SmartLab
Intelligent Interactions

MŰEGYETEM 1782

TMiT

MiSTR

Multi-Modal iEEG-to-Speech Synthesis with Transformer-Based Prosody Prediction and Neural Phase Reconstruction

Mohammed Salah Al-Radhi, Géza Németh, Branislav Gerazov

malradhi@tmit.bme.hu

EKÖP
Egyetemi Kutatói Ösztöndíj Program

August 19, 2025

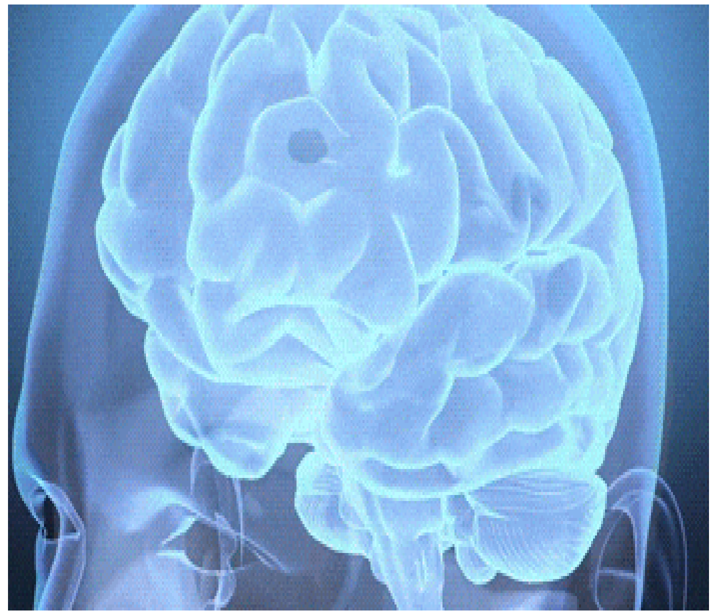
ENFIELD

What is Brain Activity?

- it refers to the electrical, chemical, and metabolic signals generated by **neurons**.
- **Neurons** communicate through electrical impulses called **action potentials**.

Type of Brain Signals:

- **Electrical Signals:** Measured as voltage fluctuations (EEG).
- **Metabolic Signals:** Changes in oxygen and glucose levels (fMRI, PET).



How can we measure brain activity?

EEG
(Electroencephalography)
Captures electrical activity of the brain

fMRI
(Functional Magnetic Resonance Imaging)
Tracks oxygenated blood flow

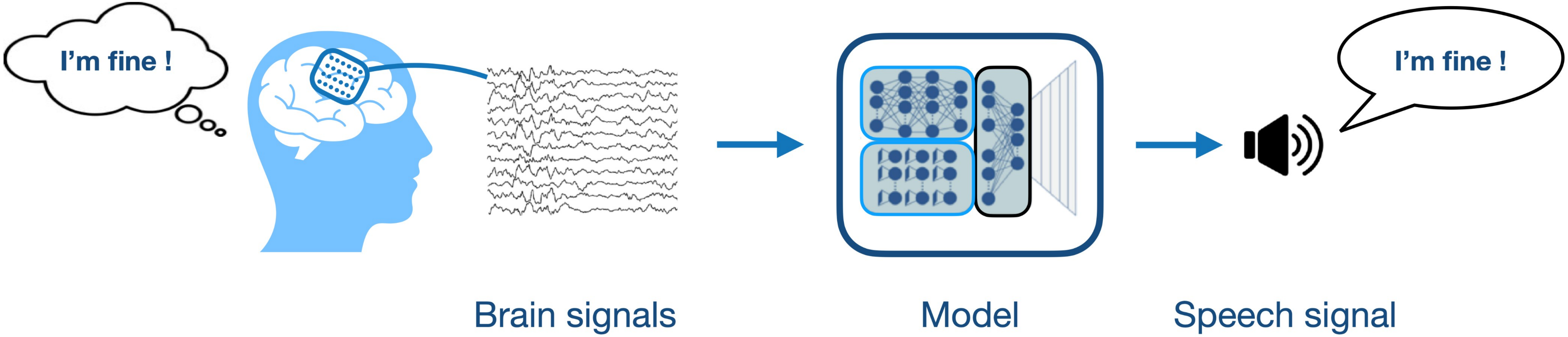
MEG
(Magnetoencephalography)
Measures magnetic fields produced by neural currents

Comparison table

Characteristic	EEG	fMRI	MEG
Temporal Resolution	High (millisecond scale)	Low (seconds)	High (millisecond scale)
Spatial Resolution	Low (centimeters)	High (millimeters)	Moderate (centimeters to millimeters)
Cost	Relatively low	High	Very high
Portability	Good can be used in various settings	Poor requires a large, stationary scanner	Poor requires a magnetically shielded room for best results
Sensitivity	Sensitive to surface electrical activity	Sensitive to changes in blood flow related to neural activity	Sensitive to magnetic fields from deeper brain structures
Noise Immunity	Susceptible to electrical noise	Less affected by noise, but can be influenced by motion and magnetic artifacts	Sensitive to magnetic noise, thus requires shielding

Is it possible to decode speech from brain signals?

Is it possible to decode speech from brain signals?



Challenges

- **Challenges in Brain-to-Speech:**
  - Neural signals are noisy, non-stationary, and vary across individuals and sessions.
  - Aligning neural features with prosodic and linguistic cues is complex.
  - Limited high-quality, annotated datasets hinder robust model training.

Challenges & Motivation

❑ Challenges in Brain-to-Speech:

- Neural signals are noisy, non-stationary, and vary across individuals and sessions.
- Aligning neural features with prosodic and linguistic cues is complex.
- Limited high-quality, annotated datasets hinder robust model training.

