Whisper activity detection using CNN-LSTM based attention pooling network trained for a speaker identification task

Abinay Reddy Naini\textsuperscript{1}, Malla Satyapriya\textsuperscript{2}, Prasanta Kumar Ghosh\textsuperscript{1}

\textsuperscript{1}SPIRE LAB, EE Dept., IISC, Bangalore, 560012, India
\textsuperscript{2}RGUKT, Kadapa, 516330, India

INTERSPEECH 2020
Table of contents

- Introduction
- Proposed approach
- Dataset
- Experiments and Results
- Summary
Table of contents

1. Introduction
2. Proposed approach
3. Dataset
4. Experiments and Results
5. Summary
Voice Activity Detection

- It is a process of detecting a speech region in a noisy audio file.¹
It is a process of detecting a speech region in a noisy audio file. An important pre-processing step in almost every speech application.
Whispered speech

Introduction

Whispered speech production:
- Private conversations
- Pathological situations

Pathological cases: Laryngectomy (Vocal folds removed surgically).

Challenges

Whispering is less intelligible compared to neutral speech.

Handling the absence of glottal vibration
Whispered speech production: private conversations, pathological situations.

Pathological cases: Laryngectomy (Vocal folds removed surgically).
Introduction

Whispered speech

- Whispered speech production: *private conversations, pathological situations*.
- Pathological cases: Laryngectomy (Vocal folds removed surgically).

Challenges

- Whispering is **less intelligible** compared to neutral speech.
- Handling the absence of glottal vibration

1 Tartter and Vivien C (1989), 2 Ito et al. (2005), 3 Sharifzadeh et al (2010)
Whispered speech

“This was easy for us.”
"This was easy for us."

Whispered speech differs significantly from natural speech.

1 Tartter and Vivien C (1989), 2 Ito et al. (2005)
Whispered speech differs significantly from natural speech.

1 Tartter and Vivien C (1989), 2 Ito et al. (2005)
"This was easy for us."

Whispered speech differs significantly from natural speech.

Lack of Pitch\(^1\).

Formant shift\(^2\).

1 Tartter and Vivien C (1989), 2 Ito et al. (2005)
Introduction

Clean vs Noisy spectrum

Typical SAD performs poorly in the Un-voiced region compared to the voiced region.

1. Presence of noise makes WAD a challenging task.

1 P. Renevey et al. (2001), 2 Nisha et al. (2015)
Typical SAD performs poorly in the Un-voiced region compared to the voiced region.¹

¹ P. Renevey et al. (2001), 2 Nisha et al. (2015)
Typical SAD performs poorly in the Un-voiced region compared to the voiced region.¹

Presence of noise makes WAD a challenging task.²

¹ P. Renevey et al. (2001)² Nisha et al. (2015)
Existing WAD method (Baseline scheme)

Using Long-Term Log Energy Variation of Sub-Band Signal (LTLEV)

- In this, the optimal sub-band sequence is obtained for each noise in the training phase.
- In the testing phase, these optimal sub-bands are used to filter the signal. Then the LTLEV of the filtered signal is used for WAD decision.

1 Nisha et al. (2015)
Existing WAD method (Baseline scheme)

Using Long-Term Log Energy Variation of Sub-Band Signal (LTLEV)\(^1\)

- In this, the optimal sub-band sequence is obtained for each noise in the training phase.
- In the testing phase, these optimal sub-bands are used to filter the signal. Then the LTLEV of the filtered signal is used for WAD decision.

Limitations of (LTLEV)\(^1\)

- It can only be used in seen noise conditions.
- Because of supervised training for obtaining optimal sub-bands, LTLEV requires annotated data indicating whispered speech regions.

---

\(^1\) Nisha et al. (2015)
Table of contents

1 Introduction

2 Proposed approach

3 Dataset

4 Experiments and Results

5 Summary
Proposed WAD using CNN-LSTM model (APW)

\[
\mathbf{K} = \sum_t \mathbf{w}_t \mathbf{f}_t
\]

Attention weights \(\mathbf{w}_t\) provide the whispered speech regions in an audio file.
Proposed WAD using CNN-LSTM model (APW)

\[ K = \sum_t w_t f_t \]  

Attention weights \( w_t \) provide the whispered speech regions in an audio file.
Proposed WAD using CNN-LSTM model (APW)

Training: This network is trained for the speaker identification task by optimizing the categorical cross-entropy loss.
Proposed approach

Proposed WAD post-processing scheme

Post-processing: we have used the weights predicted from LSTM block for WAD decision.
Proposed approach

LSTM attention weights processing

Weights $\{W_i\}$ (Nx1)

- Median Filter
- Binary Quantization with threshold $n_1$
- Moving Average Window

- Binary Quantization with threshold $n_2$
- Moving Average Window

- Binary Quantization with threshold $q_1$
- Summation

- Binary Quantization with threshold $q_2$

- WAD Decision

Graphical representation of the proposed approach with time series data over 25 seconds.
Table of contents

1 Introduction

2 Proposed approach

3 Dataset

4 Experiments and Results

5 Summary
Dataset

- **wSPIRE dataset** - 102 speakers each speaking 50 sentences in whisper mode. (Zoom and Moto-g5 mobile recording devices)

Dataset

- **wSPIRE dataset** - 102 speakers each speaking 50 sentences in whisper mode. (Zoom and Moto-g5 mobile recording devices)
- **wTIMIT dataset** - 48 speakers each speaking 450 sentences in whisper mode (50 sentences from each speaker is considered).
- **CHAINS dataset** - 36 speakers each speaking 37 sentences in whisper mode.

