Robust Whisper Activity Detection Using Long-Term Log Energy Variation of Sub-Band Signal

G. Nisha Meenakshi and Prasanta Kumar Ghosh, Senior Member, IEEE

Abstract—The goal in the whisper activity detection (WAD) is to find the whispered speech segments in a given noisy recording of whispered speech. Since whispering lacks the periodic glottal excitation, it resembles an unvoiced speech. This noise-like nature of the whispered speech makes WAD a more challenging task compared to a typical voice activity detection (VAD) problem. In this paper, we propose a feature based on the long term variation of the logarithm of the short-time sub-band signal energy for WAD. We also propose an automatic sub-band selection algorithm to maximally discriminate noisy whisper from noise. Experiments with eight noise types in four different signal-to-noise ratio (SNR) conditions show that, for most of the noises, the performance of the proposed WAD scheme is significantly better than that of the existing VAD schemes and whisper detection schemes when used for WAD.

Index Terms—Long-term signal measure, sub-band selection, whisper activity detection, whispered speech.

I. INTRODUCTION

WHISPER ACTIVITY DETECTION (WAD) is the task of identifying the whispered speech regions in a given noisy whispered speech recording. Whispered speech is one of the natural speech modes where the vocal folds do not vibrate and therefore lacks the periodicity and the pitch [1]–[3]. In addition to the loss of voiced excitation, whispered speech differs from the neutral speech by the shift of formants at lower frequencies [4], [5]. As observed by Ito et al., [6], a whispered speech sounds more like an unvoiced speech [2]. In a typical voice activity detection (VAD) problem, it is known that detecting unvoiced speech in the presence of noise is more challenging than detecting voiced speech [7], [8]. Additional features are often used to distinguish the unvoiced regions from the noise [9]. Thus, detecting whispered speech from a noisy recording becomes a challenging task.

Features used in the typical VAD algorithms include short-time energy [10]–[12] and zero-crossing rate [13]. Cepstral features are also used [8]. Apart from the short-term analysis, long-term speech characteristics have also been used, including long-term spectral divergence [14], full-band and multi-band long-term signal variability (LTSV) [15], [16] and KL divergence measure [17].

Whispered speech offers a rich source of research issues including speaker identification [18]–[20], speech recognition [6], [21], [22] and reconstruction of the neutral speech from the whispered speech [3], [23]. Works related to the identification of a test utterance as whispered or neutral speech [24] and detection of whisper-islands [25], [26] are also considered. Although different types of problems are addressed based on the whispered speech, WAD has received little attention. In the work by Sharifzadeh et al. [23], WAD is done as a preprocessing step, using a modified G.729 voice activity detector [27] where the short-term signal power and zero-crossing rate are used with adaptive thresholds.

In this paper, we propose a new signal characteristic—long-term logarithmic energy variation (LTCV) of the signal from a set of optimally chosen sub-bands, for WAD. Although whispering resembles unvoiced sound, the envelope of the whispered speech varies over time. We exploit this fact to discriminate noisy whisper from noise. The LTCV captures the degree of non-stationarity using the variation in the short-time energy profile. We also find that the difference in the variations in the short-time energy profile of the whispered speech and noise is more prominent in a few optimally selected sub-bands compared to the full-band signal. Hence, the LTCV is computed using the signal from the optimal sub-bands. We also propose a sub-band selection procedure in this regard.

II. DATASET

We use the CHAINS (Characterizing Individual Speaker) speech corpus [28] for the whispered speech. The CHAINS corpus consists of a total of 36 speakers (20 males and 16 females) comprising 3 different accents and 6 speaking conditions. We consider the 33 individual sentences from randomly chosen 10 speakers, 5 male \((f_{rm}01, f_{rm}02, f_{rm}03, f_{rm}04, f_{rm}01)\) and 5 female \((f_{rf}01, f_{rf}02, f_{rf}03, f_{rf}04, f_{rf}01)\) from the ‘Whispered Speech’ speaking condition. 24 of these sentences are taken from the TIMIT sentences and 9 from the CSLU Speaker Identification Corpus. Initially available at a sampling frequency of 44100 Hz, the data is downsampled to 16000 Hz. Thus, the total duration of the whispered speech considered for this work is \(~ 866.84~{\text{seconds}}\) (14.5 minutes).