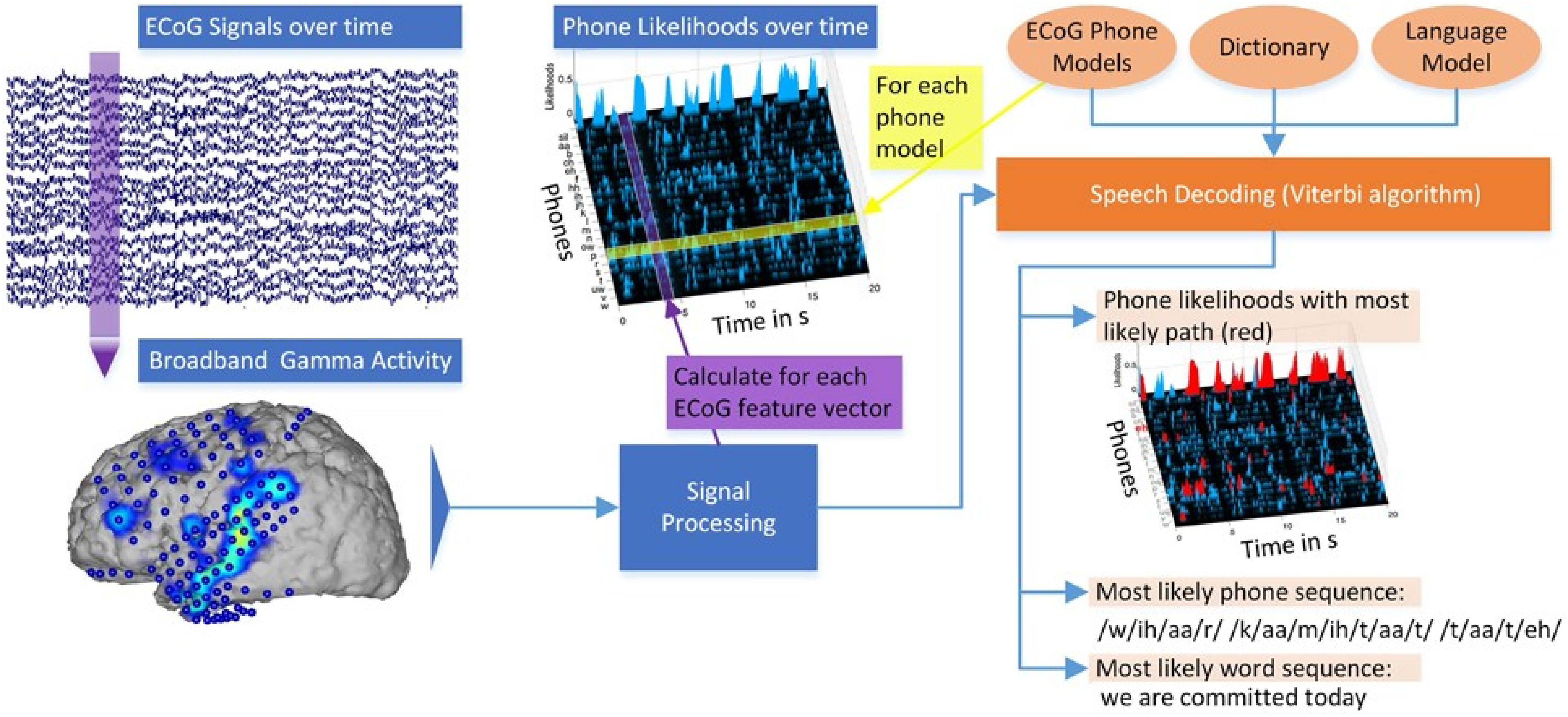
❑ Motivation for Our Approach:

- Extract richer neural features to better capture speech dynamics.
- Incorporate prosody for more natural and expressive reconstructions.
- Reduce vocoder phase artifacts to improve intelligibility and perceived quality.

Previous Methodology

Phoneme Approach

- Decode discrete phoneme sequences directly from neural signals.
- Use a phoneme-to-speech synthesizer to generate audio output.
- Focus: symbolic **linguistic units**, not acoustic detail or continuous motion.

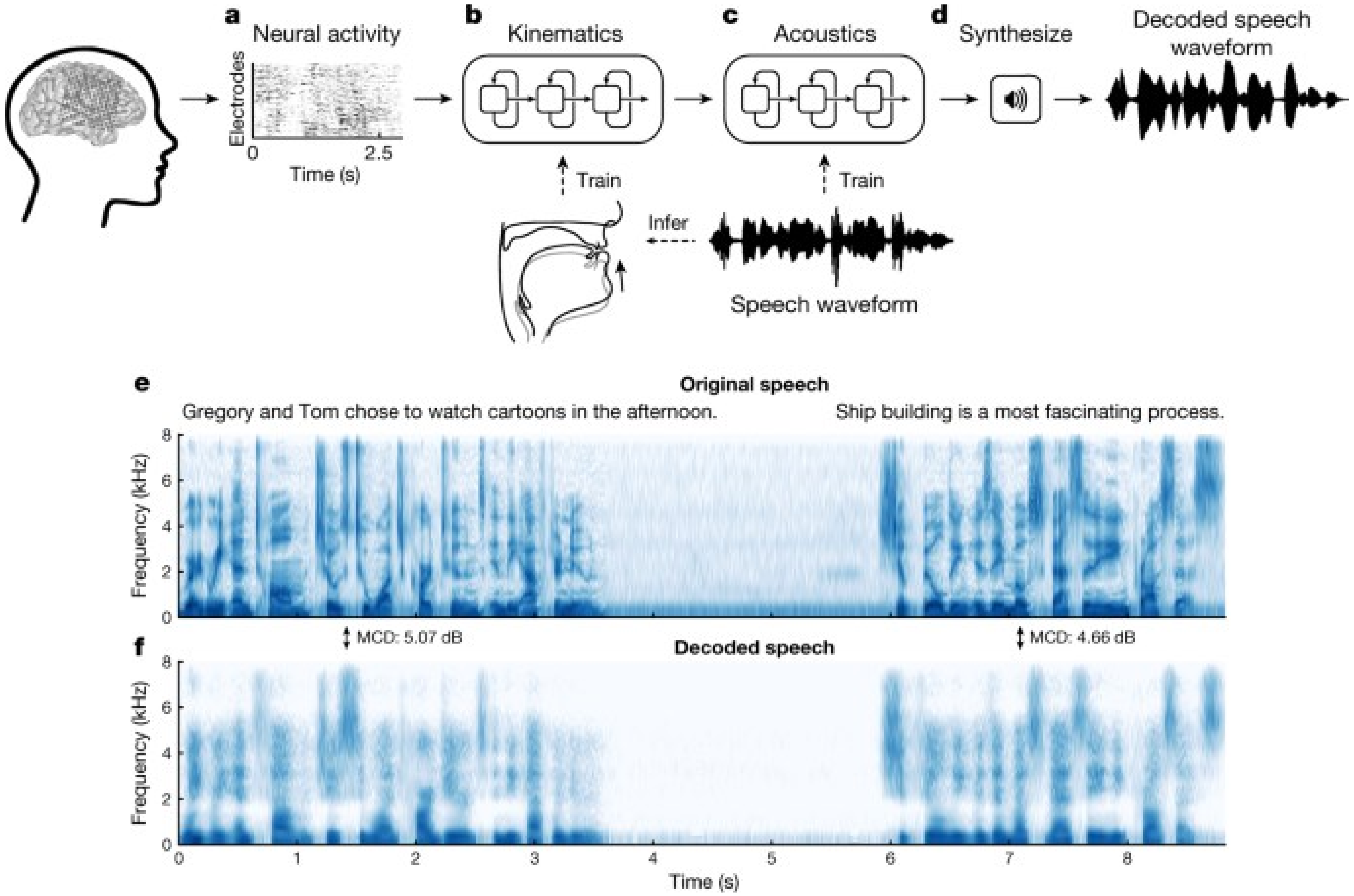


❖ Ignores speech features like tone and emotion, focusing only on semantic content, which reduces naturalness.

Herff C, Heger D, de Pesters A, Telaar D, Brunner P, Schalk G and Schultz T. Brain-to-text: decoding spoken phrases from phone representations in the brain. Front. Neurosci. 9:217, 2015

Kinematic Approach

- Decode **continuous motor trajectories** (e.g., tongue, lips, jaw) from neural signals.
- Reconstruct speech by driving a synthesizer using these trajectories.
- Focus: **movement patterns**, not direct visual articulation.

