Gesture2Speech: How Far Can Hand Movements Shape Expressive Speech?

Anonymous submission

Abstract

Human communication seamlessly integrates speech and bodily motion, where hand gestures naturally complement vocal prosody to express intent, emotion, and emphasis. While recent text-to-speech (TTS) systems have begun incorporating multimodal cues such as facial expressions or lip movements, the role of hand gestures in shaping prosody remains largely underexplored. We propose a novel multimodal TTS framework, Gesture2Speech, that leverages visual gesture cues to modulate prosody in synthesized speech. Motivated by the observation that confident and expressive speakers coordinate gestures with vocal prosody, we introduce a multimodal Mixture-of-Experts (MoE) architecture that dynamically fuses linguistic content and gesture features within a dedicated style extraction module. The fused representation conditions an LLM-based speech decoder, enabling prosodic modulation that is temporally aligned with hand movements. We further design a gesture-speech alignment loss that explicitly models their temporal correspondence to ensure fine-grained synchrony between gestures and prosodic contours. Evaluations on the PATS dataset show that Gesture2Speech outperforms state-of-the-art baselines in both speech naturalness and gesture-speech synchrony. To the best of our knowledge, this is the first work to utilize hand gesture cues for prosody control in neural speech synthesis. Demo samples are provided at URL: https://tinyurl.com/3wv58sbw.

Introduction

Expressive speech synthesis is essential in applications such as dubbing educational content, podcasts, talk shows, and interviews, where intelligibility, natural prosody, and temporal alignment are critical for effective communication (Brannon, Virkar, and Thompson 2023). Speakers rely on nonverbal bodily cues, particularly hand gestures to convey emphasis, rhythm, and affective intent. These gestures exhibit tight temporal and emotional coordination with speech rhythm and tone, making them a rich source of prosodic and affective cues (Wagner, Malisz, and Kopp 2014). While neural TTS systems produce high-quality, intelligible speech, they still lack embodied prosodic control for expressive, multimodal communication (Hu et al. 2021; Sahipjohn et al. 2024). Current models infer prosody from text or reference audio, limiting their ability to capture the richness of human expression (Han et al. 2025; Casanova et al. 2024; Shimizu et al. 2024). Given the multimodal nature of communication, TTS systems should leverage cues beyond text and audio (Jiménez-Bravo and Marrero-Aguiar 2024; Zhang et al. 2021). Hand gestures remain an underexplored modality despite offering valuable prosodic cues. Their temporal synchrony with pitch accents, emphasis, and duration reflects speaker intent and expressive style (Feyereisen and De Lannoy 1991). However, the relationship between gesture and prosody is complex and speaker-dependent. Gesture intensity and timing may not always correlate with prosodic prominence such as pitch accents or energy peaks. Nonetheless, incorporating gesture cues into TTS can improve prosody modeling and produce temporally aligned, expressive speech, especially in dubbing and conversational scenarios.

In prior work, the gesture modality has primarily been leveraged for applications such as sign language recognition and translation, human-robot interaction, and gesture generation (Madhiarasan and Roy 2022; Papastratis et al. 2021; Li, Zhong, and Wang 2023). The generation of gestures from speech, commonly referred to as co-speech gesture generation, has received significant attention (He et al. 2024; Ahuja et al. 2020b). More recently, multimodal generation frameworks have explored the simultaneous synthesis of speech and gestures in an integrated manner (Mehta et al. 2024, 2023). However, the reverse paradigm, where gestures are used as a modality to generate prosodically controlled speech in TTS, remains largely underexplored. While our experiments are conducted using the PATS dataset, which contains high-quality multimodal recordings with aligned gesture and speech, we acknowledge its limited cultural and emotional scope. in real-world scenarios, full-body visibility or high-resolution hand tracking may not always be available. Our framework is designed to process full pose keypoints, relying on upper-limb dynamics.

