UKRI Centre for Doctoral Training NLP

# Self-supervised models for dysarthric speech

## *Understanding representations through visual analysis and probing*

**Ariadna Sanchez, Simon King**

**Dysarthria:** disorder caused by a damage in the nervous system, which causes difficulties to move speech motor muscles. This often causes slurred or slow speech, and imprecise articulation.

## The short story

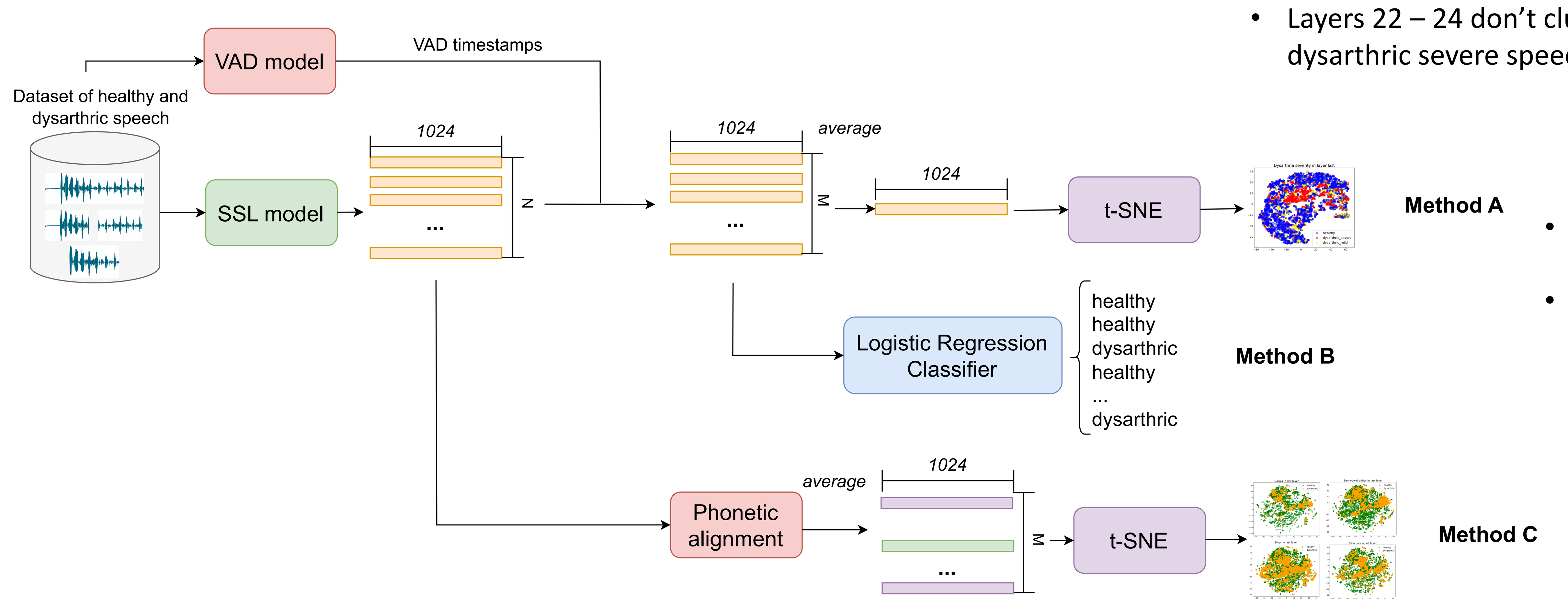
1. Previous work deploys self-supervised learning (SSL) models as feature extractors for dysarthric speech classification.
2. There is not enough work exploring how these features represent dysarthric speech. We use visual analysis and probing techniques used in prior work for healthy speech.
3. Visualisation techniques suggest that severely dysarthric speech forms a cluster in the latent space but occupies a smaller region than healthy speech.
4. Logistic regression models classify severely dysarthric speech with high accuracy, but not mildly dysarthric speech.