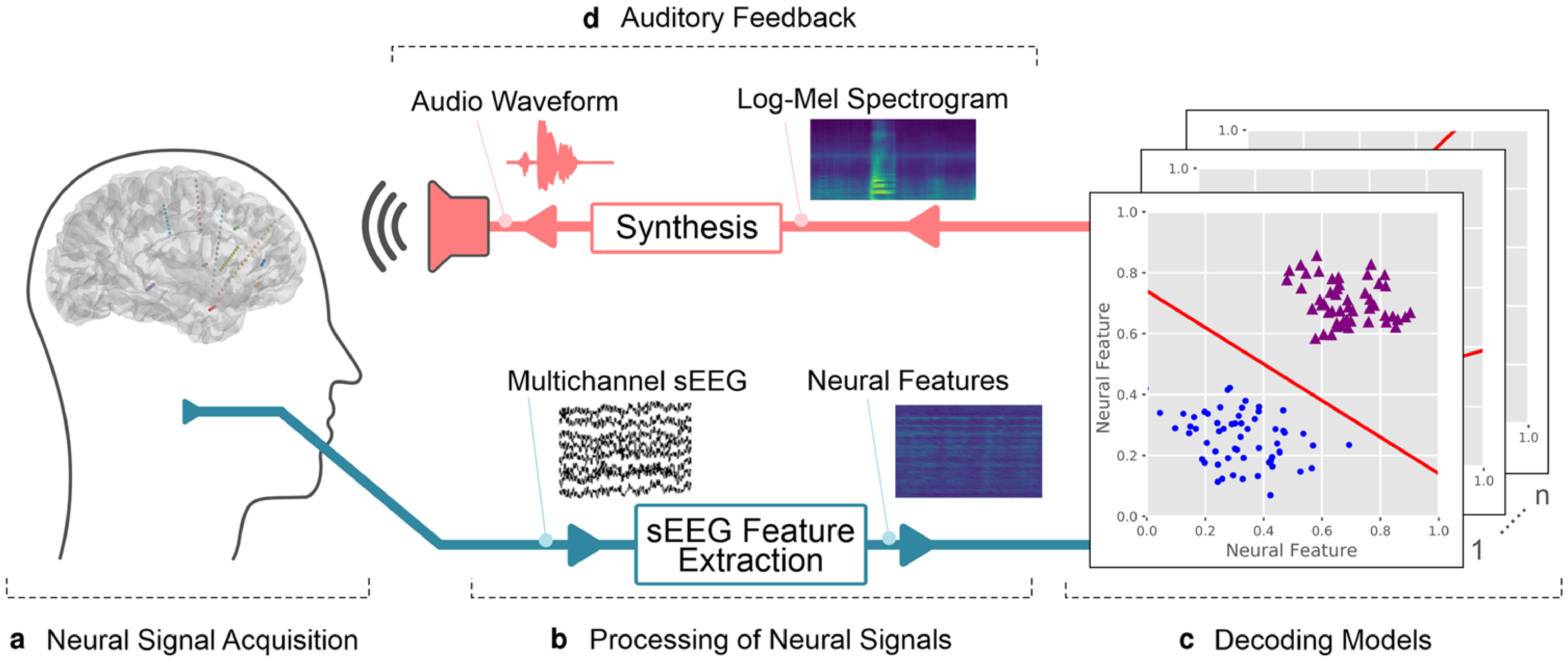


❖ Relies on accurately decoding motor representations, which are noisy and may not capture coarticulation or speech dynamics fully.

Anumanchipalli, G.K., Chartier, J. & Chang, E.F. Speech synthesis from neural decoding of spoken sentences. Nature 568, 493–498, 2019

Spectrogram Approach

- Predict time–frequency acoustic features (e.g., mel-spectrogram bins) from neural activity.
- Reconstruct speech using a simple vocoder from predicted spectrograms.
- Focus: **fine acoustic detail** for naturalness, bypassing phoneme/articulator steps.



❖ Requires high-quality neural signals and is heavily dependent on vocoder performance.

Angrick, M., Ottenhoff, M. C., Diener, L., Ivucic, D., Ivucic, G., Goulis, S., ... & Herff, C. Real-time synthesis of imagined speech processes from minimally invasive recordings of neural activity. *Communications biology*, *4*(1), 2021

Articulation Approach

- Map neural activity to **visual articulatory data** (e.g., ultrasound tongue imaging, EMA).
- Convert articulatory representations into acoustic features for speech synthesis.
- Focus: **structural, image-based articulator shapes**, not just trajectory control.