In this paper, we propose Gesture2Speech, a multimodal TTS framework that integrates gesture input alongside text, speech, and motion-derived video cues to generate expressive speech aligned with gestural intent. Unlike conventional TTS systems that primarily rely on textual and prosodic features, Gesture2Speech treats hand gestures as dynamic style control signals, enabling more grounded and contextually synchronized speech synthesis. To effectively model the varying contributions of different modalities, we extend a Mixture-of-Experts (MoE) architecture (Jacobs et al. 1991), the novelty of our approach lies in applying MoE to dynamically select

experts conditioned on gestural input in a speech synthesis task, incorporating specialized expert modules for speaker style and speaker-specific visual motion features. Inspired by recent advances in style-disentangled expressive TTS (Jawaid et al. 2024) and gesture animation (Ahuja et al. 2020b), our multimodal MoE design facilitates fine-grained control over generated speech while preserving speaker identity and expressiveness.

Our contributions lie not only in the technical novelty of multimodal conditioning and expert specialization but also in drawing attention to gesture-conditioned speech synthesis, a relatively underexplored research area. Our key contributions are as follows.

- We introduce a novel framework for prosody modeling in expressive TTS, where hand gestures are used as control signals to guide speech synthesis.
- We propose a multimodal Mixture of Experts (MoE) architecture that integrates hand gesture and audio features to learn rich, disentangled style representations.
- These learned representations condition an LLM-based speech decoder, enabling the generation of speech that is temporally aligned with gestural cues.
- We propose a gesture-speech alignment loss to explicitly model and enhance the temporal synchronization between gesture dynamics and speech prosody.

Related Work

Despite progress in neural TTS, fine-grained and interpretable prosody control remains challenging. Most systems still struggle with prosodic variability and expressiveness without explicit control. Early unimodal approaches, such as Tacotron (Shen et al. 2018) and its extensions, aimed to control prosody using textual or reference audio prompts, while models like GST-Tacotron (Wang et al. 2018) and FastSpeech (Ren et al. 2019) introduced style tokens or predicted prosodic features (e.g., pitch, duration) directly from text. These methods offered limited controllability and largely ignored the affective context underlying expressive delivery. To address this limitation, more recent works have explored multimodal prosody modeling, incorporating cues such as facial expressions and lip movements to enhance expressiveness in TTS (Chu, Shim, and Park 2024; Lu et al. 2022; Sahipjohn et al. 2024). However, hand gestures, an essential bodily cue that co-varies with speech prosody and emotion, have been largely overlooked as a control modality in speech synthesis. In contrast to facial or lip motion, gestures convey intent, rhythm, and affective emphasis through larger, rhythmically aligned movements, making them a promising but underexplored source of prosodic information.

Speech Generation via Gestures

The interplay between gestures and speech has long intrigued researchers in multimodal communication. Early studies focused primarily on gesture generation conditioned on speech (Alexanderson et al. 2020), while more recent work investigates integrated speech and gesture generation (Nyatsanga et al. 2023; Wang et al. 2021; Zhang et al. 2025). Alexanderson and Székely (Alexanderson et al. 2020) proposed a framework that jointly generates spontaneous speech and gesture from text, demonstrating the tightly coupled nature of these modalities. However, gestures were not directly used to modulate acoustic parameters. More unified frameworks, such as those by Mehta et al. (Mehta et al. 2024, 2023), used flow matching to synthesize both gestures and speech from a shared latent space, hinting at the potential for bidirectional gesture-conditioned speech synthesis. Nevertheless, these approaches remain largely exploratory and do not explicitly target fine-grained prosodic control. Our work diverges from these by directly using gesture motion features as conditioning signals for prosody generation, enabling tighter temporal and expressive alignment between physical motion and synthesized speech. This formulation bridges the gap between multimodal modeling and embodied expressivity.

