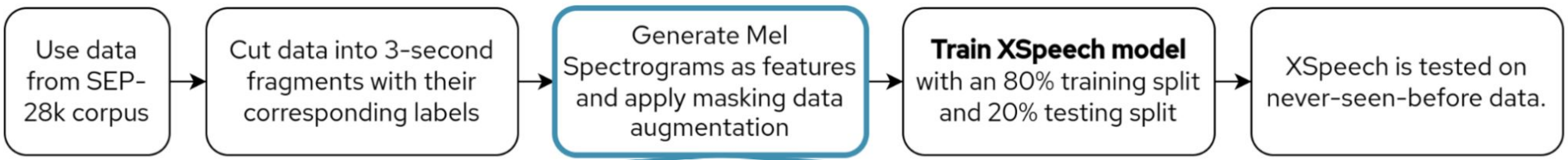
XSpeech: A Novel Deep Learning Approach to Classifying Stutters

All images by author unless otherwise noted

Research Question

How can we use deep learning to classify phonological patterns in stuttering?

[3] XSpeech Feature Processing



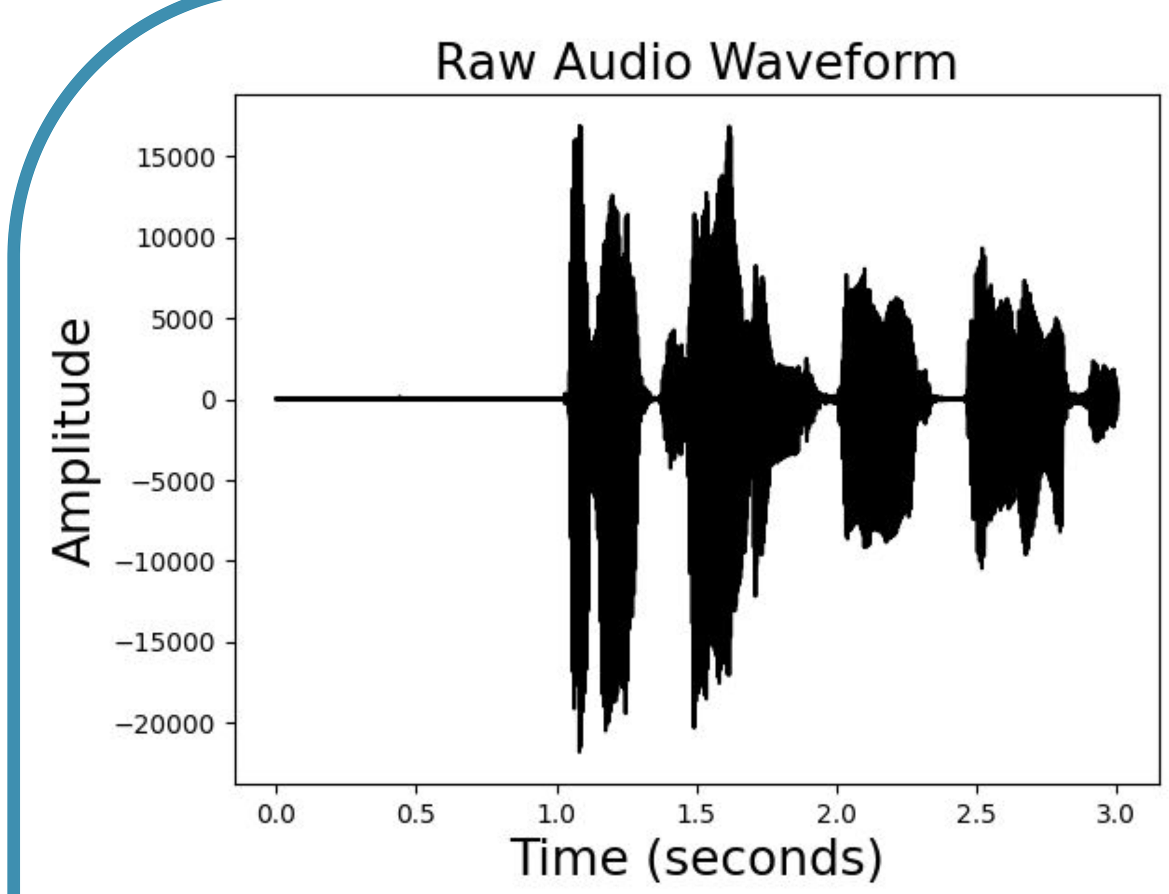


Fig. 1