The noise samples are taken from the NOISEX-92 corpus [29]. We choose stationary, non-stationary and speech-like noises to test the robustness of the LTCV based WAD algorithm. We use eight types of noises namely, White (WH), Pink (PK), Machine (MC), Factory (FR), Destroyerengine (DE), Babble (BB), Machinengun (MG) and High frequency (HF) noises. Noise data is also resampled to 16000 Hz. Noise is added to the whispered speech to generate the noisy whispered speech samples.
III. LTLEV

It is observed that the whispered speech has more short-time energy variation compared to that of various noises. Hence, LTLEV is used for detecting whisper in a given noisy whispered speech. Let the given speech signal of length $N$ samples with a sampling frequency $F_s$ be represented as $x[n], 0 \leq n \leq N - 1$. Let the frequency range (0 to $F_s/2$) be divided into $K$ sub-bands and let $x_k[n]$ be the $k$th sub-band signal of the given audio stream $x[n]$. Let us consider $S$ representing a set of sub-band indices. Then, the signal constructed from the sub-bands specified in $S$ is given by, $x_S[n] = \sum_{k \in S} x_k[n], 0 \leq n \leq N - 1$. Note that, when $S = \{1, 2, \cdots, K\}, x_S[n] = x[n]$. The logarithm of the short-time energy of the $m$th frame of $x_S[n]$ is calculated over a window of duration $N_w$ samples and shift $N_s$ samples as follows:

$$E_S[m] = \log \left( \frac{1}{N_w} \sum_{n=m-1}^{m+M-1} (x_S[n])^2 \right).$$  

(1)

LTLEV is computed as the variance of the $E_S[m]$ over a long term window comprising $M$ frames starting from the $m$th frame.

$$LTLEV[m] = \frac{1}{M} \sum_{m'=m}^{m+M-1} (E_S[m'] - \bar{E_S}[m'])^2$$  

(2)

where, $\bar{E_S}[m] = \frac{1}{M} \sum_{m'=m}^{m+M-1} E_S[m']$. Logarithm of the short-time energy in eqn (1) removes the effect of the amplitude scaling of the signal on LTLEV (computed using eqn (2)). In other words, an amplification of the signal does not change the LTLEV values. This causes LTLEV of the noise to be SNR independent. It is to be noted that the number of long-term frames $M$ and the set of sub-bands $S$ are two key parameters in computing LTLEV.

IV. SUB BAND SELECTION IN LTLEV

We describe the procedure to select the optimal set $S^*$ of sub-bands that best discriminates the noisy whisper from the noise. The sub-band selection is done for a fixed value of $M$. Let the whisper activity labels on a noisy whispered dataset be denoted by $WAD_L$.

The goal is to find the $S^*$ which gives the maximum whisper-noise classification accuracy on a given dataset. To find the optimal sub-bands, we follow a forward sub-band selection procedure as summarized in Algorithm 1. For a chosen set of sub-bands, the threshold $\eta$ is found by generating a Receiver operating characteristic (ROC) curve and finding the equal error rate (EER denoted by $E_e$) using the LTLEV feature and the corresponding $WAD_L$. This is repeated in each iteration to minimize the EER by selecting the best sub-band.

Algorithm 1 Forward Sub band selection

1: Inputs: $M, K, x[n], WAD_L$
2: Initialization: $S = [], ME = [], \eta = [], S^* = [1, 2, \ldots, K]$
3: for $i = 1$ to $K$ do
4: \hspace{1em} $E_e = []; \eta = []$
5: \hspace{1em} for $j \in S^*$ do
6: \hspace{2em} $S = [S, j]$
7: \hspace{2em} Obtain $LTLEV$ using eqn (1), (2)
8: \hspace{2em} $[E_e, \eta] \leftarrow LTLEV, WAD_L$ (using ROC)
9: \hspace{1em} end for
10: $j^* = \arg \min_j E_e$
11: $S_i \leftarrow j^*$ ($S, i^{th}$ element of the vector $S$)
12: $S^* \leftarrow S \setminus \{j^*\}$, $\bar{\eta} \leftarrow \bar{\eta} - E_e, ME_i \leftarrow E_e$
13: end for
14: $i^* = \arg \min_i ME_i$, $S^* \leftarrow S_i, i = 1, \ldots, i^*$
15: $\eta^* \leftarrow \bar{\eta}$, $ME^* \leftarrow ME_i$
16: return $S^*, \eta^*, ME^*$

end

Fig. 1 shows the normalized histogram of the LTLEV of the noisy whispered speech and different noises at 0 dB SNR and $M = 50$ frames. The sub-bands (shown in Fig. 1) used to generate the histograms are obtained using the forward sub-band selection procedure for each noise separately. The classification accuracy (corresponding to EER) is also shown for each noise type.

