DISTANT MULTI- SPEAKER VOICE ACTIVITY DETECTION USING RELATIVE ENERGY RATIO

Gang Chen ∗1, Kenichi Kumatani ∗2, John McDonough #3, Bhiksha Raj #4

∗ Disney Research, Pittsburgh
4720 Forbes Ave., Lower Level, Suite 110, Pittsburgh, PA 15213, USA
1gang.chen@disneyresearch.com, ∗kenichi.kumatani@disneyresearch.com

# Language Technologies Institute, Carnegie Mellon University
Gates-Hilman Complex, 5000 Forbes Ave., Pittsburgh, PA 15213, USA
3johnmcd@cs.cmu.edu, ∗bhiksha@cs.cmu.edu

ABSTRACT

While single-speaker voice activity detection is a well-studied problem, multi-speaker voice activity detection (MSV AD) for distant speech recognition remains a challenging task. In this work, we propose a new MSV AD system for identifying voice activity of an individual speaker from distant speech data captured with a microphone array. In contrast to normal energy-based approaches, our MSV AD algorithm employs information from the interfering channels in a hierarchical manner in order to adaptively adjust the threshold. We demonstrate the effectiveness of our MSV AD algorithm through experiments on the Speech Separation Challenge corpus [1]. A MSV AD technique with the cross-meeting normalized energy criterion [2] provided a missed detection rate (MDR) of 7.4% with a false alarm rate (FAR) of 28.0%. By incorporating the proposed criterion in the algorithm, the MDR and FAR were further reduced to 4.3% and 19.6%, respectively. Our algorithm also achieved speech recognition performance comparable to manual segmentation results. Moreover, our method requires no parameter training and has low computational complexity.

Index Terms— voice activity detection, microphone array, beamforming, speech separation, distant speech recognition

1. INTRODUCTION

Segmenting far-field audio signals into speech and non-speech parts is an essential pre-processing step for distant speech recognition (DSR) [3]. Such a technique is particularly important for interactive speech applications where a push-to-talk button or similar device cannot be used to delimit the beginning and end of an utterance.

A great deal of research effort has been devoted to the problem of voice activity detection (VAD) in the case of a single speaker. This problem is well-addressed through a couple of features such as energy features [4, 5, 6, 7], zero-crossing rates [8], distance measures of the cepstral features [9] and a degree of non-stationary and uncertainty [10].

However, when multiple speakers are simultaneously talking, identifying voice activity of an individual speaker still remains challenging due to overlapping speech as well as background noise. A speech signal from an interfering speaker leaks into a target channel, causing the crosstalk effect. In such a case, it is insufficient for speech recognition to merely detect whether speech is present on any given channel or not. We would rather detect voice activity of a speaker of interest and ignore speech from the other speakers.

Conventional VAD algorithms on multi-channel speech use cross-correlation based features between channels [2, 11, 12]. These features have been shown to be effective in detecting cross-talk and reducing false alarms caused by undesired speakers for close-talking microphone data.

In this work, we consider the multi-speaker VAD problem not in the context of several speakers wearing body-mounted microphones [13] but in the situation where speech of several speakers is captured with a microphone array and processed with beamforming [3, §13]. The purpose of beamforming is to enhance the speech signal of each desired speaker, while suppressing noise, reverberation and the other interference signals. After beamforming, the target signal is emphasized and interference signals are attenuated. As a result, the interfering speaker’s speech is overwhelmed by background noise, which makes cross-correlation-based features unsuitable for MSVAD systems.

Accordingly, we develop a novel feature to measure a degree of existence of target speaker’s speech in each channel. This feature incorporates the relative energy relationship between the target and interference signals obtained with beamformers as well as normalized absolute total energy from all the beamformers. More specifically, the energy ratio of the target signal to the total interference signals is compared with the background noise term normalized with the total energy from all the beamformers.

Figure 1 shows the configuration of our complete system. The speakers’ positions are first estimated from the microphone array
Given a position estimate for each speaker $i$, the quiescent weight vector such that it points at the target speaker. The blocking matrix is orthogonal to the channel of the active weight vector; see the details in [3, §10.2]. Overlapping speech from multiple speakers is then separated by minimum mutual information (MMI) beamforming and further enhanced by Zelinski post-filtering [14]. The MSV AD algorithm is performed on the beamformed data. Our MSV AD algorithm is multi-layered. First, we apply a standard energy-based method to beamformer’s output in order to discriminate speech regions from non-speech ones. Then, we detect cross-talk with the cross-meeting normalized energy [2]. Furthermore, voice activity is examined based on the new criterion described in section 2.3.

The balance of this paper is organized as follows. In section 2, we describe our complete system which includes the beamforming algorithm and the MSVAD system. Section 3 describes the corpus on which we conducted MSVAD experiments. MSVAD and speech recognition results are also presented in section 3. Finally, we conclude this work in section 4.