Mixture of Experts for Style Transfer in TTS

The Mixture of Experts (MoE) paradigm has gained traction in TTS for capturing diverse speaking styles and emotional nuances. By allocating responsibility across specialized expert networks, MoEs facilitate nuanced and adaptive control over prosodic features. Jawaid et al. (Jawaid et al. 2024) introduced a Style-MoE architecture that learns expressive speech synthesis via multiple style embeddings, enabling smoother transitions between speaking styles. Similarly, AdaSpeech 3 (Yan et al. 2021) models spontaneous and conversational speech through adaptive expert components, while Teh and Hu (Teh et al. 2023) explored ensemble-based prosody prediction as a mixture framework for expressive intonation control. Building on these insights, our framework integrates modality-specific MoE modules that fuse linguistic, acoustic, and gestural representations within a unified style space. Unlike prior MoE-based systems focused purely on speaker or style variation, our design explicitly leverages gesture-driven cues to enhance temporal alignment and affective prosody generation. This integration extends the MoE paradigm toward embodied multimodal expressivity, a key goal for human-like TTS systems.

Proposed Method

Problem Formulation

Given an input text sequence $\mathcal{T}$, a reference audio sample $\mathcal{A}_{\text{ref}}$ from a target speaker, and a sequence of gesture frames $\mathcal{V} = \{\mathcal{V}_t\}_{t=1}^{T}$, the goal is to synthesize a speech waveform $\hat{\mathcal{A}}$ that is semantically aligned with $\mathcal{T}$, retains the identity of the speaker from $\mathcal{A}_{\text{ref}}$, and reflects the temporal prosody driven by gestures in $\mathcal{V}$. We model this as a conditional generation problem with mapping $\mathcal{F}_\theta : (\mathcal{T}, \mathcal{A}_{\text{ref}}, \mathcal{V}) \mapsto \hat{\mathcal{A}}$, where the function $\mathcal{F}_\theta$ is trained to maximize the likelihood $p_\theta(\hat{\mathcal{A}} \mid \mathcal{T}, \mathcal{A}_{\text{ref}}, \mathcal{V})$. The synthesized speech must preserve linguistic content, speaker characteristics, and exhibit prosodic variation synchronized with gesture dynamics, encouraging a tightly coupled multimodal alignment across text, audio, and vision domains.

Proposed Architecture: Gesture2Speech TTS

Here, we propose a gesture-conditioned text-to-speech (TTS) system that synthesizes expressive speech conditioned not