We observe that for some noises, the overlap between the two LTLEV histograms is large. This is the case with (a) FR, (c) PK and (e) MC noises. On the other hand, there is very little overlap between the two histograms for (b) DE, (d) WH, (f) HF, (g) BB and (h) MG noises. The amount of overlap depends on the chosen sub-bands. The variation of the short-time energy profile could be different in different sub-bands. Also each sub-band may have an SNR different from the global SNR. Thus the choice of the sub-bands is crucial and needs to be done
in such a way that in the chosen bands there is more energy variation in the whispered speech compared to noise. Detailed investigation of the spectra of different noises shows that, on average, the spectrum of the whispered speech closely matches with that of the (a) FR, (c) PK and (e) MC noises. This could cause less discrimination between the two histograms for these noises. The classification accuracies for noises with minimal overlap between the two histograms are above 96% while those with higher overlap result in accuracies below 85%, as shown in Fig. 1.

V. EXPERIMENTS AND RESULTS

A. Experimental Setup

Dataset preparation: The whispered speech file for WAD is created by interleaving the sentences from CHAINS corpus with silence of duration ranging from 1.5 to 2 seconds. Each of the eight noise types are added separately to this silence-appended whispered speech signal at a given SNR. Four different SNR conditions are considered, namely, –5 dB, 0 dB, 5 dB, 10 dB. We perform WAD in a five-fold cross-validation setup with each fold having one male and one female speaker at the test set. Remaining four folds are used for training and development. Thus, there is no overlap between the subjects in the training and test sets. The average duration of the noisy whispered speech for each speaker pair is \( \sim 346,742 \) seconds (5.77 minutes). It is to be noted that, on average, \( \sim 64\% \) of a test utterance corresponds to the noisy whispered speech.

Parameter selection: The training set is used to learn the optimal parameters \( M^*, S^* \) and \( \eta^* \) by minimizing the EER. We use a \( N_w = 320 \) (20 ms) and \( N_{sl} = 160 \) (10 ms) for the short-time energy calculation. We choose different durations of long-term window, namely \( M = 10, 30, \) and \( 50 \), from which \( M^* \) is selected. We use a uniform filter bank configuration with \( K = 32 \). Chebyshev II IIR digital filters, with pass band ripple of 1 dB and stop band attenuation of 60 dB, are used.

The final WAD decision is made for every 10 ms of the noisy whispered speech. Let there exist, \( P \) such 10 ms blocks in a noisy whispered utterance. The WAD at the \( p \)th block, \( 1 \leq p \leq P \), is decided using the class labels of all \( M \) long-term windows, which overlap with the \( p \)th block. If \( c \% \) of these \( M \) labels are whispered utterance, then we determine the \( p \)th block to be the whispered speech otherwise noise. We choose different values of \( c = [10, 20, \ldots, 90] \) and find the optimal \( c^* \) by maximizing the accuracy of the WAD on the development set. We also post-process the WAD decisions with a hangover scheme as used by Davis et al. [30], but there was no significant improvement in the WAD performance. Hence, all results reported in this work are obtained without any hangover scheme.

Performance metrics: To quantify the performance of the proposed WAD algorithm, five performance metrics are used, viz., WAD accuracy (CWAD), front end clipping (FEC), mid speech clipping (MSC), carry over (OVER) and noise detected as speech (NDS) following the VAD work by Beritelli et al. and Freeman et al. [31], [32]. We would typically expect a good WAD algorithm to have lower values for FEC, MSC, OVER and NDS and higher CWAD.

Baseline schemes: We compare the performance of the proposed algorithm, with four existing schemes- two from the VAD literature, namely the LTSV [15] and G.729 Voice Activity Detector [27], and two from the whispered speech detection literature, viz. auditory-inspired modulation spectrum (AIMS) features [24] and a four dimensional entropy based feature (AD Entropy) [25], that were proposed to discriminate the whispered and the neutral speech. The parameters in LTSV are optimized on the development set of each fold for the WAD task. We use the full-band LTSV [15] as multi-band LTSV [16] did not improve the WAD performance. We use the absolute value of the logarithm of the 61 dimensional AIMS feature as the WAD performance is found to improve compared to the original AIMS features.

B. Results and Discussion

In Fig. 2, we illustrate WAD using LTLEV in additive BB noise. We see that the lower frequencies are primarily dominated by the speech-like BB noise. At the higher frequencies, we observe a higher energy variation for the whispered speech regions compared to the noisy regions, which explains the choice of the optimal sub-band to be 30 as indicated in Fig. 1(g). From Fig. 2(c), we see that the noisy whispered speech yields higher value of LTLEV feature than the noise alone. Fig. 2(d) indicates that even at a low SNR condition such as –5 dB, most of the whispered speech segments are detected correctly, while noise regions are falsely detected as speech at the boundaries of the whispered segments and noise. This is mainly due to the long-term window used for computing LTLEV.

