Curriculum Cross-Modal Transfer Learning for Imagined Speech Reconstruction from sEEG

Joshua Yin
University of Washington
Seattle, WA, USA
joshjyin@uw.edu

Kevin Luo
University of Washington
Seattle, WA, USA
kevinluoyr@uw.edu

Johnathan Zhang
University of Washington
Seattle, WA, USA
jz050701@uw.edu

Abstract

Imagined speech decoding from stereotactic EEG (sEEG) is hindered by limited data and faint signals, making it challenging to train good speech reconstruction models. We propose a curriculum transfer learning approach to improve reconstruction from sEEG recordings of imagined speech. Our method leverages the VocalMind dataset, which provides parallel sEEG recordings for vocalized, mimed, and imagined Mandarin speech [6]. We first train a neural network on vocalized speech data, then sequentially fine-tune it on mimed speech and finally on imagined speech. Two architectures are explored: a baseline convolutional net plus gated recurrent unit (CNN+GRU) and a CNN plus Transformer sequence model. Results indicate that while cross-modal curriculum transfer learning was unable to improve on the baseline for imagined speech, it was effective in transferring significant learning from vocalized and mimed training to the imagined speech reconstruction task, speeding up convergence during fine-tuning on imagined speech. Further research is needed to determine whether transfer learning from large, rich datasets helps the model actually learn certain features more effectively than a limited imagined baseline.

1. Introduction

Reconstructing intelligible speech from brain activity (colloquially known as "mind-reading") is a long-standing goal of neuroscience. Recent invasive BCIs (brain-computer interfaces) have demonstrated impressive results in decoding overt speech and restoring speech for paralyzed patients [1, 5]. However, decoding imagined speech (internal speech without vocalization) remains challenging due to weaker neural signals and the lack of large, labeled training data. In imagined speech tasks, subjects merely think of words or sentences, producing no audible output for direct supervision. Models trained solely on imagined speech often generalize poorly and produce low-quality reconstructions.

One promising approach to address data scarcity is transfer learning across related speech tasks [4]. Intuitively, vocalized (spoken aloud), mimed (silent mouthing), and imagined speech share underlying neural processes, forming a "nested hierarchy" of speech activity [4]. Studies have shown that neural patterns for overt, mouthed, and imagined speech are related, differing mainly in the presence or absence of articulatory movement and auditory feedback. This suggests that a model trained on overt speech data might capture features and mappings that are partially reusable for silent or imagined speech decoding. If we can leverage the richer vocalized speech data to inform models for imagined speech, we may overcome the generalization issues caused by limited imagined speech samples, or even utilize subtle articulatory, auditory, or other cues in the imagined sEEG that might not otherwise be picked up.

We present a curriculum transfer learning strategy that progressively trains a speech reconstruction model across modes of decreasing observability. We take advantage of and base our models heavily on the recently released VocalMind dataset [6], the first public sEEG dataset to include parallel recordings of vocalized, mimed, and imagined speech in a tonal language (Mandarin Chinese). Our method uses a three-stage training curriculum:

1. **Stage 1 (Vocalized Pre-training)**: Train a model to map sEEG signals to corresponding speech audio (represented as mel-spectrograms) using the vocalized speech data, which has the strongest signals.

2. **Stage 2 (Mimed Fine-tuning)**: Fine-tune the model on mimed speech sEEG data, using the vocalized-mode audio as a surrogate ground truth (since mimed doesn't produce sounds). This adapts the model to silent articulation neural patterns while retaining knowledge from Stage 1.

3. **Stage 3 (Imagined Fine-tuning)**: Further fine-tune

on imagined speech sEEG data (again using the same ground truth as Stage 2) to adapt to purely internal speech.

Through this curriculum, the model incrementally learns to handle the diminishing feedback cues, rather than taking on the hardest task (imagined speech) from scratch. We implement the strategy with two network variants: a baseline 1D CNN+GRU model, and a CNN+Transformer model that can capture longer-range temporal dependencies (in line with prior work [6]). We evaluate our approach on word- and sentence-level tasks from VocalMind, comparing no-transfer baselines (trained only on imagined data) to curriculum transfer models.