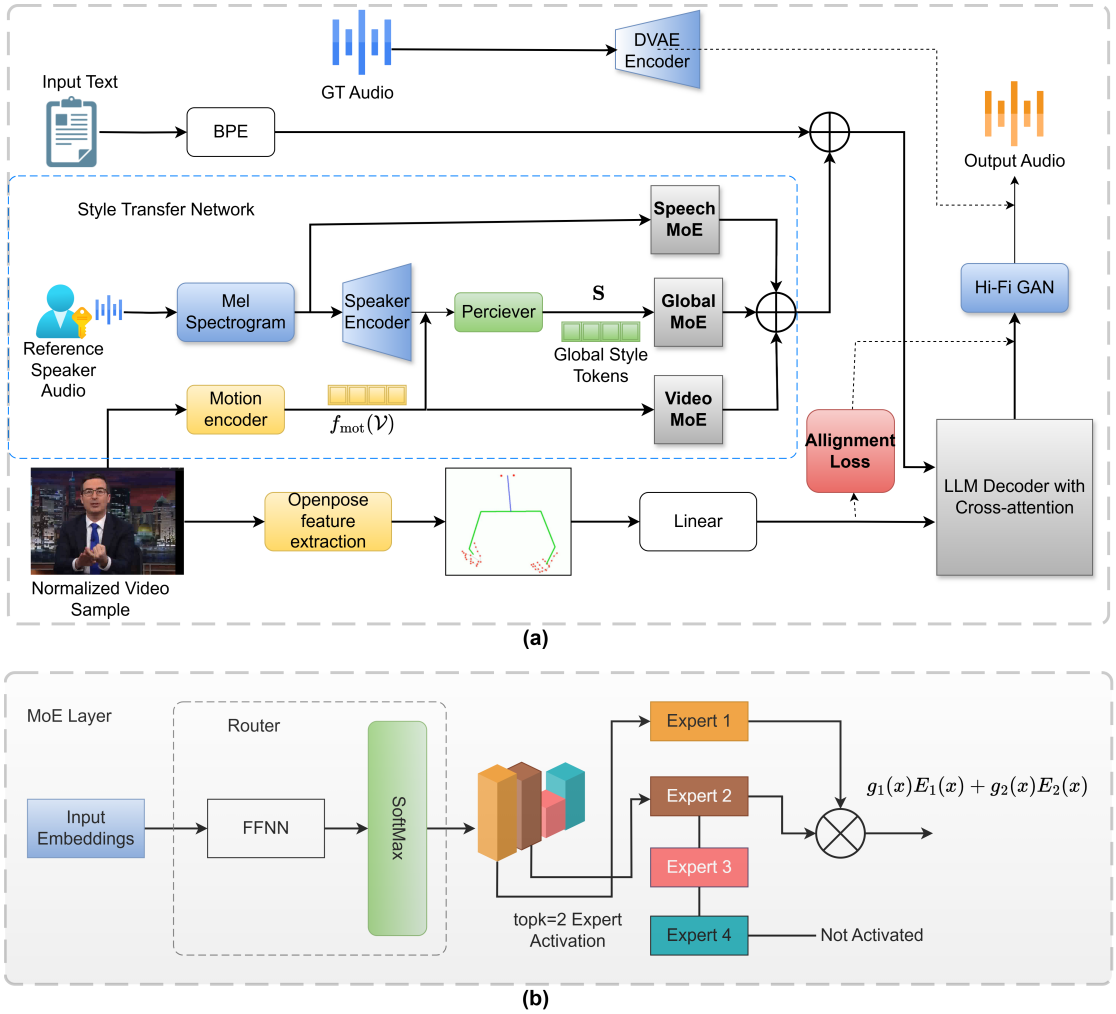

Figure 1: Overview of (a) the proposed Gesture2Speech architecture and (b) MoE layer. The system takes text, speech, and video-based gesture features as input and generates expressive speech. Multiple MoE modules enable dynamic routing of features for improved style representation and aligned with gestural intent via cross-attention.

only on textual input but also on hand gestures and motion cues derived from video. By incorporating visual-semantic information, our model aligns speech prosody with the temporal dynamics of gestures. An overview of the proposed framework is illustrated in Figure 1. The model operates on three input modalities: (1) textual features, (2) audio embeddings, and (3) motion and pose features extracted from video. Text and audio inputs are processed through a shared encoder, while motion features are handled by a dedicated visual encoder. To achieve temporal alignment and feature compression, we employ a perceiver resampler. All modalities are projected into a shared latent space of dimension $d = 1024$ to facilitate effective cross-modal fusion.

The input text $\mathcal{T}$ is first tokenized using the byte-pair encoding (BPE), producing a sequence of embeddings $\mathbf{E}_{\text{text}} \in \mathbb{R}^{L\times d}$. From the reference audio, a mel-spectrogram is computed and passed through a speaker encoder f_{spk} to obtain a speaker embedding $\mathbf{e}_{\text{spk}} \in \mathbb{R}^d$ and normalized video frames $\mathcal{V}$ are processed by a SlowFast (Zhang, Tie, and Qi 2021) motion encoder f_{mot} to produce spatiotemporal features $\mathbf{M} \in \mathbb{R}^{T\times d}$:

$$\mathbf{e}_{\text{spk}} = f_{\text{spk}}(\text{Mel}(\mathcal{A}_{\text{ref}})); \quad \mathbf{M} = f_{\text{mot}}(\mathcal{V}) \tag{1}$$

These motion features are concatenated with the broadcasted speaker embedding and fed into a Perceiver module to generate global style tokens $\mathbf{S} \in \mathbb{R}^{N\times d}$:

$$\mathbf{S} = \text{Perceiver}([\mathbf{M} \parallel \mathbf{e}_{\text{spk}}]). \tag{2}$$