Plot the amplitude of the audio clip in the temporal domain of 3 seconds per fragment. 3 seconds works the best because entire stutters can fit within the fragment while still keeping the data length low.

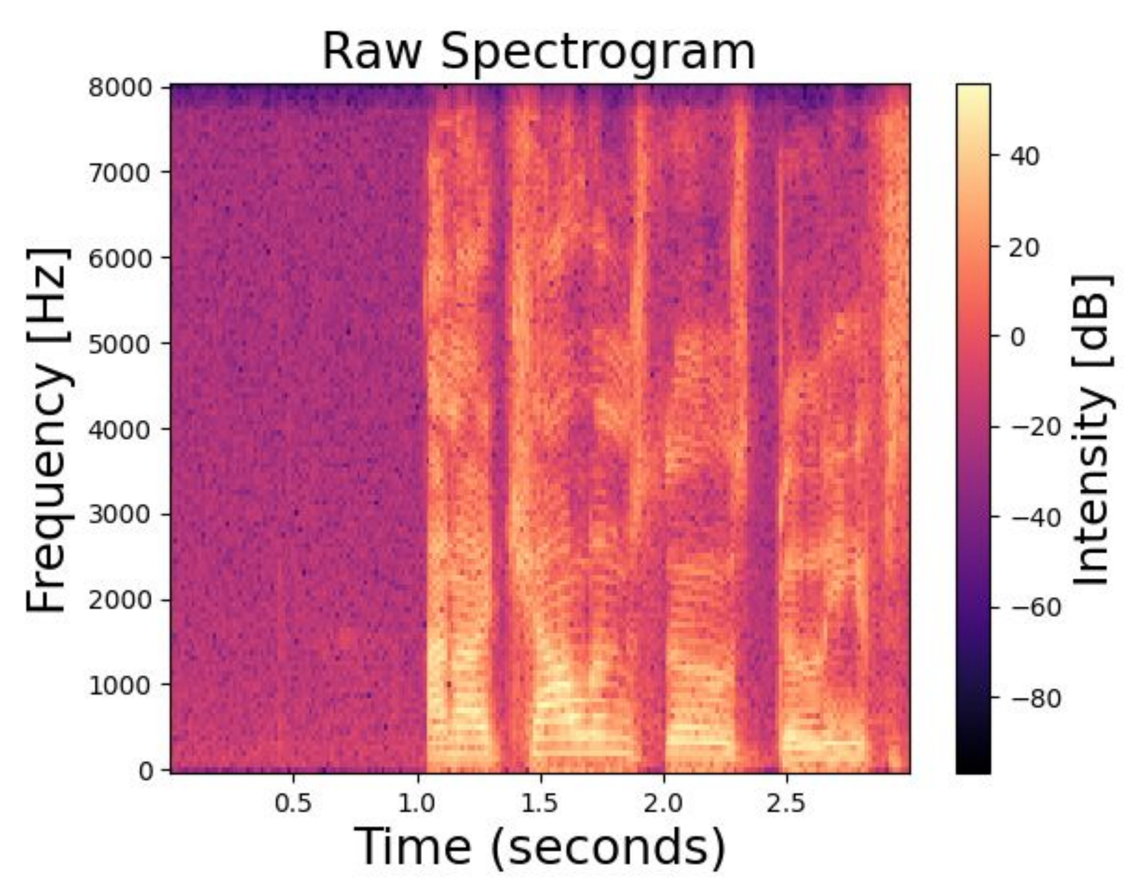


Fig. 2

The spectrogram is computed by applying a Short-time Fourier Transform (STFT) by computing Discrete Fourier Transforms (DFTs) for a rectangular sliding window with a length of 2048 samples.