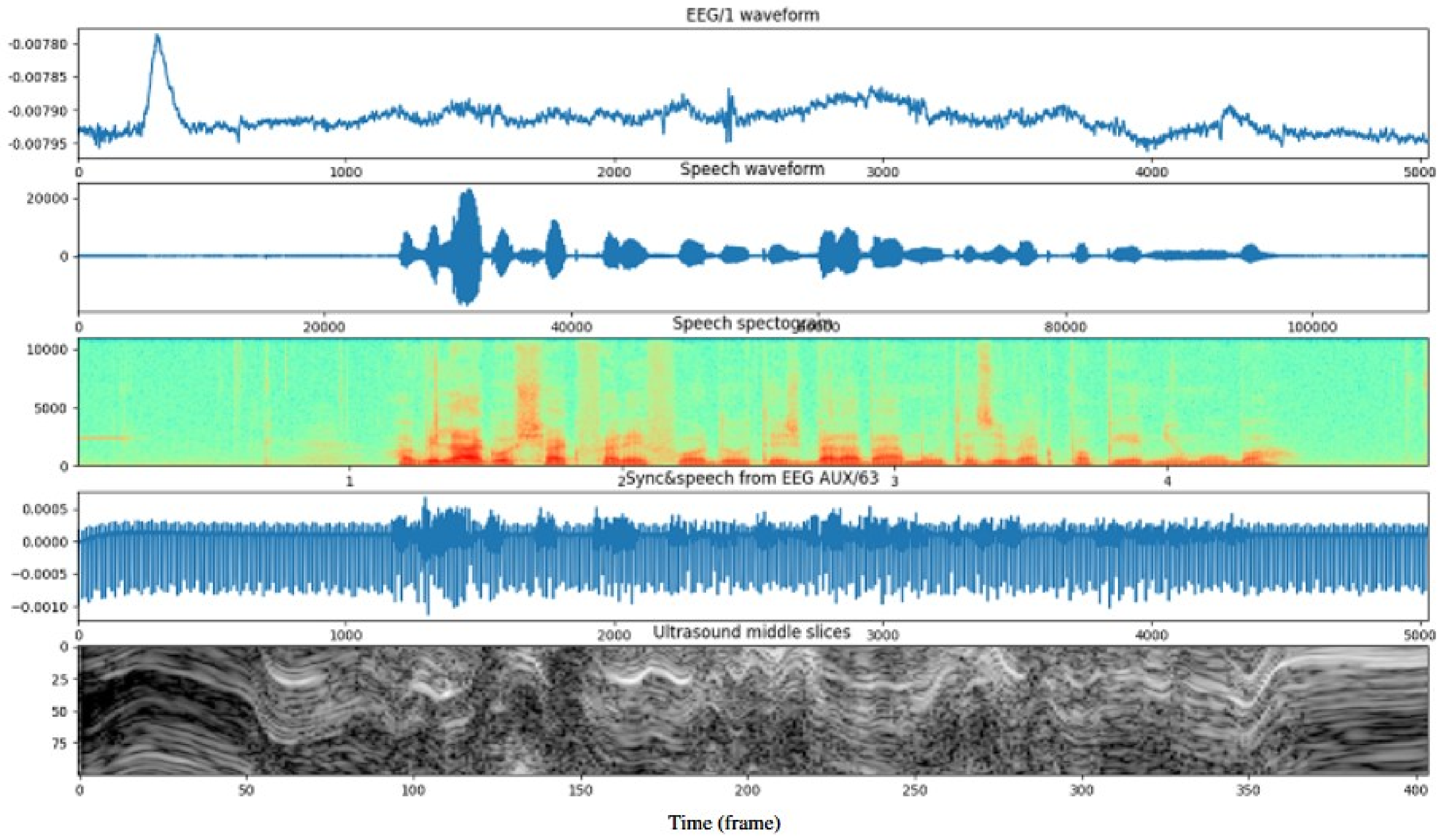


Figure 1: *Example for synchronized EEG, speech, and ultrasound tongue imaging recordings. a) EEG / 1st channel, b) speech signal, c) speech spectrogram, d) ultrasound synchronization signal and speech signal (EEG on AUX), e) temporal change of the center line of UTIs.*

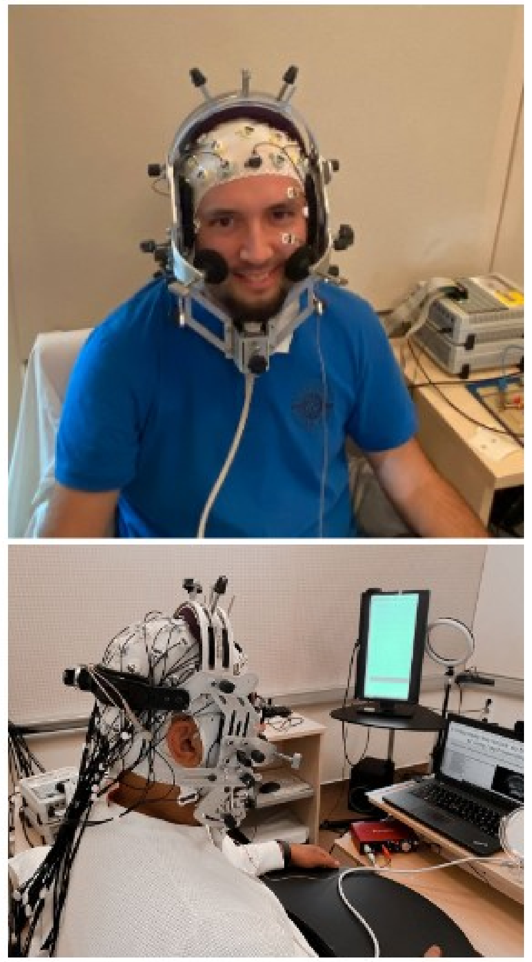


Figure 2: *Recording setup: EEG, ultrasound tongue imaging with a headset, microphone and webcam.*

❖ Mapping brain activity to articulatory movements is complex and often lacks prosody and expressiveness in the output.

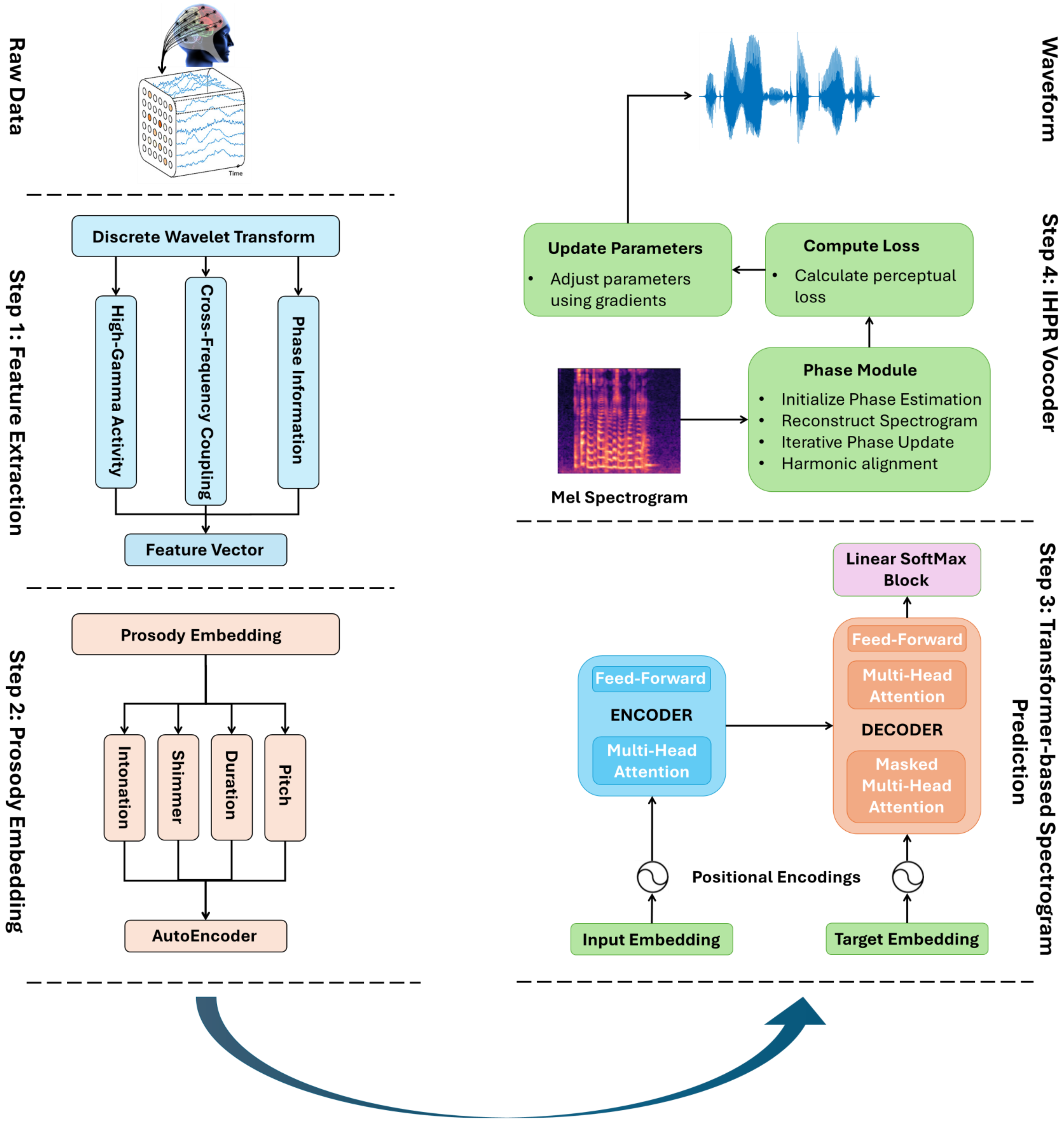
T. G. Csapó, Frigyes Viktor Arthur, Péter Nagy, Ádám Boncz, Towards Ultrasound Tongue Image prediction from EEG during speech production, Interspeech, Dublin, Ireland, 2023