Contributions: We introduce a novel application of curriculum-based transfer learning for speech BCI:

- We leverage cross-modal sEEG data (vocalized, mimed, imagined speech) to get a reasonable imagined speech decoding. To our knowledge, this is the first demonstration of curriculum learning across speech modes in an sEEG interface.
- Our curriculum-trained models outperform $<$50 epochs of direct training on imagined speech, yielding more accurate mel-spectrogram reconstructions (lower Mel-Cepstral Distance) and stronger similarity metrics (higher Pearson and DTW-aligned correlations, lower Pitch RMSE, higher Pitch Correlation). However, we fail to outperform a baseline trained to convergence.
- We provide insights into the model's learning progression and discuss how vocalized and mimed speech data contribute to improved representations for silent speech.

2. Related Work

2.1. Speech BCI and Neural Speech Decoding

Decoding speech from brain activity has advanced rapidly in recent years. Using high-density electrocorticography (ECoG) or intracortical arrays, researchers have reconstructed audible speech or text from neural signals of people speaking or attempting to speak [1,5]. Notably, Anumanchipalli et al. [1] synthesized intelligible spoken sentences from ECoG signals by decoding articulatory representations and mapping them to a vocoder. More recently, Metzger et al. [5] demonstrated a full-stack neural speech prosthesis that decoded attempted speech in a paralyzed patient at 78 words per minute, driving a text-to-speech avatar. These works underscore the potential of invasive BCIs for restoring communication. Stereotactic EEG (sEEG), which uses depth electrodes implanted for epilepsy monitoring, is a less-explored recording modality for speech BCI [2]. Arthur and Csapó [2] reported one of the first studies reconstructing speech from sEEG, using a neural vocoder to synthesize audio from intracranial signals. The VocalMind dataset introduced by He et al. [6] is a significant contribution in this area, providing over one hour of sEEG recordings with aligned speech data in Mandarin. Their baseline models showed that decoded mel-spectrograms from sEEG can achieve high similarity to the original speech, validating the dataset's quality.

2.2. Imagined and Silent Speech Decoding

Mimed speech (articulation without sound) falls between speaking and pure imagination. Both mimed and imagined speech lack auditory output, complicating training and evaluation since there is no directly recorded audio. Previous studies have sought evidence of decodable signals in these modes. Angrick et al. [3] achieved real-time synthesis of imagined phrases from ECoG-like minimally invasive recordings, although with limited vocabulary. Other works have attempted imagined speech decoding with non-invasive EEG, but reconstructing intelligible speech from scalp signals has proven very difficult. For example, a recent EEG-based study by Xiong et al. [7] introduced an EEG-to-speech model that produced some understandable words from imagined EEG, by aligning imagined EEG with actual speech EEG through dynamic time warping. Their results, while promising, highlight that additional innovations (like transfer learning and alignment) are needed to handle the noisier signals in imagined speech tasks.

2.3. Transfer Learning and Curriculum Learning in BCIs

Transfer learning can be used in BCIs to handle data scarcity, by pre-training on larger datasets or related tasks. In the context of speech BCI, researchers have considered multi-task learning or transfer between speech perception and production, or between different subjects, to improve generalization [4]. The VocalMind dataset itself was designed to facilitate cross-mode transfer learning research [6]. Our approach is inspired by curriculum learning, where a model is trained on easier subtasks before tackling the hardest task. Curriculum strategies have been effective in domains like computer vision and natural language, but is only beginning to be explored for neural decoding. To our knowledge, our work is the first to apply a curriculum across speech modes (vocalized$\rightarrow$mimed$\rightarrow$imagined) in invasive BCI. It is conceptually related to Xiong et al.'s EEG study [7] and other heterogeneous transfer learning approaches in imagined speech BCIs (e.g., EEG-based word classification via transferred features).