To model modality specific characteristics, three MoE modules are applied to the speaker embedding (i.e., Speech MoE), motion features (i.e., Video MoE), and global style tokens (i.e., Global MoE):

$$\mathbf{z}_{\text{speech}} = \text{MoE}_{\text{speech}}(\mathbf{e}_{\text{spk}}); \quad \mathbf{z}_{\text{motion}} = \text{MoE}_{\text{motion}}(\mathbf{M}); \\ \mathbf{z}_{\text{style}} = \text{MoE}_{\text{style}}(\mathbf{S}) \tag{3}$$

The outputs are concatenated to form a fused style representation:

$$\mathbf{z}_{\text{style-total}} = [\mathbf{z}_{\text{speech}} \parallel \mathbf{z}_{\text{motion}} \parallel \mathbf{z}_{\text{style}}] \tag{4}$$

Gesture features are extracted using OpenPose (Cao et al. 2021) included in experimental dataset, to obtain 2D keypoints $\{\mathbf{K}_t\}_{t=1}^{T}$, with each $\mathbf{K}_t \in \mathbb{R}^{J\times 2}$ representing J joints. These are flattened and projected to latent vectors using a learnable linear mapping, resulting in a gesture token sequence $\mathbf{G} \in \mathbb{R}^{T\times d}$.

The LLM decoder receives the concatenation of text embeddings $\mathbf{E}_{\text{text}}$ and fused style tokens $\mathbf{z}_{\text{style-total}}$, and gesture tokens $\mathbf{G}$ as input. We have used an LLM-based decoder with cross attention:

$$\hat{\mathbf{v}} = \text{LLM}_{\text{cross}}([\mathbf{E}_{\text{text}} \parallel \mathbf{G} \parallel \mathbf{z}_{\text{style-total}}]). \tag{5}$$

The output token sequence $\hat{\mathbf{v}}$ is decoded by a HiFi-GAN (Kong, Kim, and Bae 2020)[1] vocoder to produce the final waveform:

$$\hat{\mathcal{A}} = \text{HiFi-GAN}(\hat{\mathbf{v}}). \tag{6}$$

This design enables expressive, gesture-aware speech generation that respects both motion dynamics and speaker-specific prosody.

Style Transfer with Mixture-of-Experts (MoE)

To effectively capture modality-specific style information, we incorporate a sparse Mixture-of-Experts module (Jacobs et al. 1991) into the fusion pipeline. Specifically, we deploy three distinct MoE modules, each for the conditional audio embeddings, video features, and the fused representation. Each module adopts an expert routing mechanism, enabling dynamic and data-dependent expert selection during both training and inference.

Let $x_{\text{audio}} \in \mathbb{R}^{A\times d}$, $x_{\text{video}} \in \mathbb{R}^{V\times d}$, and $x_{\text{fused}} \in \mathbb{R}^{S\times d}$ denote the inputs to the speech, video, and global MoEs respectively. Each MoE transforms the input using a gated expert network:

$$\text{MoE}(x) = \sum_{i=1}^{K} g_i(x) E_i(x), \tag{7}$$

Where E_i is the i^{th} expert, and $g_i(x)$ is the gating function determining the contribution of expert i for input x. The outputs from all three MoEs are concatenated and passed to the LLM decoder along with text embeddings and openpose output embeddings. The resulting fused embeddings are then used to predict expressive prosodic features, optimized jointly using reconstruction and gesture-speech alignment losses.