Proposed Methodology

Our Contributions

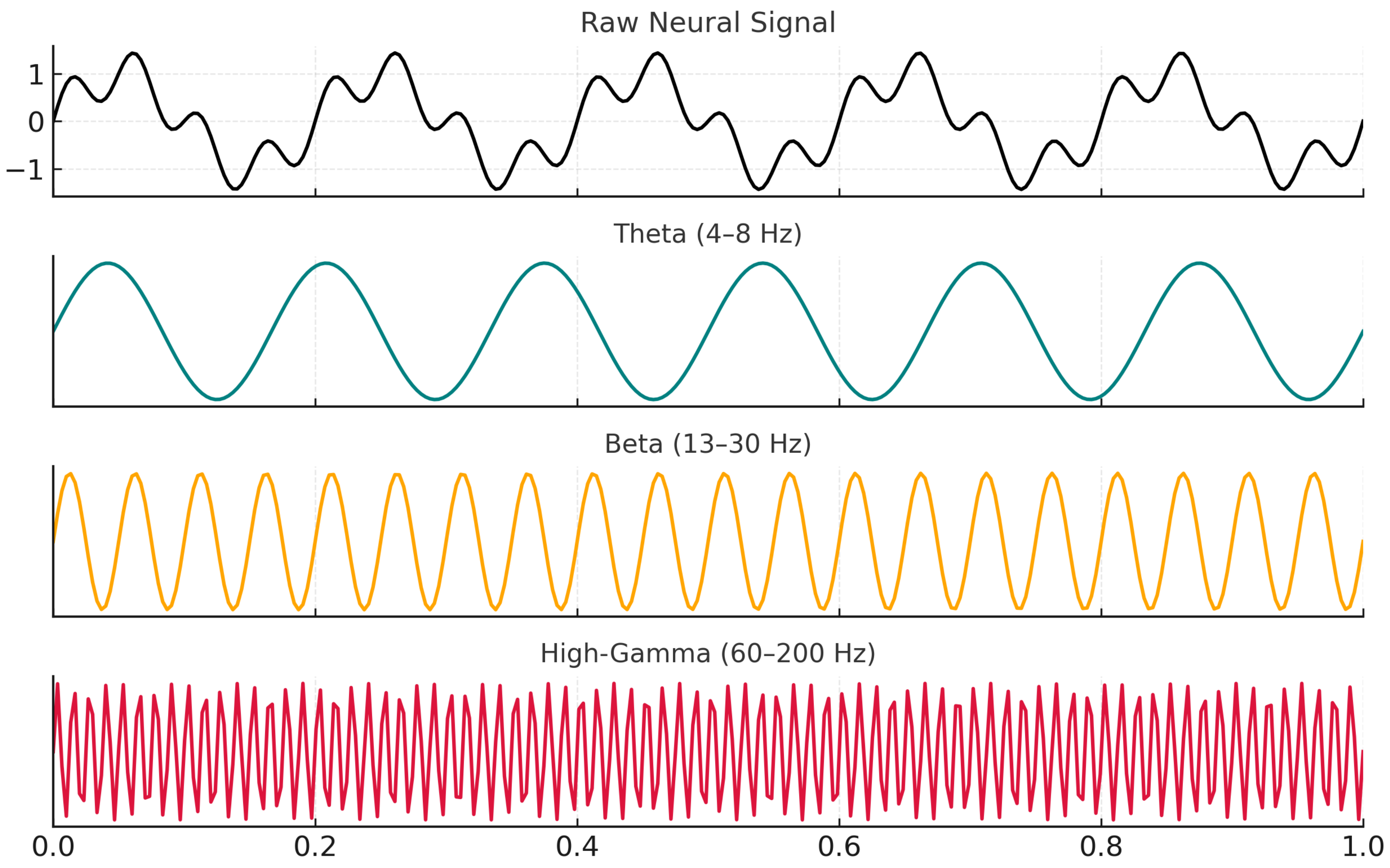
- Wavelet-based multi-modal feature extraction capturing articulatory and prosodic cues.
- Prosody-aware Transformer for accurate and expressive spectrogram prediction.
- IHPR neural phase vocoder for artifact-free, harmonically aligned speech synthesis.

MiSTR Overview Diagram



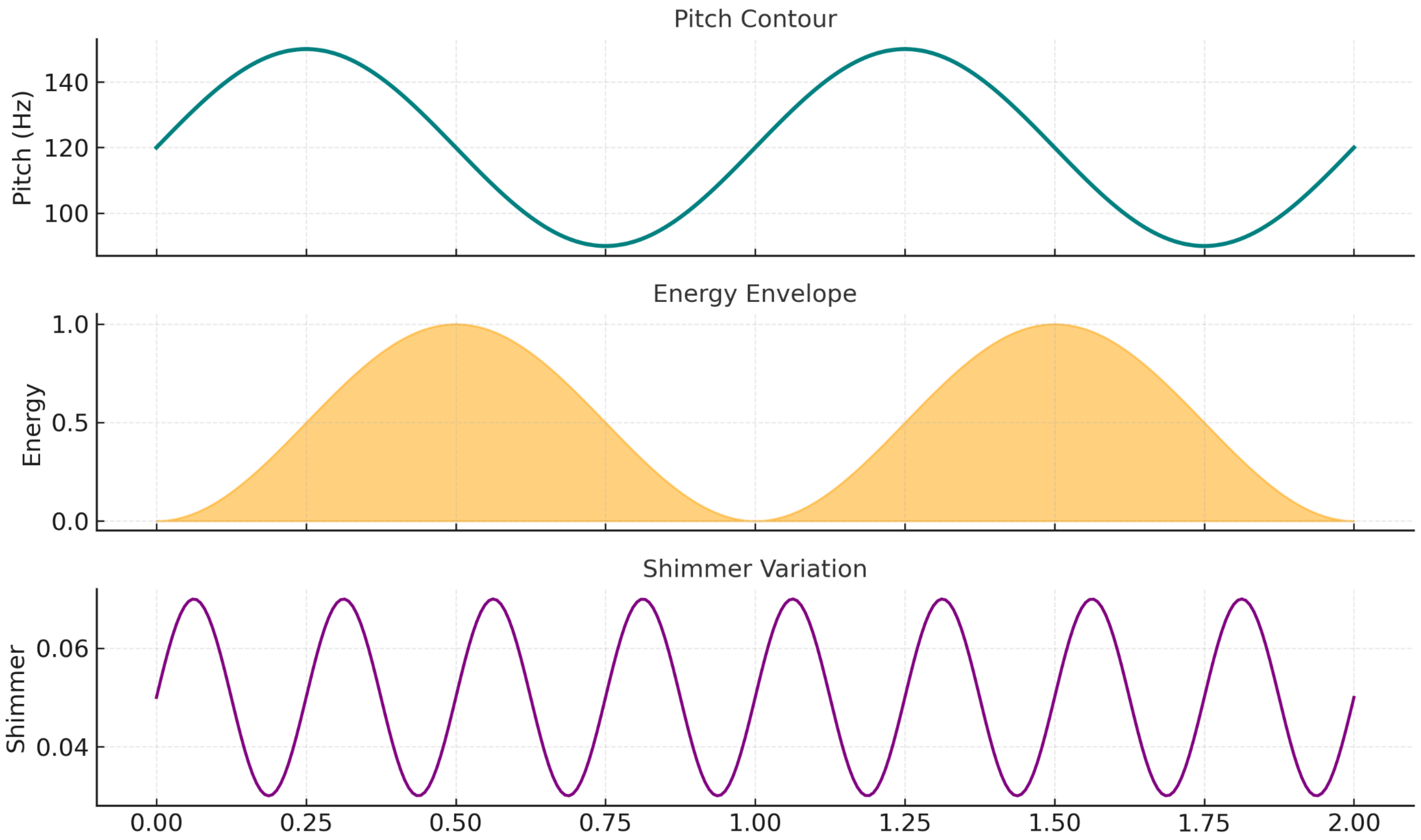
Step 1 – Wavelet-Based Feature Extraction