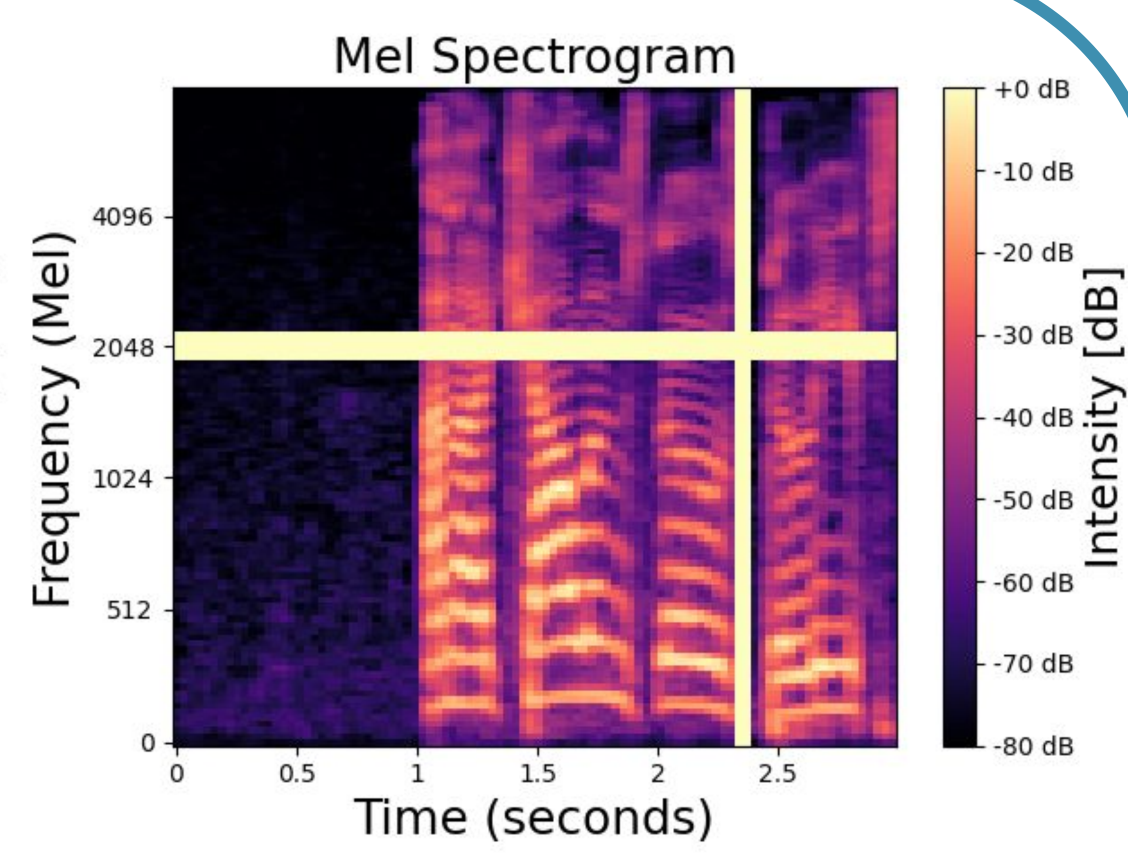


Fig. 3

Because human hearing is **non-linear**, the **mel scale** was developed as unit of pitch such that equal distances of pitch in the mel scale sounded equally distant to the listener. The Mel Spectrogram is a mapping of the frequency axis from the raw spectrogram, using the mel scale. I also applied random masks to both the time and frequency domain in order to increase model accuracy.