## Introduction

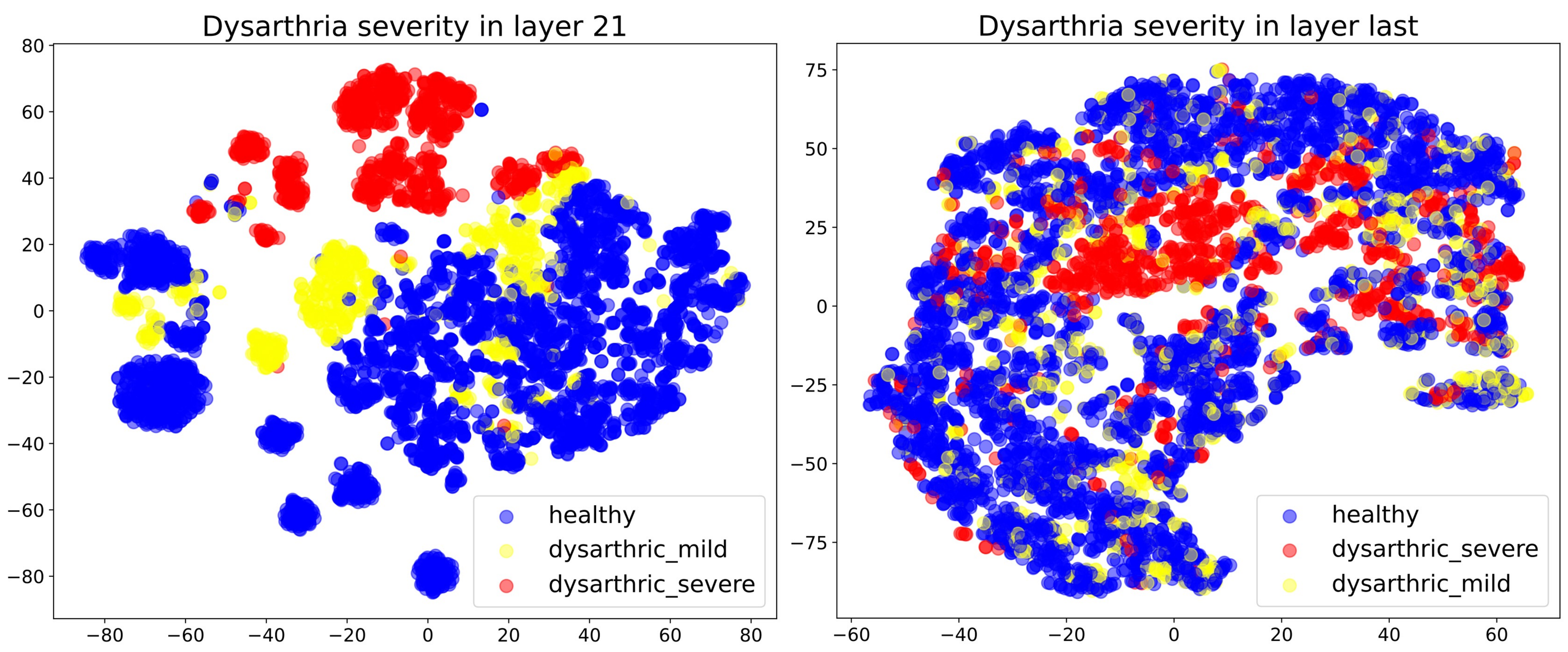
- Dysarthria, which stems from difficulties controlling motor regions, is commonly present in neurodegenerative conditions.
- Clinicians have access to vast amounts of information but not enough methods or time to find patterns between patients.
- Previous work by others trained pathology detection models with classic feature extraction models, e.g., MFCC's, filterbanks. Recently, SSL features have been explored, but there are still questions on how pre-trained SSL models represent dysarthric speech.