- We decompose neural recordings into theta (4–8 Hz), beta (13–30 Hz), and high-gamma (60–200 Hz) bands.
- This multi-band analysis preserves information about articulation, rhythm, and fine acoustic detail.
- Wavelet transforms enable localized, time-frequency analysis for capturing transient neural events.



Step 2 – Prosody Features

- Extract prosody features including pitch contour, energy variation, shimmer, and speech segment durations.
- These features are essential for producing speech that sounds expressive and human-like.
- Prosodic information complements spectral features, improving both intelligibility and naturalness.

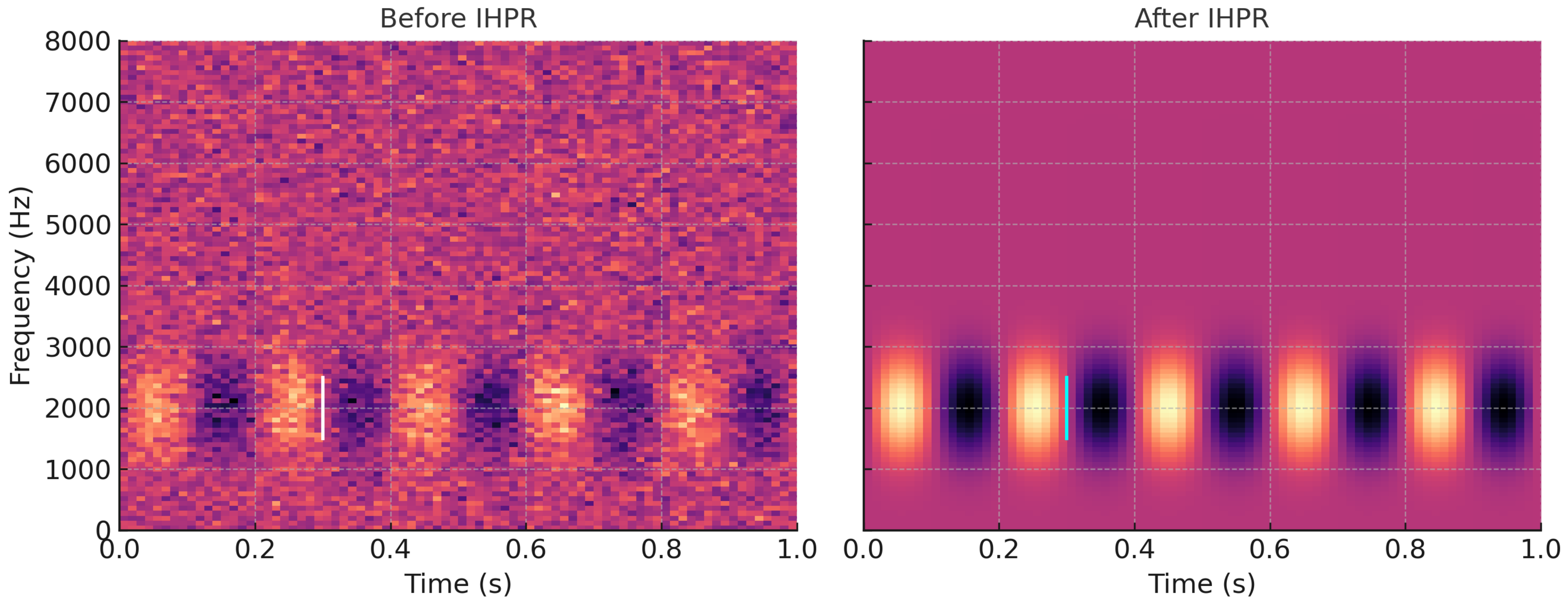


Step 3 – Transformer Spectrogram Prediction

- The Transformer architecture models long-range dependencies better than RNN-based approaches.
- Multi-head self-attention allows the model to focus on different temporal and spectral patterns simultaneously.
- This leads to more coherent spectrogram predictions, especially for complex phoneme sequences.