3. Methods

3.1. Dataset and Preprocessing

We use the VocalMind sEEG dataset [6] and identical pre-processing, focusing on its parallel recordings of vocalized, mimed, and imagined speech. In this dataset, a native 22-year Mandarin-speaking male participant with stereotatic-EEG implants performed multiple trials of reading words and sentences. Each trial is labeled with the target phrase and has an associated time-aligned audio recording for the vocalized condition. In total the dataset provides 20 words and a set of 100 sentences (each with 5-8 characters, repeated twice per speech-mode), with over one hour of sEEG data across all conditions. Because mimed and imagined trials produce no sound, the vocalized audio for the same prompted phrase is used as a proxy ground truth for training and evaluation. All audio was sampled at 48kHz and downsampled to 16 kHz, then converted to mel-spectrograms with 80 mel filterbank channels (the same setup as in the VocalMind baseline). Each mel-spectrogram frame corresponds to $\sim$20 ms of audio, yielding spectrogram sequences on the order of a few hundred time-steps for each word or sentence.

Raw sEEG signals from each implanted electrode contact were preprocessed following the procedure of He et al. [6]. This includes common preprocessing steps such as band-pass filtering the neural signals (e.g., 70–150 Hz) and downsampling (e.g., to 200 Hz) to reduce high-frequency noise. Artifact removal was also performed to ensure data quality. We further downsample the sequence in time when feeding it to the neural network by applying a 1-D convolution with stride 4 as the first network layer, reducing the input length by 4$\times$, roughly matching the ratio of original signal sampling rate to mel frame rate. The sEEG features are standardized (z-scored) per electrode.

Speech audio for each trial (for vocalized trials, the recorded audio; for silent trials, the corresponding vocalized audio of that phrase) is converted to an 80-dimensional mel-spectrogram using a short-time Fourier transform (STFT) with 64 ms window length and 20 ms hop. During training, we minimize the Mean Squared Error (MSE) between the predicted and ground-truth mel-spectrogram frames.

During evaluation, we synthesize waveforms from the predicted mel-spectrograms with a pre-trained HiFi-GAN vocoder [6]. From the waveforms, we then compute the Mel Cepstral Distortion and extract the fundamental frequency (f_0) required for the Pitch RMSE and Pitch Correlation metrics.

3.2. Neural Network Architecture

Our model is based on the baseline architecture described in the VocalMind paper [6]. The input to the network is a sequence of multi-channel sEEG features (time $\times$ channels). The network processes these inputs as follows:

Convolutional Encoder: A one-dimensional conv layer (Conv1D) with 64 output channels, filter size 4, stride 4, and padding 2 acts as the encoder. This layer reduces the temporal resolution by a factor of 4 while increasing the feature dimension. We use a ReLU activation followed by dropout (dropout rate 0.7). The output is flattened to form a sequence of feature vectors for the recurrent layer.

Recurrent Decoder (GRU): The core sequence modeling is performed by a 3-layer bidirectional GRU with 256 hidden units per direction (512 total per layer when bidirectional). The input size to the GRU is 64 (matching the Conv1D output channels). A dropout of 0.7 is applied between GRU layers. The bidirectional GRU allows the model to utilize both past and future context.

Output Layer: A fully connected layer maps the final GRU layer's output at each time step to an 80-dimensional output, corresponding to the mel-spectrogram frame at that time step. We use a linear activation (i.e., no nonlinearity) on the output since we are predicting continuous spectrogram values.

The total number of trainable parameters in this CNN+GRU model is on the order of a few million, which is modest enough to train on $\sim$1 hour of data without severe overfitting. We refer to this architecture as CNN+GRU in our experiments.

CNN+Transformer Variant: In addition to the GRU-based model, we experimented with replacing the recurrent layers with a Transformer encoder. This model uses the same Conv1D encoder for initial feature extraction. After that, instead of GRUs, we use a stack of Transformer encoder blocks (3 layers, 4 heads, embedding dim 64, feed-forward dim 256). The output from the last Transformer layer at each time step is passed through a linear layer to predict the 80-dim mel-spectrogram frame. We apply dropout with rate 0.7 in the Transformer. We call this model CNN+Transformer in experiments.