**OBJECTIVE: Analyse all layers of a pre-trained SSL model through visualisation and probing analyses previously performed in healthy speech [1, 2].**

## Methods




## Results – Method A

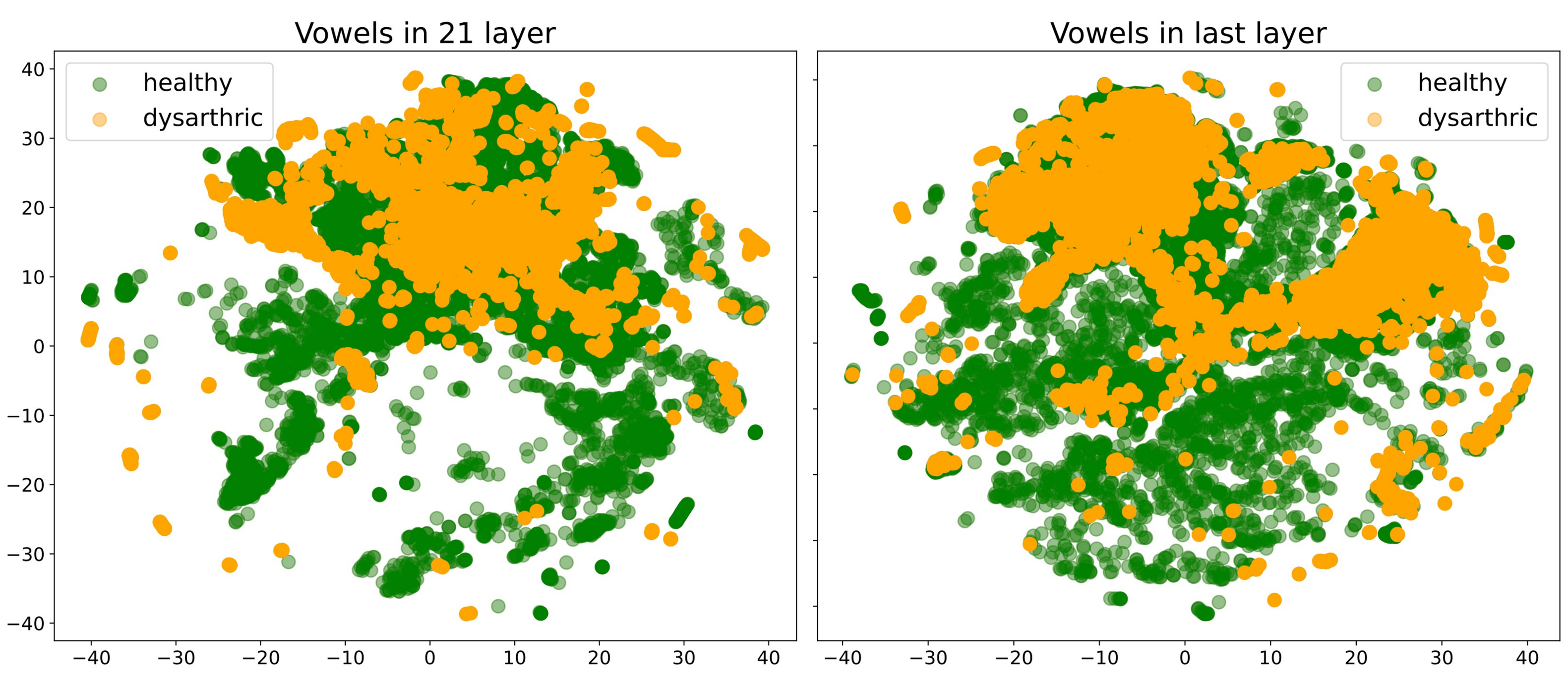



- Layers 1 – 21 appear to separate severe dysarthric speakers from healthy and dysarthric mild speakers. There are clear speaker clusters.
- Layers 22 – 24 don't cluster embeddings per speaker, but still shows that dysarthric severe speech is clustered in a specific area of the latent space.