[4] XSpeech Model Design

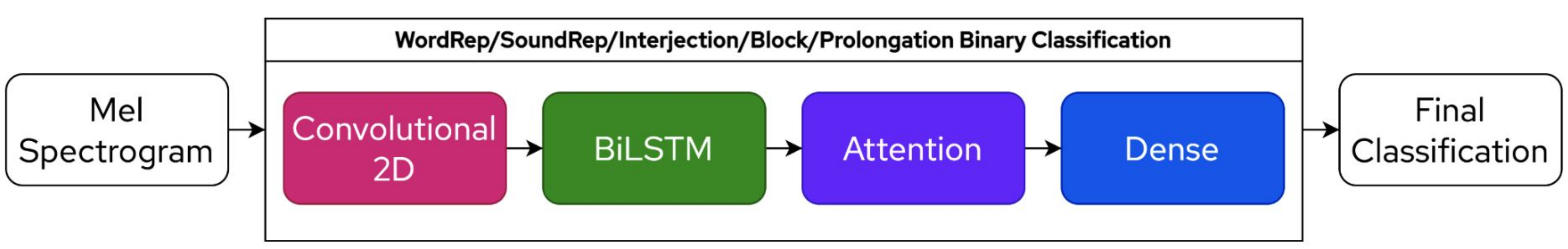


Fig. 4

XSpeech's **novel ensemble learning** approach allows 5 binary classification models to run in parallel, one for each type of stuttering. Each model provides a prediction of if the input spectrogram is an example of its assigned type of stuttering. The model that provides the highest confidence is chosen as the final result.

The **convolutional layer** extracts important features from the mel spectrogram image by sliding 64 different 3x3 windows ("filters") over the image.

$$G[m,n] = (f * h)[m,n] = \sum_j \sum_k h[j,k] f[m-j, n-k]$$

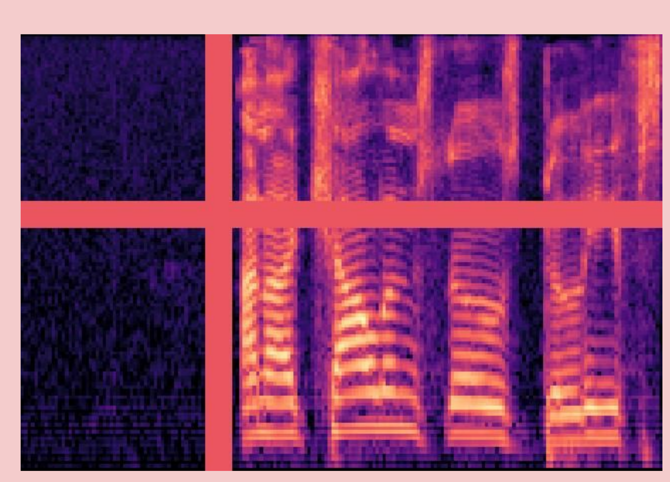


Fig. 5

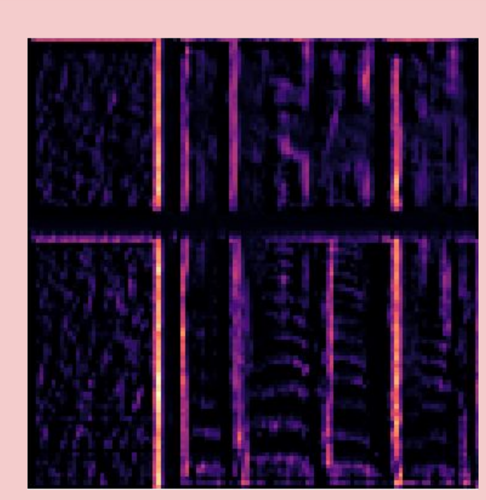


Fig. 6

The **BiLSTM** (Bi-Directional Long Short Term Memory) processes the feature map in both directions (past → future and future → past). For every timestep, the BiLSTM can use contextual information from both past and future signals in order to make a more informed prediction. The output is a 3D tensor.