---

Dataset

- **wSPIRE dataset** - 102 speakers each speaking 50 sentences in whisper mode. (Zoom and Moto-g5 mobile recording devices)

- **wTIMIT dataset**\(^1\) - 48 speakers each speaking 450 sentences in whisper mode (50 sentences from each speaker is considered).

- **CHAINS dataset**\(^2\) - 36 speakers each speaking 37 sentences in whisper mode.

We considered **NOISEX-92**\(^3\) dataset, containing eight different noises for the experiments. Among these four types of noise used for training (pink, machine, factory, destroyerengine) and the other four (White, babble (SB), machinegun (MG), high freq (HF)) for testing.

---

# Table of contents

1. Introduction
2. Proposed approach
3. Dataset
4. Experiments and Results
5. Summary
Experiments and Results

Experimental setup

Exp1:
- We have used only clean speech data for training.
- Both the proposed method and baseline tested on four testing noise types.
Experiments and Results

Experimental setup

Exp1:
- We have used only clean speech data for training.
- Both the proposed method and baseline tested on four testing noise types.

Exp2:
- In this experiment, both the baseline and the proposed method is trained using four training noise types.
- Both the proposed method and baseline tested on four testing noise types.
Experiments and Results

Training data considered

Exp1:

- We have used 80 speakers from the wSPIRE database with 40 whispered speech utterances from each of them. (For training data recorded using only zoom recorder is considered.)
- Along with the above wSPIRE data, 40 whispered speech utterances from each of 48 speakers from the wTIMIT database are also used for the training.

**Remaining 10 whispered speech utterances from the wTIMIT is used for the training of post-processing block parameters.**
Experiments and Results

Training data considered

Exp1:
- We have used 80 speakers from the wSPIRE database with 40 whispered speech utterances from each of them. (For training data recorded using only zoom recorder is considered.)
- Along with the above wSPIRE data, 40 whispered speech utterances from each of 48 speakers from the wTIMIT database are also used for the training.

Remaining 10 whispered speech utterances from the wTIMIT is used for the training of post-processing block parameters.

Exp2:
- We have used the same data mentioned above for the Exp1.
- However, we augmented this data using the same data corrupted with four training noise types at -5dB, 0dB, 5dB, 10dB, 15dB, 25dB SNR.
Different test cases considered

<table>
<thead>
<tr>
<th>Test cases</th>
<th>Testing conditions and (Data)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1A</strong></td>
<td>Seen subject &amp; environment (remaining 10 utterances from the 80 training speakers of wSPIRE)</td>
</tr>
<tr>
<td><strong>1B</strong></td>
<td>Unseen subject &amp; seen environment (10 utterances from the remaining 22 speakers of wSPIRE)</td>
</tr>
<tr>
<td><strong>2</strong></td>
<td>Unseen subjects &amp; environment (10 utterances of 36 speakers from the CHAINS database)</td>
</tr>
<tr>
<td><strong>3</strong></td>
<td>Seen subjects &amp; unseen device (10 utterances of 80 training speakers from the wSPIRE recorded using Moto-g5)</td>
</tr>
<tr>
<td><strong>4</strong></td>
<td>Unseen subjects &amp; device (10 utterances of 22 speakers from the wSPIRE recorded using Moto-g5)</td>
</tr>
</tbody>
</table>

Table 1: Different testing conditions and the type of data used
In these experiments, we have used cumulative accuracy (ratio of the number of correctly classified frames to the total number of frames) to measure the performance, which is denoted as CWAD$^1$. 

1 Nisha et al. (2015)
We can observe from the fig that the proposed WAD is performing better than the baseline in all four noisy conditions.
Even though baseline performance is improved, the proposed method showed a significant improvement in almost all conditions.
Performance comparison of baseline and the proposed method
1 Introduction

2 Proposed approach

3 Dataset

4 Experiments and Results

5 Summary
Summary

Conclusions

—we showed that robust WAD can be achieved by Attention pooling based network, which is trained for speaker classification.
—we showed that the proposed method is robust towards the varying noise and recording conditions.
—Its ability to train on limited annotated data, unlike the existing methods.
Summary

Conclusions

- We showed that robust WAD can be achieved by Attention pooling based network, which is trained for speaker classification.
- We showed that the proposed method is robust towards the varying noise and recording conditions.
- Its ability to train on limited annotated data, unlike the existing methods.

Future Work

- We want to observe how the proposed method performs when trained for other speech tasks, such as gender classification.
Authors thank the Department of Science and Technology (DST), Govt of India for their support in this work.
THANK YOU

Have Questions/Suggestions?
Write to us at spirelab.ee@iisc.ac.in