Gesture-Speech Alignment Loss

We propose a novel alignment loss based on Cross-Modal Temporal Distance (CMTD) to enforce temporal alignment between gesture apex points and speech prominences, as illustrated in Figure 2. Gesture apexes are identified as the midpoints of high-magnitude motion peaks, while speech

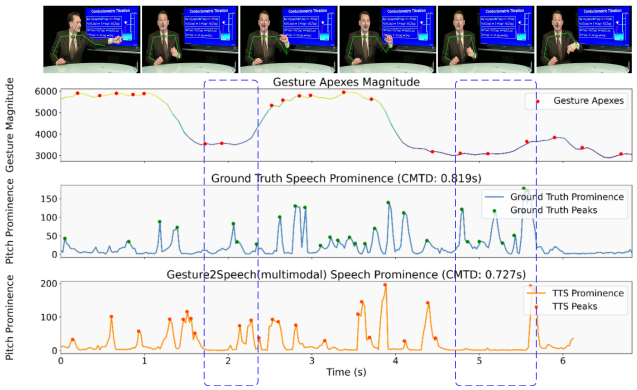

Figure 2: Comparison of gesture apexes and speech prominence peaks between ground truth and TTS-generated speech. The Cross-Modal Temporal Distance (CMTD) quantifies alignment between gestures and prosodic peaks. Ground truth CMTD: 0.819 seconds, indicating looser temporal alignment, while TTS CMTD: 0.727 seconds suggests improved synchronization with gestures.

end timings are derived from the predicted token sequence produced by the decoder.

Let t_{pred} denote predicted speech durations (in seconds) inferred from the stop token positions, and t_{gesture} denote gesture apex times extracted from motion magnitudes. The alignment loss is defined as the mean absolute error:

$$\mathcal{L}_{\text{AL}} = \frac{1}{B}\sum_{i=1}^{B}\left| t_{\text{pred}}^{(i)} - t_{\text{gesture}}^{(i)} \right|, \tag{8}$$

where B is the batch size.

Our final loss function combines standard text cross-entropy loss, mel distortion loss, duration loss $\mathcal{L}_{\text{dur}}$ and alignment loss $\mathcal{L}_{\text{AL}}$:

$$\mathcal{L} = \mathcal{L}_{\text{text}} + \mathcal{L}_{\text{mel}} + \lambda_{\text{dur}}\mathcal{L}_{\text{dur}} + \lambda_{\text{AL}}\mathcal{L}_{\text{AL}} \tag{9}$$

This encourages natural speech generation while preserving alignment between gesture intent and prosodic realization.

Experimental Setup

We conduct all experiments using a subset of the PATS dataset (Ahuja et al. 2020a,b; Ginosar et al. 2019)[2], which provides transcribed poses with aligned audios and corresponding transcripts. Our experiments focus on five speakers, namely, Alamaram, Angelica, Kubinec, Oliver, and Seth. We restrict clip durations to 4–15 seconds to ensure meaningful gesture extractions and resample the video to 25 fps. Audio is downsampled from 44.1 kHz to 22.05 kHz for efficient processing. The dataset contains 17,747 samples, totaling approx. 34.1 hours. We adopt a 90:10 train-test split for all model variants.

Baselines

As baselines, we adopt two state-of-the-art multilingual and zero-shot expressive TTS models XTTS-V2 (Casanova et al.

[1]https://github.com/jik876/hifi-gan (MIT License)

[2]https://github.com/chahuja/pats (CC BY-NC 2.0 License)

2024)[3] and GPT-SoVITS (RVC-Boss 2024)[4] neither of which incorporates explicit gesture-speech alignment. Both models provide strong prosody modeling and high-fidelity speech synthesis, making them effective starting points for multimodal extensions. We first experimented with GPT-SoVITS by injecting pose-derived gesture embeddings into the GPT module alongside the text representation. However, this led to hallucinations in the generated speech and failed to capture gesture-speech intent accurately. Subsequently, we integrated gesture information into the XTTS-V2 pipeline by extracting visual-semantic features from hand gestures and upper-body motion. These features were fused with text and audio representations via a cross-attention mechanism within the LLM-based decoder, allowing the model to attend the relevant motion cues while generating speech along with gesture speech alignment loss. To further enhance multimodal fusion, we incorporate sparse Mixture-of-Experts and hierarchical MoE modules[5] at critical fusion points. These modules dynamically route modality-specific information to specialized expert networks, improving both expressiveness and generalization. This progression from unimodal baselines to a hierarchically fused multimodal architecture forms the backbone of our Gesture2Speech architecture.