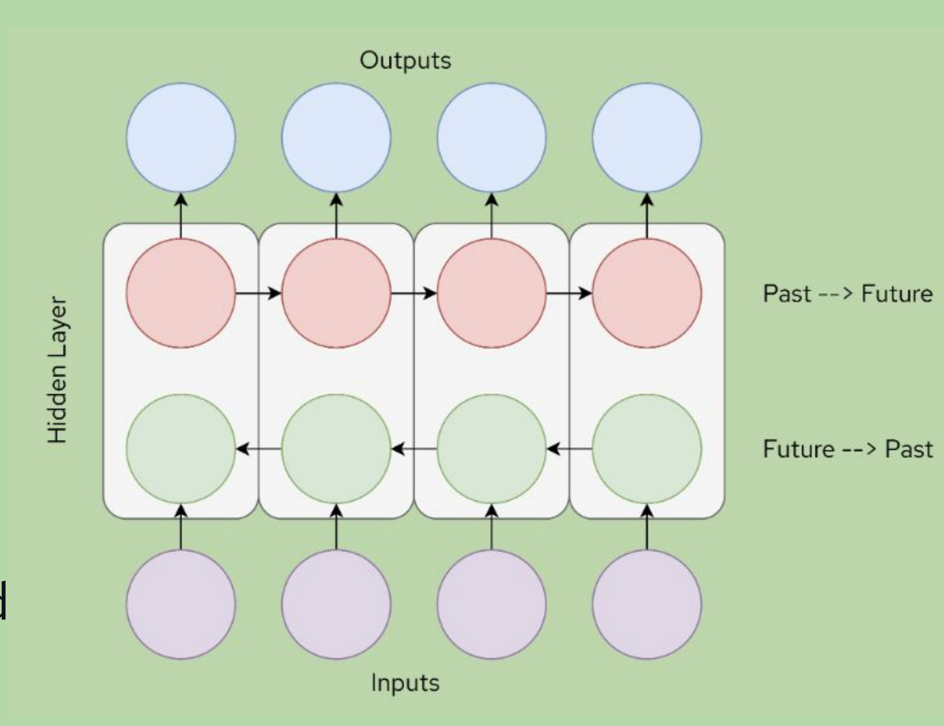


Fig. 7

While the BiLSTM provides contextual information for each timestep, the **Attention** mechanism allows XSpeech to focus on the most salient and contextually important parts of the input data (which is the BiLSTM's 3D tensor). This is similar to how a human can selectively pay attention to different signals around them.

The **Dense** layers downsample the output of the Attention layer (which is a large 3D tensor) into just one number between 0 and 1 for the final prediction.

[1] Introduction

Stuttering can be generally characterized as involuntary disfluencies during speech.

- Over **70 million** people in the world suffer from stuttering. (Ghai & Mueller, 2021)
- People who stutter (PWS) can suffer from anxiety and depression because of perceived inability to communicate with others. (Iverach et al., 2009)

Classification	Phonological Pattern
Sound Repetition (SoundRep)	Repetition of a phoneme "a-a-and"
Word Repetition (WordRep)	Repetition of a full word "and and"
Interjection	Addition of filler words "I think that - uhmm…"
Block	Unintended pause "I think …*pause*… that"
Prolongation	Stretched out sound "soooooo"

- Previous research has used **statistical** models such as hidden Markov models and support vector machines to classify stuttering. However, these approaches are less accurate (with around 70-85% accuracy) and cannot classify 5 types of stuttering.
- I propose to use a novel deep learning approach (**XSpeech**) to classify all 5 types of stuttering with >90% accuracy, helping PWS to **understand and correct their stutter**, allowing for easier and freer communication.

