Feature Selection for Labelling of Whispered Speech in ASMR Recordings using Edyson

Pablo Pérez Zarazaga, Zofia Malisz
Speech, Music & Hearing, KTH Royal Institute of Technology, Sweden
pablopz@kth.se, malisz@kth.se

Abstract

Whispered speech is a challenging area for traditional speech processing algorithms, as its properties differ from phonated speech and whispered data is not as easily available. A great amount of whispered speech recordings, however, can be found in the increasingly popular genre of ASMR in streaming platforms like Youtbe or Twitch. Whispered speech is used in this genre as a trigger to cause a relaxing sensation in the listener. Accurately separating whispered speech segments from other auditory triggers would provide a wide variety of whispered data, that could prove useful in improving the performance of data driven speech processing methods. We use Edyson as a labelling tool, with which a user can rapidly assign labels to long segments of audio using an interactive graphical interface. In this paper, we propose features that can improve the performance of Edyson with whispered speech and we analyse parameter configurations for different types of sounds. We find Edyson a useful tool for initial labelling of audio data extracted from ASMR recordings that can then be used in more complex models. Our proposed modifications provide a better sensibility for whispered speech, thus improving the performance of Edyson in the labelling of whispered segments.

Introduction

The growing popularity of speech applications and the larger availability of speech data have contributed to a surge in the development of data driven methods leading to continuing improvements in speech technology. However, the analysis of whispered speech still poses a challenging task for typical speech processing methods. Whispering is a natural mode of speech, characterized by features that considerably differ from normal phonated speech (Jovičić, 1998; Tartter, 1989). For instance, the most notable feature of whispered speech is the absence of pitch, produced by the lack of vibration of the vocal folds in the production of the voice. This causes a degradation in the performance of traditional speech processing methods when such features are not considered. Additionally, the limited availability of whispered speech recordings hinders the training of data driven methods. Several datasets exist that provide whispered speech data, such as wTIMIT (Lim, 2011) or wSPIRE (Singhal et al., 2021). However, recording a speech dataset is a costly process, and the size of these datasets is significantly smaller than their phonated speech counterparts. This is one of the reasons why few data driven methods are applied to whispered speech, and the existing ones rely on previously trained models for different applications (Naini et al., 2020). For that reason, the gathering of whispered data is a crucial task that will improve the performance of speech applications on these types of signals.

Autonomous sensory meridian response (ASMR) is a phenomenon in which a relaxing sensation is transmitted through visual and auditory triggers. This phenomenon has significantly grown in popularity over the last years in streaming platforms such as Youtube and Twitch (Andersen, 2015; Barratt & Davis, 2015; del Campo & Kehle, 2016; Smith et al., 2017). Among the auditory triggers used in ASMR content, whispered speech is one of the most popular types of recordings. Additionally, in order to enhance the transmitted perception of intimacy that users find relaxing, ASMR is usually recorded using high sensitivity microphones. This provides a vast number of whispered recordings that can be exploited to improve speech processing methods. The whispered segments, however, are usually interleaved with several other types of noise that make up other triggers: lip and tongue smacking, tinkering with objects, blowing air into the microphone etc. It is, therefore, necessary to accurately separate the whispered segments that could be used in speech applications from the noise samples that would degrade their performance.

Labelling of Whispered Speech

Edyson is a tool that allows for automatic labelling of audio data with human-in-the-loop (Fallgren & Edlund, 2021). This tool relies on several dimensionality reduction techniques to represent the input audio features in a two-dimensional space. The data is then presented in an interactive interface, where the user can easily analyse and assign labels to segments of the audio signal. This allows for accurate labelling of large amounts of data in a fraction of the time required if the task were to be performed manually. The selected input features in Edyson are Mel-frequency cepstral coefficient (MFCC) and the linear spectrogram of the audio segments, which perform well distinguishing normal phonated speech from noise signals. However, the representation of whispered speech is closer to that of noise, hence making it difficult to separate whispered segments from pure or unwanted noise.

In this demo, we propose several features that, in combination with the current inputs to Edyson, can improve the sensibility of the tool towards whispered speech. ASMR recordings containing mixtures of whispered speech and noise-like triggers were selected from Youtube, and the performance of the proposed methods is compared using phonated and whispered speech segments. Additionally, we analyse different parameter configurations that will adapt the performance of Edyson distinguishing different types of noise. In conclusion, we find Edyson a useful tool for initial labelling of large quantities of audio data, which can then be used in the training of more complex models of whispered speech. This proves especially useful considering the growing amount of unlabelled whispered data that is publicly available online. This will help us gain a better understanding of this area of speech and motivate the development of methods applied to whispered speech and whispered speech applications.

References

Andersen, J. (2015). Now you’ve got the shiveries: Affect, intimacy, and the ASMR whisper community. *Television & New Media*, *16*(8), 683–700.

Barratt, E. L., & Davis, N. J. (2015). Autonomous Sensory Meridian Response (ASMR): a flow-like mental state. *PeerJ*, *3*, e851.

del Campo, M. A., & Kehle, T. J. (2016). Autonomous sensory meridian response (ASMR) and frisson: Mindfully induced sensory phenomena that promote happiness. *International Journal of School & Educational Psychology*, *4*(2), 99–105.

Fallgren, P., & Edlund, J. (2021). Human-in-the-Loop Efficiency Analysis for Binary Classification in Edyson. *Proc. Interspeech 2021*, 3685–3689. https://doi.org/10.21437/Interspeech.2021-45

Jovičić, S. T. (1998). Formant feature differences between whispered and voiced sustained vowels. *Acta Acustica United with Acustica*, *84*(4), 739–743.

Lim, B. P. (2011). *Computational differences between whispered and non-whispered speech*. University of Illinois at Urbana-Champaign.

Naini, A. R., Satyapriya, M., & Ghosh, P. K. (2020). Whisper Activity Detection Using CNN-LSTM Based Attention Pooling Network Trained for a Speaker Identification Task. *INTERSPEECH*, 2922–2926.

Singhal, B., Naini, A. R., & Ghosh, P. K. (2021). wSPIRE: A Parallel Multi-Device Corpus in Neutral and Whispered Speech. *2021 24th Conference of the Oriental COCOSDA International Committee for the Co-Ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, 146–151. https://doi.org/10.1109/O-COCOSDA202152914.2021.9660449

Smith, S. D., Katherine Fredborg, B., & Kornelsen, J. (2017). An examination of the default mode network in individuals with autonomous sensory meridian response (ASMR). *Social Neuroscience*, *12*(4), 361–365.

Tartter, V. C. (1989). What’s in a whisper? *The Journal of the Acoustical Society of America*, *86*(5), 1678–1683.