Speaker-Dependent Voice Activity Detection Robust to Background Speech Noise

Shigeki Matsuda$^1$, Naoya Ito$^2$, Kosuke Tsujino$^3$, Hideki Kashioka$^1$, and Shigeki Sagayama$^2$

$^1$Spoken Language Communication Laboratory, National Institute of Information and Communication Technology (NICT), Kyoto, Japan
$^2$Graduate School of Information Science and Technology, University of Tokyo, Japan
$^3$Research laboratories, NTT DOCOMO Inc., Japan

shigeki.matsuda@nict.go.jp, nitou@hil.t.u-tokyo.ac.jp, tsujino@nttdocomo.co.jp, hideki.kashioka@nict.go.jp, sagayama@hil.t.u-tokyo.ac.jp

Abstract

In this paper, we proposed a speaker-dependent voice activity detection (VAD) algorithm that only extracts the speech period uttered by a target user. Based on our survey of the recognition error of real speech data collected in “VoiceTra,” which is a speech-to-speech translation system for smartphones, we found that many word insertion errors are caused by background speakers’ speech. Our VAD, which consists of three GMMs (noise and speaker-independent GMMs as used in a traditional GMM-based VAD and a speaker-adapted GMM) can be used for speech detection of the target speaker. In the VAD evaluations for the test utterances with background speakers’ speech, our proposed VADs achieved better performance than the conventional VAD. Also speech recognition experiments demonstrated that an ASR system with our proposed VAD achieved better performance than an ASR system using the conventional VAD.

Index Terms: voice activity detection, speech recognition

1. Introduction

Voice activity detection (VAD) [1], which locates speech segments from continuous speech utterances, is crucial in many speech application systems, such as speech enhancement systems, automatic speech recognition (ASR) systems, and hearing aid systems. Traditional VAD algorithms work well in clean or high signal to noise ratio (SNR) environments. However, in strong noisy environments, for example, the noisy environments of train stations and airports, the performance of most of them decreases sharply. To improve noise robustness, several noise robust acoustic analysis for VAD [2, 3, 4, 5], several noise suppression and speech enhancement algorithms [6, 7] have been proposed.

Compared with traditional noise, another type of noise, i.e., background speakers’ speech, is much more difficult to remove. In real applications, background speakers’ speech interference is very common. For example, in our NICT network-based speech translation application “VoiceTra” [9] for smartphones such as iPhones and android terminals, people often use their systems in noisy environments with many background speakers. Based on our survey of the recognition error of a huge collected speech corpus, we found that many word insertion errors are caused by background speakers’ speech. VAD must improve the noisy robustness of such a system. Although many noise reduction algorithms can be applied to improve VAD’s noise robustness, no VAD algorithm is robust to background speakers’ speech. The difficulty lies in that the background speakers’ speech usually has the same or similar statistical properties of the target speaker’s speech.

In this paper, we propose a speaker-dependent VAD algorithm that only extracts the speech period uttered by a target user. Our proposed VAD is based on a Gaussian mixture model (GMM). Different from a conventional GMM-based VAD that applies both noise and speaker-independent GMMs to detect speech periods based on their acoustic probability scores, our proposed VAD uses an additional GMM adapted to the target user. Based on the three GMMs (noise and speaker-independent GMMs as used in traditional GMM-based VAD and a speaker-adapted GMM), our VAD can be easily used for speech detection of target speakers. In real applications, it is easy to collect a speech data corpus for estimating a speaker-adapted GMM, since we assume that our proposed VAD is implemented to a smartphone used by a specific user.

The remainder of this paper is organized as follows. In Section 2, we describe the details of our proposed speaker-dependent VAD algorithm. In Section 3, we evaluate the proposed VAD on speech detection and recognition experiments and compare the results with a baseline performance. Discussions and a conclusion are given in Section 4.
2. Speaker-Dependent Voice Activity Detection

Our proposed speaker-dependent VAD uses three GMMs: a noise GMM, a speaker-independent GMM, and a speaker-adapted GMM. The conventional GMM-based VAD uses only two GMMs, the noise GMM and the speaker-independent GMM. In equation (1), speech or non-speech frames are discriminated by comparing two acoustic probability scores calculated from these two GMMs:

\[ P(x_t|\lambda_{\text{noise}}) < \alpha P(x_t|\lambda_{\text{ubm}}), \] (1)

where \( x_t \) is an acoustic feature vector at time \( t \) and \( \lambda_{\text{noise}} \) and \( \lambda_{\text{ubm}} \) are the model parameters of noise and speaker-independent GMMs. \( \alpha \) represents a balance between the noise and the speaker-independent GMMs.

