Detecting whispered speech via machine learning with human-in-the-loop

Pablo Pérez Zarazaga, Zofia Malisz
KTH Royal Institute of Technology, Stockholm

Whispered speech is a very common way of communication that comes naturally to humans. It is used to reduce the audibility of the speech signal (Tartter, 1989) often when privacy of information is required. Whisper is also associated with situations of intimacy - speakers sometimes use it to explicitly transmit an intimate sensation. However, as common as it is, whisper has not been hitherto widely studied. The main acoustic characteristic of whispered speech is the lack of vibration of the vocal folds. This results in features that considerably differ from modal phonated speech e.g.: in the absence of fundamental frequency (Tartter, 1989; Jovicic,1998).

Moreover, the amount and size of datasets available for whispered speech is low. For this reason, whispered speech is usually not considered in most data-driven methods applied in speech technologies. Only recently, companies such as Amazon and Google have started to provide support for whispered speech in their voice assistants (Cotescu et al., 2020; Rekimoto, 2022) recognising whispered commands and providing responses in the same way.

In this work, we set out to make more whispered data available to speech sciences and technology. We have noticed that very large amounts of whisper data can be found in autonomous sensory meridian response (ASMR) videos uploaded to streaming platforms such as Youtube or Twitch. ASMR is an increasingly popular phenomenon that uses auditory and visual cues to create a feeling of comfort or intimacy with a relaxing effect on the listener (Barratt and Davis, 2015; Del Campo and Kehle, 2016). One of the main stimuli used in ASMR is whispered speech. It is, however, mixed with a wide variety of other acoustic triggers that make it necessary to separate the whispered sections from the noisy ones.

In this talk, we present Edyson (Fallgren et al., 2019) an audio annotating tool based on machine learning with a human-in-the-loop that can assist in the annotation of long series of noisy speech. We show how Edyson's performance can be improved on a whispered signal with a careful choice of features. After extracting features like MFCCs from the audio recording, Edyson applies dimensionality reduction techniques such as PCA or t-SNE to represent the corresponding features in a two-dimensional space. The data points are then presented to the user, who can listen to multiple segments in the same area in space and assign them labels. With a proper choice of parameter and some practice, a listener can easily separate hours of whisper from noise in several minutes.

Barratt, E. L. and Davis, N. J. (2015). Autonomous sensory meridian response (ASMR): a flow-like mental state. PeerJ, 3:e851.
Cotescu, M., Drugman, T., Huybrechts, G., Lorenzo-Trueba, J., and Moinet, A. (2020). Voice conversion for whispered speech synthesis. IEEE Signal Processing Letters, 27:186–190.
Del Campo, M. A. and 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., Malisz, Z., and Edlund, J. (2019). How to annotate 100 hours in 45 minutes. In Interspeech 2019 15-19 September 2019, Graz, pages 341–345. ISCA.
Jovicic, S. T. (1998). Formant feature differences between whispered and voiced sustained vowels. Acta Acustica united with Acustica, 84(4):739–743.
Rekimoto, J. (2022). Dualvoice: A speech interaction method using whisper-voice as commands. In Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems, New York, NY, USA.