2. PROPOSED MSVAD SYSTEM

2.1. Beamforming preprocessing

Let us assume that there are two active speakers and that their speech is captured with a microphone array. As shown in Figure 2, we construct two beamformers in generalized sidelobe canceller (GSC) configuration in the subband domain for each active speaker. The subband GSC beamformer consists of the quiescent weight vector $w_{q,i}$, the blocking matrix $B_i$, and the active weight vector $w_{a,i}$, where $i$ is an index over subbands. The beamformer’s output for an input snapshot $X$ is then expressed as $Y_i = (w_{q,i} - B_i w_{a,i})^H X$. Given a position estimate for each speaker $i$, the quiescent weight vector is computed to provide unity gain for the look direction; i.e., it points at the target speaker. The blocking matrix is orthogonal to the quiescent weight vector such that $w_{q,i}^H B = 0$. By doing so, the distortionless constraint for the look direction is maintained. The active weight vector is typically adjusted so as to achieve the minimum distortionless constraint for the look direction is maintained. The active weight vector is typically adjusted so as to achieve the minimum distortionless constraint for the look direction.

\[
Y_i = (w_{q,i} - B_i w_{a,i})^H X
\]

2.2. Cross-meeting normalized energy

The cross-meeting normalized energy was introduced in [2] to detect cross-talk in individual headset microphones (IHMs). It incorporates the energy of all the $K$ IHM channels and is defined as follows:

\[
E_i^{\text{norm}}[n] = \frac{E_i[n]}{\sum_{k=1}^{K} E_k[n]}
\]

where $E_i[n]$ is the signal energy for a target channel $i$ at a frame $n$. It measures the target channel’s energy relative to the total energy across all the channels which is also useful for overlapping speech detection. Notice that the channel in (1) corresponds to each beamformer in our development. Our MSV AD system marks a current frame as voice activity if $E_i^{\text{norm}}[n]$ is more than $1/K$ for the target channel.

2.3. Target Speech Presence Score

When multiple speakers are talking, the average output power of the beamformers for the interfering speakers often exceeds that of the beamformer pointing at the target speaker. Thus, the normalized energy of the target speaker (1) would be below a threshold. In this case, voice activity will be missed with the normalized energy measure only. The voice decision criterion should incorporate the relative energy relationship between the energy of target and the interfering speakers as well as the absolute total energy from all the beamformers. In this work, we propose the target speech presence score (TSPS) to incorporate both relative and absolute energy information, which is defined as follows:

\[
\text{TSPS}_i[n] \triangleq \log \left( \frac{E_k[n]}{\sum_{k \neq i} E_k[n]} \right) - \log \left( \frac{E_0}{\sum_{k=1}^{K} E_k[n]} \right)
\]

$E_i$ is the energy of the target signal obtained with the beamformer and $E_k(k \neq i)$ is the energy of the interference speech signal from the other beamformers. $E_0$ is a constant reflecting the static reference energy level of the recording system. This TSPS measures a degree of the presence of target speaker’s speech given the conditional information of interfering channels’ speech. In our system, TSPS is only calculated when $E_i^{\text{norm}} \leq 1/K$. When the energy of the target signal is lower than the average energy across all the beamformers’ outputs, the interference speakers are possibly more dominant than the target speaker. In that case, we can more accurately detect voice activity of the target speaker by comparing the energy of the target signal normalized by the total energy of the interference signals only with the total energy over all the beamformers based on (2).

2.4. System framework

Given the above definitions, we can plausibly adopt the following strategy for determining when speech from a competing speaker is present on the signal from the desired speaker’s beam.

The individual steps leading to a decision are the following:

**Step1** Compare the short-time signal energy in the desired beam with the noise threshold. If the signal energy is lower than the threshold, decide no speech activity, otherwise go to next step. (Energy threshold is set to top $70\%$ of previous unvoiced frame energy distribution, where the range of $K$ is from 70 to 99 for obtaining the Receiver Operating Characteristic (ROC) curve.)

**Step2** Calculate cross-meeting normalized energy. If $E_i^{\text{norm}} > \frac{1}{K}$, decide speech activity, otherwise go to next step.

**Fig. 2.** Generalized sidelobe cancelling (GSC) beamformers.
Step 3 If $TSPS > 0$, decide speech activity, otherwise decide no speech activity.

We also use the post-processing techniques to further smooth the segmentation results. A segment of 0.25 seconds long is padded at the start of each speech segments since it is difficult to detect the exact speech onset. In the second step, two speech segments with less than 0.3 seconds gap in between are merged.

3. EXPERIMENTS

In this section, we present MSVAD and speech recognition results obtained with the MSVAD systems described in section 2.

3.1. Experimental setup

All the experiments of this paper were conducted on the PASCAL Speech Separation Challenge (SSC) data collected by the Augmented Multi-party Interaction (AMI) project [1]. The data set contains recordings of five pairs of speakers where each pair of speakers reads approximately 30 sentences taken from the 5000-word vocabulary of the Wall Street Journal (WSJ) task. There are a total of 43.9 minutes of speech in the development set and a total of 11,598 word tokens in the reference transcriptions. The data from two simultaneously active speakers were recorded with two circular, eight-channel microphone arrays. The diameter of each array was 20 cm, and the sampling rate of the recordings was 16 kHz.