Learning Rate and Progressive Unfreezing Schedule: After tuning, we decided on the following hyperparameters for training both the CNN+GRU and CNN+transformer models: Stage 1 was trained for 100 epochs with a learning rate of 1×10^{-3}, Stage 2 for 20 epochs with a learning rate of 5×10^{-4}, and Stage 3 for 10 epochs with a learning rate of 1×10^{-5}. The Stage 2/3 unfreeze schedule for both the GRU and transformer is expressed as $[0, 10, 20]$ to indicate that the topmost layer is unfrozen at epoch 0, the middle layer unfrozen at epoch 10, and the bottom layer at epoch 20 (so for imagined fine-tuning only the top layer is unfrozen and for mimed fine-tuning, the top two layers are progressively unfrozen).

3.3. Evaluation Metrics

Following the original VocalMind study, we report the *same five objective metrics* used to assess mel-spectrogram reconstruction:

- **Mel-Cepstral Distance (MCD)** [6]: Quantifies spectral differences between predicted and reference spectrograms; lower values indicate higher similarity.
- **Pearson Correlation Coefficient (Correlation)**: Measures linear correlation between predicted and ground-truth spectrograms; higher (closer to 1) is better.
- **DTW-Aligned Correlation (DTW Correlation)**: For mimed and imagined speech, the prediction is aligned to the corresponding *vocalized* spectrogram with Dynamic Time Warping before computing Pearson correlation; higher is better.
- **Pitch Root-Mean-Square Error (Pitch RMSE)**: After DTW alignment of fundamental-frequency (f_0) contours extracted with the DIO algorithm, lower RMSE indicates better pitch accuracy.
- **Pitch Correlation**: Pearson correlation between the aligned f_0 contours; higher is better.

Results for all five metrics are shown in the Appendix section.

4. Experiments

4.1. Experiment Setup

We perform separate experiments for word-level and sentence-level decoding. 6-fold cross-validation is used as in the VocalMind baseline [6]: for words, ensure each fold's test set contains one instance of each word; for sentences, each test set contains a subset of sentences not seen in training.

All models were implemented in PyTorch and trained on an NVIDIA H100 GPU. Curriculum models train in three stages (vocalized $\rightarrow$ mimed $\rightarrow$ imagined) within each fold. For baselines, we train the same architectures on imagined speech only, and utilize early-stopping.

Model Conditions:

- Imagined-Only (Baseline): CNN+GRU trained from scratch on imagined data.
- Curriculum (Vocal$\rightarrow$Mimed$\rightarrow$Imagined): CNN+GRU with staged transfer learning.
- Imagined-Only (Transformer): CNN+Transformer trained from scratch on imagined data.
- Curriculum (Transformer): CNN+Transformer with staged transfer.
- Vocalized-Pretrained Only (GRU): CNN+GRU pretrained on vocalized, directly fine-tuned on imagined (no mimed stage).
- Vocalized-Pretrained Only (Transformer): CNN+Transformer pre-trained on vocalized, directly fine-tuned on imagined (no mimed stage).

All models are evaluated on the imagined speech test sets for final comparison.

4.2. Results

The curriculum transfer models nearly match the performance of the direct imagined-only baselines on sentences in terms of correlation and DTW correlation. The transformer model also nearly matched the baseline for pitch correlation, while the GRU model was further off. However, both the GRU and transformer curriculum models underperformed the baseline significantly in terms of MCD and Pitch RMSE (the latter is of particular importance for languages like Mandarin).

Imagined word reconstruction has similar trends, but the baseline transformer completely flopped in both types of correlation, presumably because the transformer architecture is not optimized for shorter sequences like this. It is interesting to note that the curriculum transfer learning helped rectify this abysmal performance from the transformer, which suggests that the vocalized/mimed pre-training may be useful in helping the transformer understand imagined context in a more nuanced way.

Finally, we note that during training the curriculum model tended to outperform the baseline trained on $<$50 epochs. Since the curriculum only fine-tunes on the imagined corpus for 10 epochs at a very low learning rate, this suggests that there is significant cross-modal transfer learning occurring, enough so that even with a totally frozen CNN we can achieve results comparable to the baseline. It is an open question whether the cross-modal freezing preserved features that would otherwise not be learned by the baseline itself in later epochs, but the significantly improved correlation of the transformer seems to suggest it might. More research is needed into whether or not this curriculum can be leveraged on new, rich, datasets to support models that learn better than a limited baseline.