[2] Related Information

SEP-28k

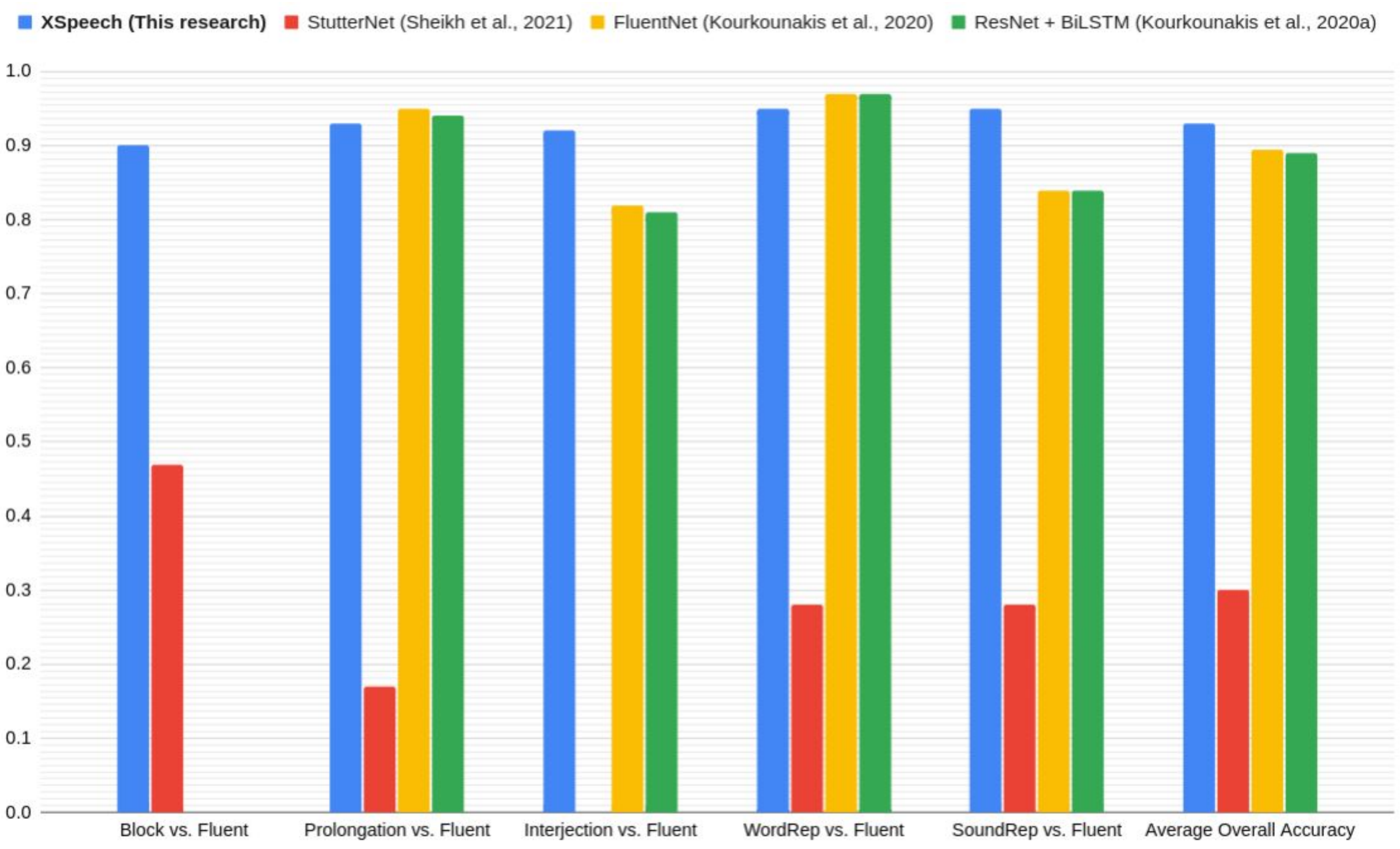
This public dataset contains around 23.3 hours of natural events of stuttering, recorded from public podcasts of people who stutter. The data is pre-labeled with the correct classification of each stuttering event.