The proposed technique has a speaker-adapted GMM \( \lambda_{\text{user}} \) estimated by a speaker-adaptation technique in addition to the two GMMs used in the conventional VAD. The speech or non-speech frames are discriminated by comparing the acoustic score as in equation (2):

\[ P(x_t|\lambda_{\text{noise}}) + P(x_t|\lambda_{\text{ubm}}) < \alpha P(x_t|\lambda_{\text{user}}). \] (2)

The speaker-independent GMM calculates a higher acoustic score for speech in the noise than the speaker-adapted GMM, and therefore a period of noise that contains background speakers’ speech is judged as non-speech frames. On the other hand, the speaker-adapted GMM calculates a higher acoustic score for the user’s speech than the noise and speaker-independent GMMs.

### Table 1: Noise types used in experiments

<table>
<thead>
<tr>
<th>For training</th>
<th>For testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Platform at train station</td>
<td>High-speed railway</td>
</tr>
<tr>
<td>Airport lobby</td>
<td>Public bus</td>
</tr>
<tr>
<td>Concourse of a station</td>
<td>Food department</td>
</tr>
<tr>
<td>Boiler room</td>
<td>Driving</td>
</tr>
<tr>
<td>Auction market</td>
<td>Elevator hall</td>
</tr>
<tr>
<td>Health club</td>
<td>Supermarket checkout</td>
</tr>
<tr>
<td>Hotel lobby</td>
<td>Front desk</td>
</tr>
<tr>
<td>Pub</td>
<td></td>
</tr>
</tbody>
</table>

The speaker-adapted GMM is adapted from the speaker-independent GMM using maximum a posteriori adaptation (MAP) [10] and a multi-class maximum likelihood linear regression (MLLR) [8]. In Section 3, we compare these two adaptation techniques on VAD evaluation.

#### 2.1. Hangover processing

The detected speech periods are fragmented and only applied to the acoustic score comparison as described above. To combine into a period of a lump of speech, hangover processing is applied (Fig. 1):

- **Step 1** To remove fragmentary speech frames, if the number of contiguous speech frames is less than \( C \), those frames are set as non-speech frames.
- **Step 2** To fill short pauses in an utterance, if the number of non-speech frames between speech frames is less than \( G \), the non-speech frames are changed to speech frames.
- **Step 3** To remove too short speech periods, if the number of contiguous speech frames is less than \( M \), those periods are set as non-speech frames.
- **Step 4** \( R \) contiguous speech frames are added to the beginning and ending of the speech period.

### 3. Experiments

#### 3.1. VAD evaluation

We evaluated the VAD performance of our proposed speaker-dependent VAD algorithm. The model parameters of speaker-independent GMM were estimated using five hours of dialogue speech from the ATR travel arrangement task database (TRA) and 25 hours of read speech of phonetically balanced sentences (TRA-BLA). The number of speakers in the training data was about 400 with about 26 k sentences. The training data were contaminated with the 15 types of noises listed in Table 1.
at four types of SNRs: 30, 25, 20, and 15 dB. For testing, we used the basic travel expression corpus (BTEC) with 20 male speakers and 20 female speakers; each speaker uttered about 300 sentences. We used 200 sentences for each speaker for speaker adaptation from the speaker-independent GMM. Answer labels used on VAD evaluations were generated by Viterbi algorithm using phoneme sequences of transcriptions. Since we assume that our proposed VAD is implemented to a smartphone used by a specific user, it is easy to collect users’ speech. These adaptation sentences were also contaminated with the same noises and SNRs as the above speaker-independent GMM training. The remaining 100 sentences, which were used for testing, were contaminated with the five types of noises listed in Table 1 at four types of SNRs: 30, 25, 20, and 15 dB. The scheduling task from the ATR Spontaneous Speech Database (APP) was used as the background speakers’ speech, which was added to the test sentences at -12 dB. -12 dB means that the distance from a microphone to a background speaker was 20 cm when the distance from the microphone to a target user was 5 cm. The acoustic parameters consisted of 12 MFCCs, 12 ∆MFCCs, and ∆pow extracted from 20-ms long frames with 10-ms frame shifts. The number of Gaussians in the speaker-independent GMM and the speaker-adapted GMM obtained by speaker adaptation was 2,048 and number of Gaussians in the noise GMM was 128.