5. Discussion

Curriculum transfer learning substantially improves imagined speech reconstruction from brain activity, achieving comparable performance to a model trained solely on imagined data—while requiring significantly fewer epochs

for fine-tuning. By leveraging data from overt speech (vocalized) and covert speech (mimed) as intermediate training steps, the model gains a stronger initial mapping of neural signals to acoustics. This is especially valuable given the limited and noisier data available for pure imagination. The improvements suggest that the model trained with our curriculum may be better at extracting and preserving latent speech content present in the sEEG signals, even when no sound is produced.

Our findings align with neuroscience evidence that covert speech shares representational structure with overt speech [4]. Incorporating the mimed stage proved surprisingly inconsequential—skipping it led to no significant change in performance, suggesting that silent articulations do not provide crucial bridging information.

It is worth noting, however, that VocalMind's data is collected from a single subject, whereas real-world BCIs require generalization across users. Whether training a model using sEEG data collected from multiple speakers with a curriculum approach—and then adapting it to a new user's imagined speech data—can enable effective transfer remains an open question. Still, collecting sEEG data remains a challenging and resource-intensive process to collect sEEG data, let alone find participants willing to contribute large volumes of data.

One promising direction for improving generalization is to incorporate unsupervised pre-training on a large corpus of resting-state sEEG. By initializing the encoder with representations learned from this broader neural activity we may be able to enhance downstream performance during speech fine-tuning.

6. Conclusion

We presented a curriculum transfer learning method to improve imagined speech reconstruction from sEEG signals, by sequentially training on vocalized, mimed, and imagined speech data. In evaluations on the VocalMind dataset, our approach yielded similar spectrogram and audio reconstruction performance compared to models trained solely on imagined speech. These results highlight that knowledge acquired from overt speech can be successfully leveraged to decode covert and internal speech, addressing a key hurdle in speech BCI development.

Acknowledgments

We would like to thank Professor Ranjay Krishna for introducing us to the intricate and profound world of deep learning and for being an amazing professor and mentor!

References

[1] G. K. Anumanchipalli, J. Chartier, and E. F. Chang. Speech synthesis from neural decoding of spoken sentences. *Nature*, 568:493–498, 2019. 1, 2

[2] F. V. Arthur and T. G. Csapó. Speech synthesis from intracranial stereotactic eeg using a neural vocoder. *Infocommunications Journal*, 16:47–55, 2024. 2

[3] M. Angrick et al. Real-time synthesis of imagined speech processes from minimally invasive recordings of neural activity. *Commun. Biol.*, 4(1055), 2021. 2

[4] P. Z. Soroush et al. The nested hierarchy of overt, mouthed, and imagined speech activity evident in intracranial recordings. *NeuroImage*, 269:119913, 2023. 1, 2, 5

[5] S. L. Metzger et al. A high-performance neuroprosthesis for speech decoding and avatar control. *Nature*, 620:1037–1046, 2023. 1, 2

[6] T. He et al. Vocalmind: A stereotactic eeg dataset for vocalized, mimed, and imagined speech in tonal language. *Sci. Data*, 12(657), 2025. 1, 2, 3, 4

[7] W. Xiong, L. Ma, and H. Li. Synthesizing intelligible utterances from eeg of imagined speech. *Front. Neurosci.*, 19, 2025. 2