Optimal subband selection: Fig. 3 shows the histograms of the optimally selected sub-bands in case of the eight noises considered. Optimal set of sub-bands are combined from all folds of different SNRs for a noise type for generating the histogram. It is clear from Fig. 3 that the optimal sub-bands are different for different noises. It is also clear that the histograms are sparse, i.e., only a few sub-bands are selected for each noise category. This suggests that the information for discriminating whisper and noise are present only in a few sub-bands. For example, in the case of BB and MG noises, 30th and 29th sub-bands are consistently selected for all four SNRs and five folds indicating that energy variation in a single band is capable of distinguishing whispered speech from MG noise and BB noise as shown in Fig. 1. For FR, DE, PK, WH, and HF noises, 1st band is consistently

\[ \text{Fig. 2. (a) Whispered speech with BB noise added at } -5 \text{ dB SNR consisting of 3 whispered speech utterances- 'If it doesn’t matter who wins, why do we keep score?’, ‘Stop each car if its little’, ‘Play in the street up ahead.’, from speaker 02, with ground truth whisper activity labels shown in red (label } = 0 \text{ indicates noise) (b) Spectrogram of the noisy whispered speech with window 20 ms and overlap 10 ms, (c) log (LTLEV) with threshold } \eta, \text{ (d) WAD decisions (in black dashed line) with ground truth labels overlaid.} \]
Fig. 3. The histogram of the optimally chosen sub-band indices obtained using Algorithm 1 for eight different noises.

Fig. 4. CWAD for five schemes - LTLEV, LTSV, G.729, 4D-Entropy, AIMS - at four different SNRs and eight noisy conditions. The bars show the average CWAD with errorbars indicating one standard deviation (SD) across five folds.

Fig. 5. Detailed WAD performance measures for five schemes - LTLEV, LTSV, G.729, 4D-Entropy, AIMS - at 0 dB SNR. For numerical values visit http://www.ee.iisc.ernet.in/new/people/faculty/prasantg/softwares.html

selected for all SNRs and folds. It is to be noted that, for the PK noise, no band above the 5th one is selected. For the MC noise, the optimal set of sub-bands varies significantly across SNRs and folds. Interestingly there is a large overlap between the histograms of noisy whisper and MC noise as shown in Fig. 1(e). It should be noted that the $M^*$ turned out to be 50 for all types of noises in different SNRs and for different folds. This suggests that longer the window, better is the discrimination between the whisper and noise. We find that the $e^*$ increases with increase in SNR on the development sets, with the values ranging from 30-60% for FR and MC noises and 10-70%, 10-80%, 20-80%, 40-90%, 70-90%, 80-90% for PK, DE, HF, WH, BB and MG noises respectively.

WAD performance: The proposed LTLEV based WAD scheme is first compared with the four schemes in terms of the CWAD as shown in Fig. 3. It is clearly seen that the proposed LTLEV based WAD algorithm, has the highest CWAD value in all noise conditions and for all SNRs, except for the MC noise, where G.729 has comparable performance with that of the LTLEV. Even at a low SNR condition of $-5$ dB, we see an absolute improvement of 15.1%, 23.71%, 12.64%, 28.03%, 5.07%, 23.48%, 27.2%, 18.61% over the best baseline scheme for FR, DE, PK, WH, MC, HF, BB, MG noises respectively. This indicates the superiority of the proposed signal measure for the WAD task. At lower SNRs, lesser CWAD is observed for the LTLEV in cases of the PK, MC and FR noises compared to the other noises. This corresponds to the amount of overlap in the distributions as seen from the Fig. 1. Among the other four schemes considered for comparison, we see that the 4-D Entropy feature performs better in all SNRs of the MG and WH noises and in the higher SNR conditions of the BB noise while the G.729 performs better in all SNRs of the FR and MC noises and in the higher SNR conditions of the DE and PK noises.

VI. CONCLUSIONS

We propose a feature, namely LTLEV, that captures the long-term energy variation of sub-band signal for WAD. We also present a forward sub-band selection algorithm to select the optimal sub-bands to maximally discriminate the noise and noisy whisper. We find that the proposed LTLEV based WAD is superior to the four baseline schemes, namely G.729, LTSV, 4D Entropy and AIMS based algorithm. It is found that features proposed to discriminate neutral speech from the whispered speech, i.e. 4D-Entropy and AIMS, are not very successful to discriminate noisy whispered speech from noise. The better performance of LTLEV compared to LTSV suggests that the spectral variation is less robust in discriminating whisper from noise compared to the log-energy variation of the sub-band signal. Further investigation is required to understand the difference in the variability for a neutral and a whispered speech. These are parts of our future work.
REFERENCES