Front-end MSVAD feature extraction is implemented using the following settings: window length 16 milliseconds, window shift 10 milliseconds and pre-emphasis factor 0.97. The feature extraction used for the ASR experiments reported here was based on cepstral features estimated with a warped minimum variance distortionless response (MVDR) spectral envelope of model order 30 [3]. Front-end analysis involved extracting 20 cepstral coefficients per frame of speech. Features are transformed with cepstral mean normalization (CMN), linear discriminant analysis (LDA). HMM training was conducted along the lines suggested in [16].

3.2. Segmentation results

Table 1 shows the list of MSVAD systems used in our experiments. In addition to the minimum mutual information beamformer (denoted as $\text{MMIBF}$), the conventional delay-and-sum beamformer (denoted as $\text{DSBF}$) is also considered for investigation. System M0 applies the proposed MSVAD system to the $\text{DSBF}$ output. System M1 only uses the first step of the proposed MSVAD system for the $\text{MMIBF}$ output. In system M2, the second step of the proposed MSVAD system is performed in addition to system M1. In system M3, all the MSVAD features are applied to the $\text{MMIBF}$ output. For comparison, we perform experiments with the cross-correlation feature (denoted as $\text{Corr}$) described in [2, 11] instead of the TSPS feature (referred to as system M4).

To evaluate the performance of our MSVAD system, we calculated the percentage of frames correctly recognized as speech (True Positive) and falsely recognized as speech (False Alarm) relative to the total number of frames. The receiver operating characteristic (ROC) curves of MSVAD systems are shown in Figure 3.

It is clear from Figure 3, system M1 provided the worst MSVAD performance because it uses the single channel only. This is not surprising since there are still crosstalk effects in the target channel although the beamformer managed to suppress interfering speakers speech. System M2 incorporates energy information from all the channels and improved the detection performance significantly. Our proposed system M3 uses the TSPS as an additional step and to help reject false hypotheses. It is also clear from Figure 3 that our MSVAD system can further improve MSVAD performance compared to system M2. It is worth noting that system M0 which uses the delay-and-sum beamformer performs slightly worse than our proposed system M3. These results indicate that our proposed MSVAD feature is not affected by the beamforming algorithms very much although it improves detection performance on data processed with the MMI beamformer. The poor performance of system M4 confirms that the cross-correlation based features are not suitable for multiple-speaker MSVAD of microphone array recordings. We were led to conclude that MSVAD performance by the cross-correlation feature becomes insignificant because the outputs of the beamformers are uncorrelated well.

The proposed MSVAD system M3 finishes segmentation of 1 second of speech in 0.06 seconds on average, as measured on a 3.2GHz Intel Xeon machine. Our proposed MSVAD system is computationally efficient and suitable for real-time processing.

3.3. Parameter selection

As mentioned in section 2.3, the TSPS (2) requires a constant $E_0$. The value of $E_0$ is crucial for our proposed MSVAD system. The MSVAD performance of system M3 given various $E_0$ values are tabulated in Table 2. We use $\log(E_0) = 14$ throughout this paper because it provided the lowest frame error rate.

3.4. Speech recognition results

ASR experiments were performed on segments generated by MSVAD systems M0-M3 and manual labeling. ASR results described here
are obtained with the best parameter setting of each system. Results are shown in Table 3. It is clear from Table 3 that the baseline system M1 provides the word error rate (WER) of 48.7% which is worse than the manual segmentation result. It can be considered that the ASR performance of system M1 does not degrade drastically compared to manual segmentation. This could be partially attributed to the contribution of the beamformers, which suppressed interfering speaker’s speech. We can also see from Table 3 that system M2 with the cross-meeting normalized energy can reduce WER by 5.1% compared to system M1. By adding the TSPS feature to system M3, the WER was further reduced and almost the same recognition performance as manual segmentation was achieved. These results indicate that incorporating energy information from interfering channels improves MSVAD segmentation and ASR performance. Notice that our proposed MSVAD feature TSPS significantly improves recognition performance on the MMI beamformer as well as the conventional delay-and-sum beamformer.

Although the VAD performance of system M0 is slightly worse than that of system M3 as mentioned in Section 3.2, the difference of the WERs is relatively large. The WER of M0 is 8.3% worse than that of system M3. This is because the different beamforming techniques are employed in two systems. MMI beamforming provides better suppression performance than the delay-and-sum beamformer. Thus, ASR performance of system M0 still degrades significantly, albeit its VAD performance is comparable to system M3. Another ASR experiment was run on the MMI beamformed data using segmentation labels from system M0 and provides a WER of 41.2%. This result confirms our argument.

4. CONCLUSIONS

In this work, we developed a computationally efficient VAD system for multiple speakers. Our system incorporates energy information from the interfering channels to adaptively adjust the decision threshold of the target channel. We showed through the experiments on the real data that adding the proposed feature significantly improves VAD performance. ASR results show that the WER of our VAD system is close to manual segmentation.

5. REFERENCES