Appendix

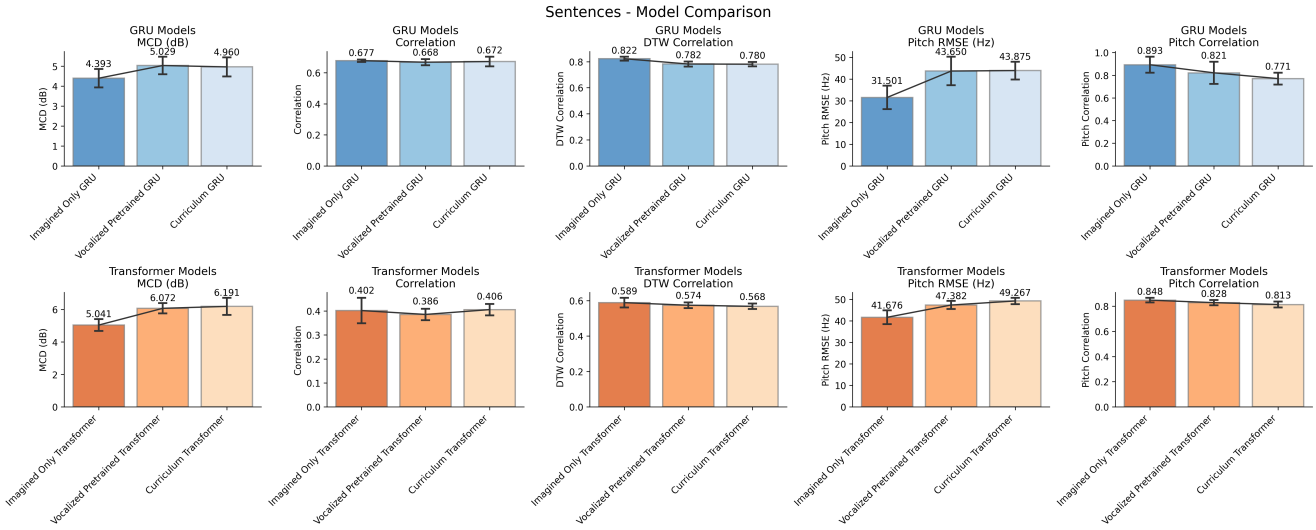

Figure 1. Performance comparison between GRU and Transformer models on sentence-level decoding, based on evaluation metrics across vocalized, mimed, and imagined speech.

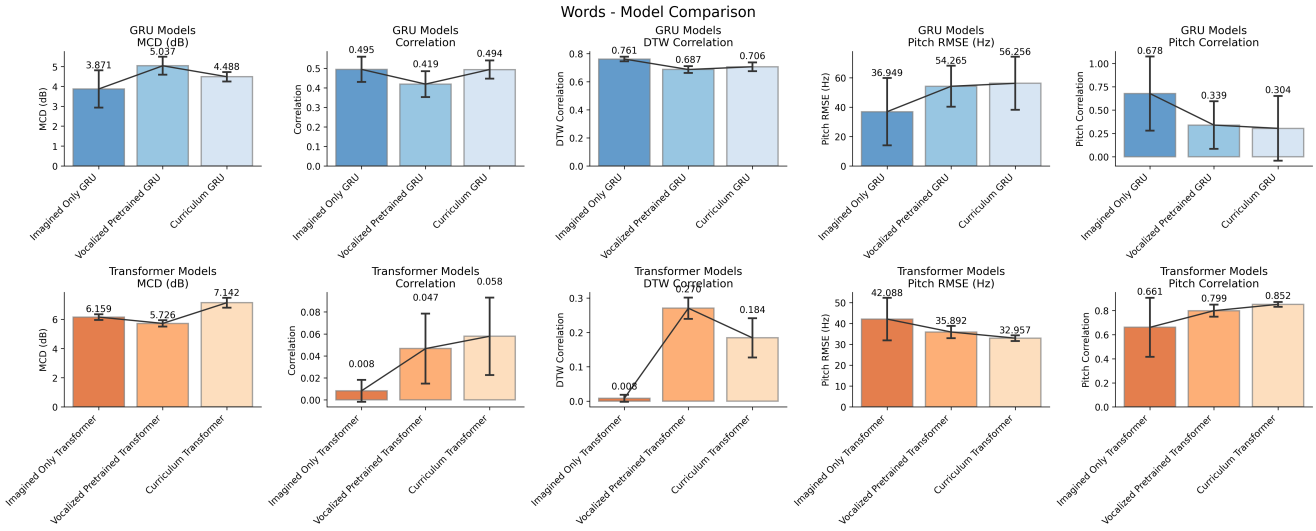

Figure 2. Performance comparison between GRU and Transformer models on word-level decoding, based on evaluation metrics across vocalized, mimed, and imagined speech.