We tested two different speaker-adaptation techniques: 1) MAP adaptation of mean vectors and 2) multi-class MLLR adaptation. Weighting parameter τ shown in equation (3) for the MAP adaptation was set to 10.0 based on preliminary experiments.

\[ \hat{\mu} = \frac{N \mu}{N + \tau} + \frac{\tau \mu}{N + \tau} \]  \hspace{1cm} (3)

where \( \hat{\mu} \) and \( \mu \) are the adapted mean vector and the mean vector calculated from the adaptation data. \( \mu \) is a mean vector in the speaker-independent GMM. \( N \) is occupation count of the adaptation data and \( \tau \) is a weighting of a priori knowledge to the adaptation data.

The number of classes and occupation thresholds used in the MLLR adaptation were set to 64 and 5000.0, respectively. The balance parameter \( \alpha \) in equation (2) was set to 1.4, and the parameters used in the hangover processing were set to \( C = 3, G = 30, M = 20, \) and \( R = 10 \) from preliminary experiments.

Figure 2 shows the experimental results. A conventional VAD that consisted of noise and speaker-independent GMMS which is used as the baseline in this study only achieved the best performances without background speakers’ speech (red solid lines) in all test sets with different SNRs.
Figure 3: Experimental results of ASR performance

However, the baseline VAD performance was heavily degraded when there is background speakers’ speech (red dot lines). Especially for relatively clean test utterances such as 30 dB, the conventional VAD degraded the performance. On the other hand, our proposed VADs using speaker-adapted GMMs estimated by MAP or MLLR adaptation outperformed the conventional VAD when there is background speakers’ speech (blue and green dotted lines). In addition, in the proposed VADs, when the MLLR adaptation (blue dot lines) is used, better performance was obtained than that of using the MAP adaptation (green dot lines). The reason may due to that the MAP adaptation does not estimate all Gaussian distributions when the amount of adaptation data is limited. Since acoustic scores of Gaussian distributions that were not adapted are the same as a speaker-independent GMM, it is difficult to discriminate user’s speech from background speakers’ speech.

3.2. ASR evaluation

We evaluated the recognition performance of an ASR system with our proposed VAD using a speaker-adapted GMM estimated by MLLR adaptation. In acoustic modeling, a state-tying structure with 5670 states was generated using the MDL-SSS technique [11] where each state had 10 Gaussian components. This acoustic model was trained using the ATR travel arrangement task database (TRA), the ATR Spontaneous Speech Database (APP), and the read speech of phonetically balanced sentences (TRA-BLA and APP-BLA). The total number of utterances was about 170 k.

Word bi-gram and composite word tri-gram language models [12] are used in our ASR system. And the language model are trained from the spontaneous speech database (SDB), a language database (LDB), and a spoken language database (SLDB) and the total number of words is 6.1 M and the lexicon size is 34 k.

Experimental results are shown in Figure 3. From this figure, we can see that when there was no background speakers’ speech, our proposed VAD achieved the same performance as an ASR system with a conventional VAD. Experiments using test utterances with background speakers’ speech demonstrated that our proposed VAD effectively reduced the errors of an ASR system than using the conventional VAD. Especially the our ASR system effectively reduced word insertion errors compared with the baseline ASR system.

4. Conclusion

In this paper, we proposed a speaker-dependent VAD algorithm that only extracted the speech period uttered by a target user. In the VAD performance evaluation for the test utterances with background speakers’ speech, our proposed VADs achieved better performance than the conventional VAD. ASR experiments using test utterances with background speakers’ speech also demonstrated that our proposed VAD effectively reduced the errors of an ASR system using a conventional VAD. In future work, we will evaluate the performance of an ASR system using our proposed VAD for real speech data collected in the “VoiceTra” service.

5. References