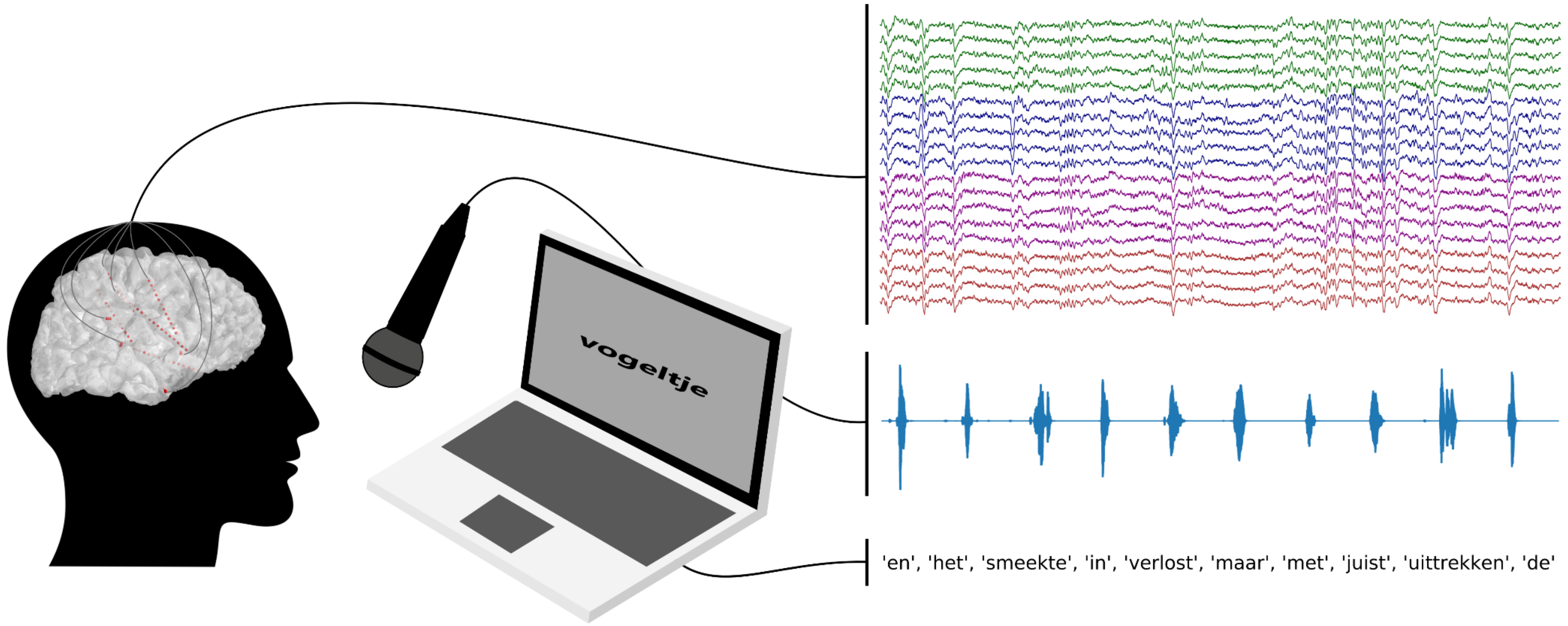
Step 4 – IHPR Phase Vocoder

- IHPR (Iterative Harmonic Phase Refinement) enforces phase continuity across harmonics.
- **Before IHPR:** noticeable phase misalignments produce metallic or distorted speech.
- **After IHPR:** harmonics are aligned, reducing artifacts and improving perceived quality.



Dataset

- Epilepsy patients
- Sessions ~2 hours
- 10 participants, native speakers of Dutch
- mean age 32 years (range 16–50 years); 5 male, 5 female).
- Speaking Dutch words aloud while audio and intracranial EEG data are recorded simultaneously
- Lab streaming layer (ref)
  - Neural stream
  - Audio stream
  - Marker stream

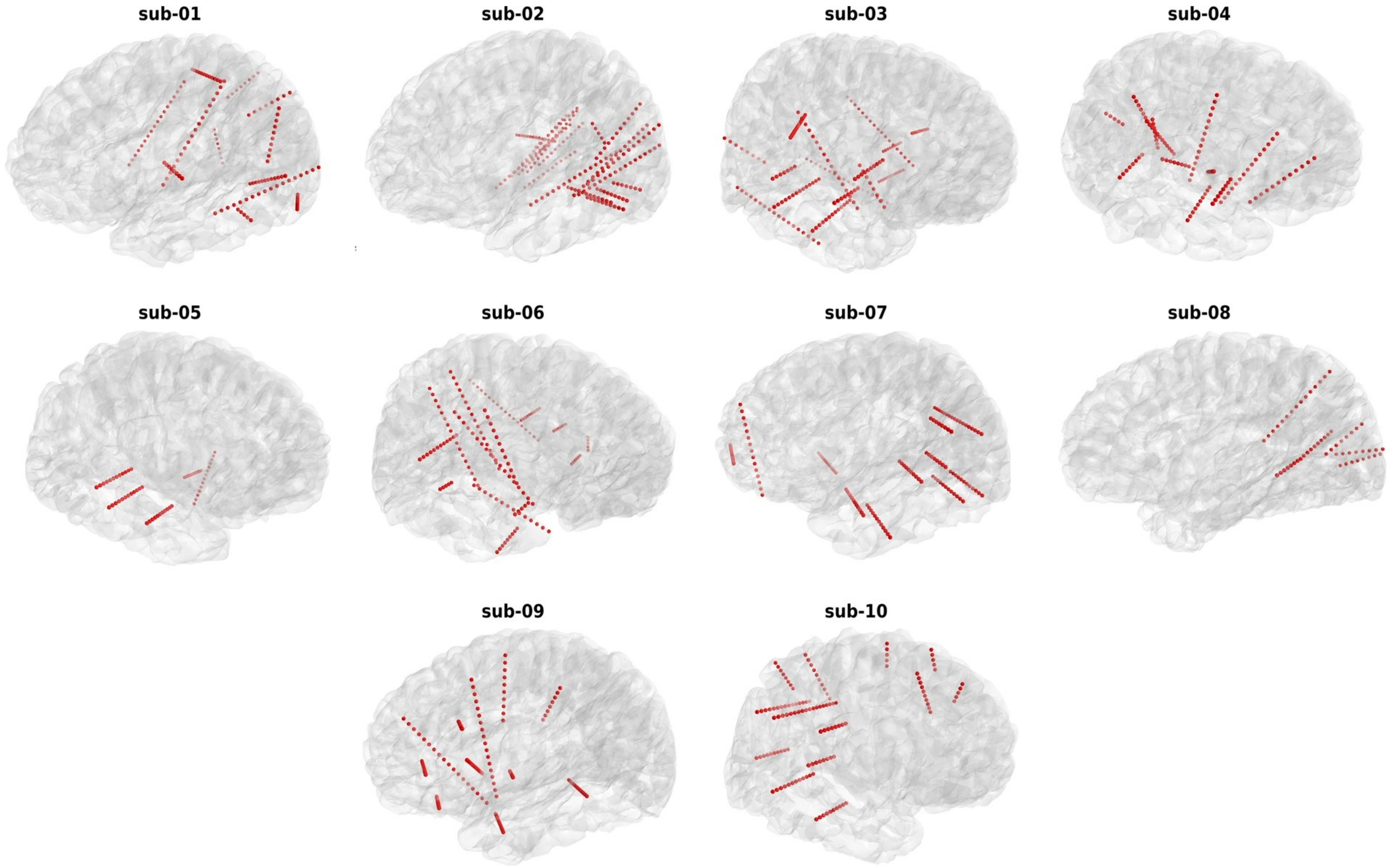


Herff, C. & Verwoert, M. Dataset of Speech Production in intracranial Electroencephalography, Open Science Framework, https://doi.org/10.17605/OSF.IO/NRGX6 (2022).

Participants

➢ Electrode locations of each participant in the surface reconstruction of their native anatomical MRI.