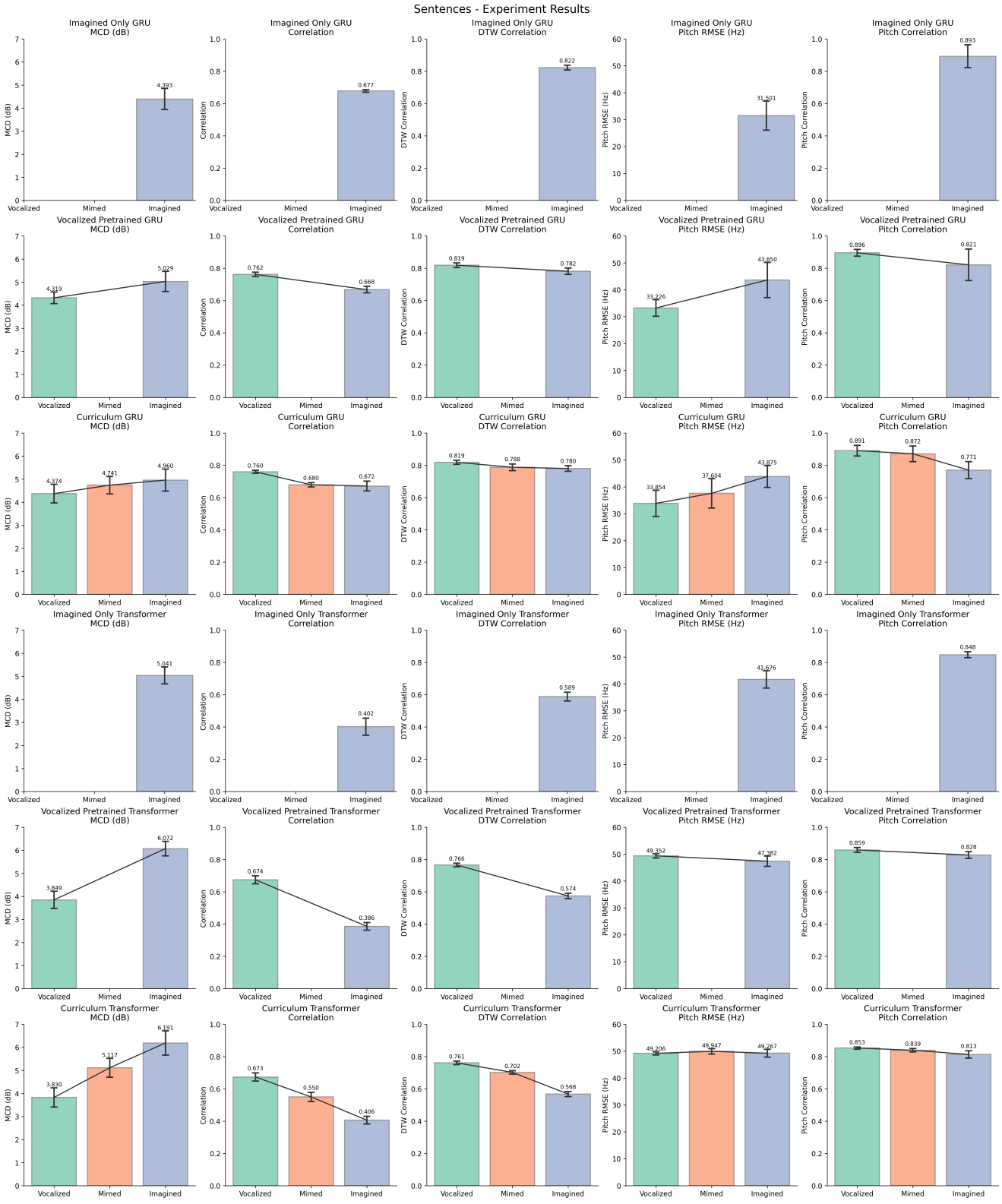

Figure 3. The evaluation metrics for each of the 6 models on sentence-level decoding, shown across vocal, mimed, and imagined speech.

