>

Table 3: Instantaneous audio-visual VAD accuracy on CUAVE and MOBIO datasets for different acoustic SNR levels. Bold numbers indicate best accuracy for individual features and fused systems at each SNR level. Note that LBP-based features were not included in the FDR and LDA classifiers due to computational time needed to exhaustively search for the number of features to keep.

4.1. Results

Features were extracted as in Sec. 2 and classifiers were constructed as in Sec. 3. For evaluation, SNR accuracy was averaged over five runs, each time selecting at random \( \frac{1}{3} \) of the samples for training and the remaining \( \frac{2}{3} \) are used for testing (where we define a sample as the set of visual and acoustic features packaged together as a visual frame as in Sec. 2.3). From each of these sets we ensured that 50% of the samples were positive (speech) and the remaining were negative (non-speech) samples so that our baseline (random) accuracy is 50.0. While in the real world this may not be true (and likely is not), we enforced this prior as an implementation detail. We also acknowledge that using this approach it is possible that samples from the two sets may fall temporally close to one another (thus potentially resulting in inflated accuracy). However there is still diversity by including multiple people and relative realized gains can still be observed.

We performed evaluation on all classifiers mentioned in Sec. 3. As mentioned, we use random forests as our classifier type due to its performance over other tested types (see Table 2). We measure classification accuracy with binary class (speech, nonspeech) accuracy as defined in equation 1.

\[
\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}
\]  

Table 3 shows the average accuracy for each SNR value as well as the averaged accuracy for all of the base features/classifiers and examined fusion classifiers.

First, these results demonstrate that our proposed system generally outperforms other modalities and fusion methods. Surprisingly, weighted majority voting also performed well in low-SNR environments. However, in mid-SNR environments (0-20dB), those typically encountered in real-world scenarios, our system outperformed the weighted majority classifier relatively on average by 2.9% on CUAVE and 2.8% on MOBIO. Second, these results verify that in low SNR conditions, visual features (DCT and LBP) generally contribute the most to performance. We also observe that in audio, MFCCs work and ∆ Energy seems most robust to noise.

These performance gains come with minimal runtime costs when compared to more expensive feature projection methods like FDR and LDA. During the evaluation, informal tests demonstrated that our two-tiered classifier had runtime performance similar to a majority voters classifier. Extracting the twelve features (see Table 1) and performing classification could be done on average at 21.33 \( \text{fps} \) using a two-tiered classifier and 21.36 \( \text{fps} \) using a majority weighted voting classifier on a desktop machine and videos of size 640x480.

5. Conclusions and Future Work

In this work we explored various features and classifiers from the acoustic and visual modalities for the task of voice activity detection. We also looked at several fusion techniques and proposed an augmented multi-tier classifier approach. On average, our proposed classifier boosted performance by a relative 4-5% over other fusion methods in environments with 0-20dB audio SNR on two public datasets. Moving forward we plan to incorporate temporal information to improve accuracy by response smoothing to filter out low-pass classification noise and better prepare the system for high-level AV event detection.

6. References


[9] Intel, “AVCSR Toolkit”.
http://sourceforge.net/projects/opencvlibrary/


http://docs.opencv.org/modules/ml/doc/random_trees.html