## Implementation

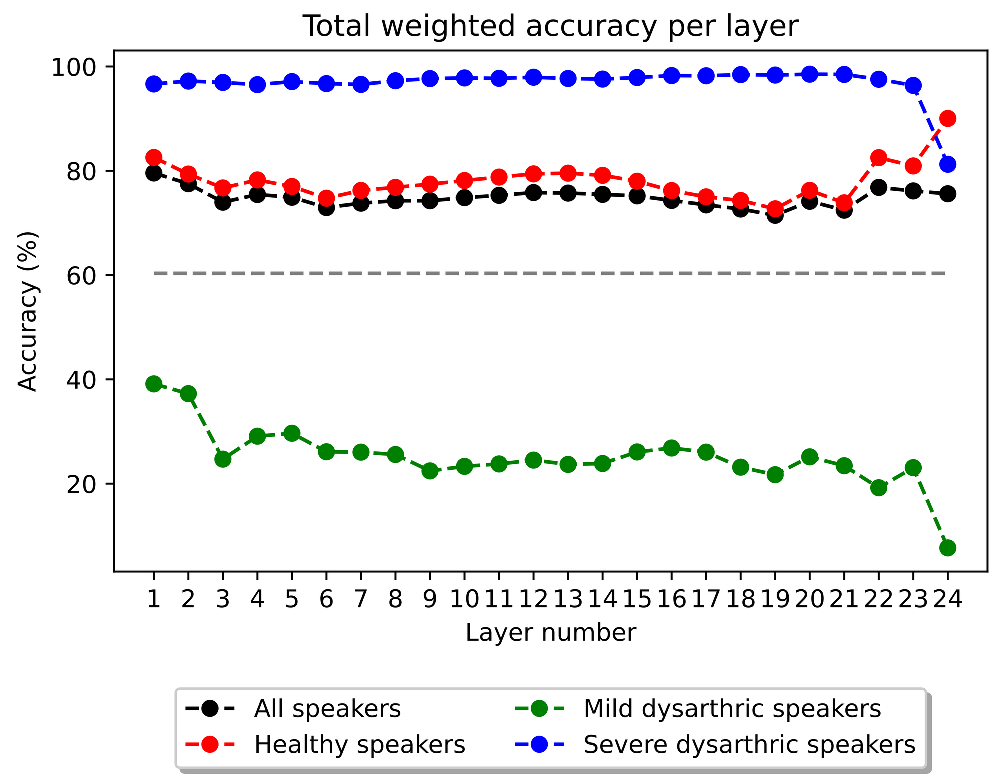
- SSL model: XLSR-53 [3], a pre-trained SSL model trained with 53 languages, with a total of 24 layers.
- Dataset: TORGO [4], a dataset in American English composed of dysarthric speech and matched healthy control speakers. Commonly used in dysarthric speech applications.

## Results – Method C




- Averaged embeddings per phonetic representation.
- For vowels, semivowels, and nasals, dysarthric speech occupies a smaller region of the latent space than healthy speech.
- Correlates with the reduction in articulation and difficulty in mobility presented by speakers with dysarthria.

## Results – Method B




- Binary classification, with accuracy divided by type of speech.
- High accuracy when detecting severely dysarthric speakers for layers 1-23.
- Mild dysarthric speakers' accuracy detection is below the input-independent bottom line (dashed grey line).
- However, most mildly dysarthric speakers in the TORGO dataset do not present audible dysarthric traits in their speech.

## Future Work

- Investigate other datasets in different languages and with more granularity in severity types.
- Explore SSL embeddings in text-to-speech/voice repair applications, e.g., as additional embeddings in the text-to-speech architecture.

*[1] Wells, D., Tang, H., Richmond, K. (2022) Phonetic Analysis of Self-supervised Representations of English Speech. Proc. Interspeech 2022, 3583-3587, doi: 10.21437/Interspeech.2022-10884*
*[2] de Seyssel, M., Lavechin, M., Adi, Y., Dupoux, E., Wisniewski, G. (2022) Probing phoneme, language and speaker information in unsupervised speech representations. Proc. Interspeech 2022, 1402-1406, doi: 10.21437/Interspeech.2022-373*
*[3] Conneau, A., Baevski, A., Collobert, R., Mohamed, A., Auli, M. (2021) Unsupervised Cross-Lingual Representation Learning for Speech Recognition. Proc. Interspeech 2021, 2426-2430, doi: 10.21437/Interspeech.2021-329*
*[4] Rudzicz, F., Namasivayam, A. K., & Wolff, T. (2012). The TORGO database of acoustic and articulatory speech from speakers with dysarthria. Language resources and evaluation, 46, 523-541.*

ariadna.sanchez@ed.ac.uk

*This work was supported by the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UKRI (grant EP/S022481/1).*