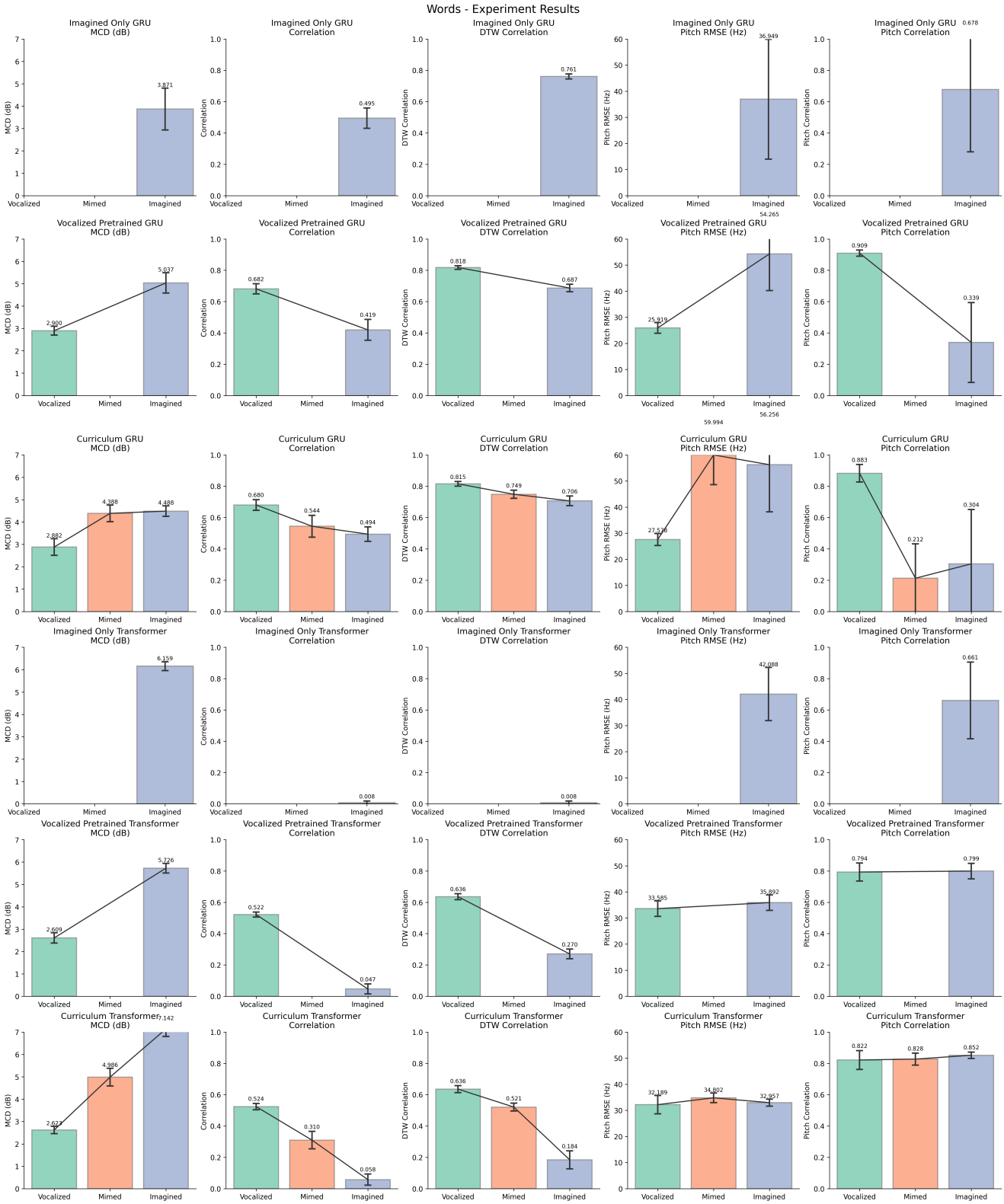

Figure 4. The evaluation metrics for each of the 6 models on word-level decoding, shown across vocal, mimed, and imagined speech.