Feature Extraction

A computer records noise with frequencies that humans do not use when understanding speech (e.g breathing noises, background static, etc). So, we need to filter out the unwanted noise and amplify frequencies that are found in human speech.

[5] Results

Fig. 8



Accuracy

- XSpeech can classify **all 5 types of stuttering**
- Individual accuracies for each separate binary classification task were all **90% or higher**.
    - Classifying **Blocks** had the lowest accuracy of **90%.**
    - Classifying **Word Repetitions** had the highest accuracy of **96%**
- Average overall accuracy of **93.2%.**

Comparison

- XSpeech was compared against other state-of-the-art DL stuttering classification models from recent literature.
- Achieved a **breakthrough** of over 90% overall accuracy.
- XSpeech **outperforms all other models** when classifying **Interjections, SoundRep, and Blocks.**

Future Work

- XSpeech can be used to automate **speech therapy** in order to help the therapist and the patient understand the patient's stuttering patterns.

[6] Works Cited

Al-Banna, A., Edirisinghe, E. A., & Fang, H. (2022). Stuttering Detection Using Atrous Convolutional Neural Networks. *2022 13th International Conference on Information and Communication Systems (ICICS)*. https://doi.org/10.1109/icics55353.2022.9811183

Arısoy, E., Sethy, A., Ramabhadran, B., & Chen, S. (2015). Bidirectional recurrent neural network language models for automatic speech recognition. *ICASSP 2015*. https://doi.org/10.1109/icassp.2015.7179007

Ghai, B., & Mueller, K. (2021). Fluent: An AI Augmented Writing Tool for People who Stutter. *arXiv*. https://doi.org/10.1145/3441852.3471211

Iverach, L., Jones, M., McLellan, L. F., Lyneham, H. J., Menzies, R. G., Onslow, M., & Rapee, R. M. (2016). Prevalence of anxiety disorders among children who stutter. *Journal of Fluency Disorders*, *49*, 13–28. https://doi.org/10.1016/j.jfludis.2016.07.002

Kourkounakis, T., Hajavi, A., & Etemad, A. (2020a). Detecting Multiple Speech Disfluencies Using a Deep Residual Network with Bidirectional Long Short-Term Memory. *ICASSP 2020*. https://doi.org/10.1109/icassp40776.2020.9053893

Kourkounakis, T., Hajavi, A., & Etemad, A. (2020b). FluentNet: End-to-End Detection of Speech Disfluency with Deep Learning. *arXiv (Cornell University)*. http://export.arxiv.org/pdf/2009.11394

Ooi, C. A., Hariharan, M., Yaacob, S., & Chee, L. S. (2012). Classification of speech dysfluencies with MFCC and LPCC features. *Expert Systems With Applications*, *39*(2), 2157–2165. https://doi.org/10.1016/j.eswa.2011.07.065

Sheikh, S. A., Sahidullah, M., Hirsch, F., & Ouni, S. (2021). StutterNet: Stuttering Detection using Time delay Neural network. *2021 29th European Signal Processing Conference (EUSIPCO)*. https://doi.org/10.23919/eusipco54536.2021.9616063

Sheikh, S. A., Sahidullah, M., Hirsch, F., & Ouni, S. (2022). Machine learning for stuttering identification: Review, challenges and future directions. *Neurocomputing*, *514*, 385–402. https://doi.org/10.1016/j.neucom.2022.10.015

Smith, A., & Weber, C. (2017). How stuttering Develops: The Multifactorial Dynamic Pathways Theory. *Journal of Speech Language and Hearing Research*, *60*(9), 2483–2505. https://doi.org/10.1044/2017_jslhr-s-16-0343