➢ Each red sphere represents an implanted electrode channel.



https://www.nature.com/articles/s41597-022-01542-9

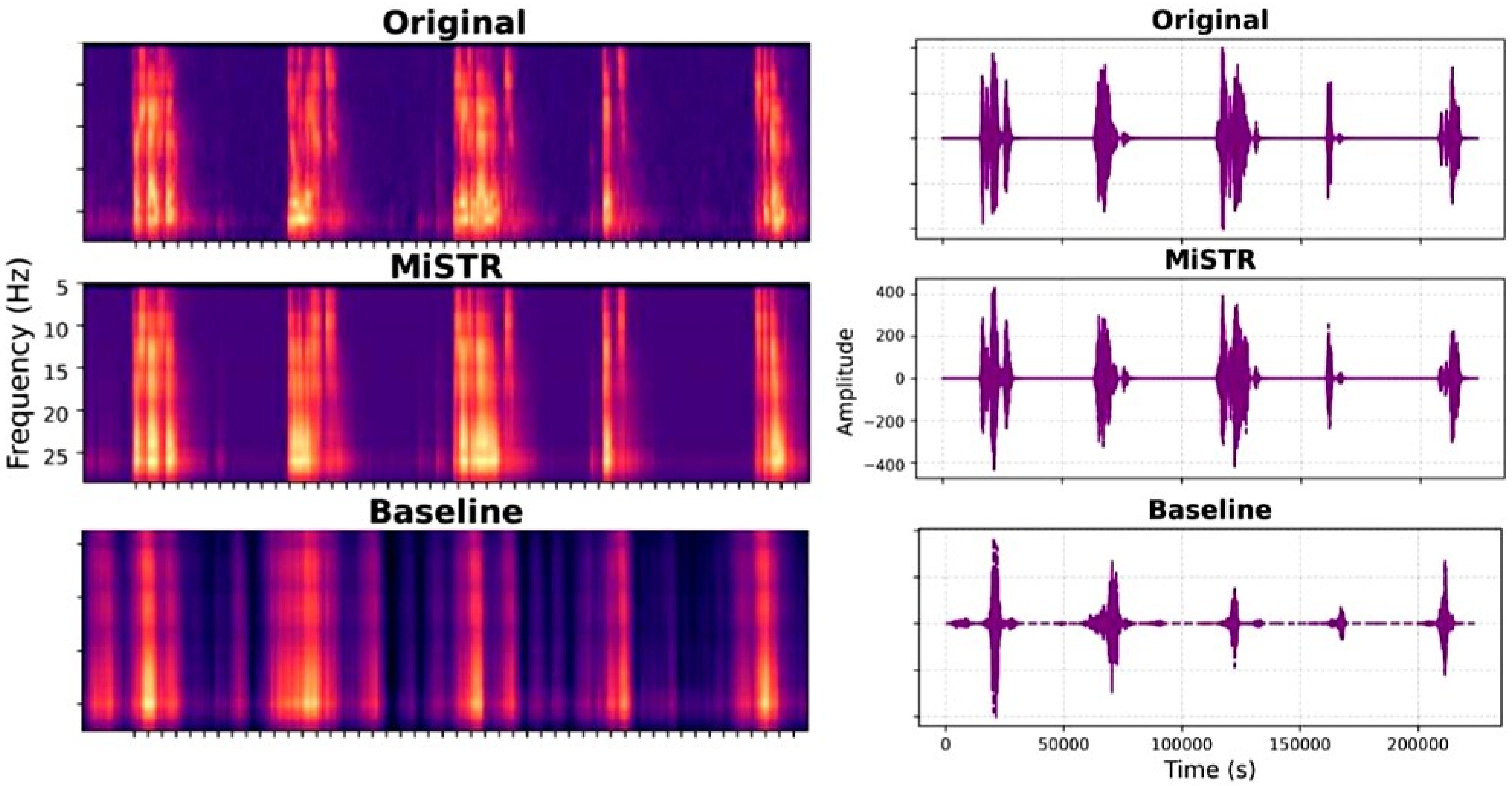
RESULTS

Evaluation Metrics

Model	PC ↑	MCD ↓	STOI ↑	HNR dB ↑	MOSA-Net ↑
Regression [21]	0.72	5.39	0.61	6.2	2.14
bLSTM [8]	0.78	5.23	0.48	8.5	2.12
CNN [20]	0.81	4.95	0.52	10.4	2.41
3D-CNN [19]	0.83	5.04	0.56	9.8	2.57
Seq2Seq [17]	0.85	**3.9**	0.59	10.7	3.21
Encoder-Decoder [11]	0.87	4.34	0.64	11.1	2.82
MiSTR (Ours)	**0.91**	3.92	**0.73**	**12.7**	**3.38**

- Our model, MiSTR, outperforms state-of-the-art baselines across multiple objective measures.
- Significant improvements observed in STOI (intelligibility) and PESQ (perceived quality) scores.
- Demonstrates that integrating prosody and phase refinement yields substantial performance gains.

Visual Comparisons



- MiSTR shows clearer high-frequency structure and stronger harmonic bands vs. baseline.

Conclusion and Future Directions

- ✓ **MiSTR** achieves speech reconstructions that are both intelligible and natural-sounding, outperforming baseline spectrogram-only pipelines.
- ✓ Demonstrated the benefits of combining multi-modal wavelet-based features, prosody-aware Transformers, and a phase-aligned vocoder to reduce artifacts.
- ✓ Validated on real neural speech data, showing improved prosody preservation and harmonic alignment.

❑ **Future Work:**

- Explore end-to-end neural decoding pipelines that bypass intermediate spectrogram prediction.
- Integrate diffusion-based neural vocoders and other generative models for further gains in naturalness.
- Extend to continuous speech and speaker-independent scenarios.

Take-Home Message

- **Combining prosody-aware modeling with harmonic phase refinement is key to bridging the gap between intelligibility and naturalness in brain-to-speech synthesis.**

References

- Angrick, M., Ottenhoff, M. C., Diener, L., Ivucic, D., Ivucic, G., Goulis, S., ... & Herff, C. (2021). Real-time synthesis of imagined speech processes from minimally invasive recordings of neural activity. *Communications biology*, *4*(1), 1055.
- Angrick, M., Ottenhoff, M., Diener, L., Ivucic, D., Ivucic, G., Goulis, S., ... & Herff, C. (2022, May). Towards closed-loop speech synthesis from stereotactic EEG: a unit selection approach. *In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1296-1300). IEEE.
- Herff, C., Diener, L., Angrick, M., Mugler, E., Tate, M. C., Goldrick, M. A., ... & Schultz, T. (2019). Generating natural, intelligible speech from brain activity in motor, premotor, and inferior frontal cortices. *Frontiers in neuroscience, 13*, 1267.
- Herff, C., Krusienski, D. J., & Kubben, P. (2020). The potential of stereotactic-EEG for brain-computer interfaces: current progress and future directions. *Frontiers in neuroscience, 14*, 123.
- Verwoert, M., Ottenhoff, M. C., Goulis, S., Colon, A. J., Wagner, L., Tousseyn, S., ... & Herff, C. (2022). Dataset of speech production in intracranial electroencephalography. *Scientific data, 9*(1), 434.

Thank you

Mohammed Salah Al-Radhi

malradhi@tmit.bme.hu

GitHub: https://github.com/malradhi/MiSTR

Demo : https://malradhi.github.io/MiSTR/

ENFIELD

Happy to collaborate!