Model Configurations

Our proposed Gesture2Speech system builds upon the XTTS-V2 architecture, incorporating a multimodal framework enriched by multiple sparse Mixture-of-Experts modules to enable adaptability to gesture-aware speech synthesis. The core autoregressive speech generation is handled by a transformer-based LLM configured with 30 layers, each having a hidden size of 1024 and 16 attention heads. We integrate three distinct MoE modules: a Multimodal MoE operating on the fused gesture-text-audio embeddings, a Speech MoE focusing on spectrogram features, and a Video MoE tailored for visual-semantic gesture features. Each MoE is composed of either 8 or 16 experts, where each expert is a four-layer feed-forward network with Leaky ReLU activation (Xu et al. 2015). The choice of expert count is informed by prior work such as Switch Transformer (Fedus, Zoph, and Shazeer 2022) and V-MoE (Riquelme et al. 2021), which demonstrate that this range strikes a good tradeoff between routing stability and computational overhead. Expert routing is performed using top-2 routing with randomized fallback and adaptive capacity constraints to ensure balanced utilization during training and inference. To further enhance multimodal representation, we employ Hierarchical Mixture-of-Experts (H-MoE) modules. All the H-MoEs are initialized with an expert configuration of `num_experts=(4, 4)`, enabling efficient handling of modality-specific complexities.

Furthermore, the system leverages a HiFi-GAN vocoder configured to accept input at 22.05 kHz and produce output at 24 kHz, with conditioning vectors applied at each upsampling layer to maintain temporal and acoustic fidelity. All models are trained from scratch using an NVIDIA A100 80GB GPU for 100 epochs with a batch size of 48. We use the Adam optimizer with a learning rate of 5e-6. During inference, a probability of dropping condition is 0.1 and temperature of 0.7 are applied to to control randomness in outputs.

Results and Discussion

We consider five models in our evaluation, as shown in Table 1: (1) Gesture2Speech: XTTS-V2, (2) Gesture2Speech: GPT-SoVITS, (3) Gesture2Speech: Unimodal MoE, (4) Gesture2Speech: Hierarchical MoE, and (5) the proposed Gesture2Speech: Multimodal MoE.

Evaluation Metrics

To assess gesture-speech coordination, we employ two custom-designed metrics tailored to capture the quality of cross-modal alignment: Gesture Offset and Gesture-Audio Mutual Information.

Gesture Offset measures the average temporal misalignment between peaks in gesture motion and corresponding peaks in speech pitch prominence. Gesture peaks are identified by detecting apex points in the norm of gesture vectors, while speech peaks are derived from the pitch contour of the audio signal. The computed apex points from both modalities are temporally aligned, and the gesture offset is calculated as the mean absolute difference (in seconds) between these matched peaks. A lower gesture offset value reflects a tighter synchronization between gestural intent and vocal expression.

Gesture-Audio Mutual Information quantifies the statistical dependency between the temporal dynamics of gesture features and speech prosody. This metric provides a global measure of how effectively gestural input influences speech characteristics over time. Higher mutual information values indicate stronger cross-modal coupling, reflecting more expressive and gesture-aware speech synthesis. To compute this, gesture and speech peak times are discretized into uniform bins over the full audio duration, and the resulting histograms are used to estimate mutual information via non-parametric regression.

In addition to gesture-speech coordination, we evaluate the synthesized speech for intelligibility and naturalness using a suite of objective metrics. Word Error Rate (WER) and Character Error Rate (CER) are used to assess intelligibility, computed using transcriptions generated by the Whisper-base model (Radford et al. 2022). To assess prosodic similarity, we employ AutoPCP (Barrault et al. 2023)[6], which measures the prosodic similarity between the synthesized and reference speech. Hence, it serves as a direct indicator of improvement in prosody modeling, with higher scores indicating stronger prosodic similarity and, by extension, more expressive and natural-sounding TTS outputs. We also evaluate the perceptual quality of the generated speech using predicted Mean Opinion Scores (MOS). Two systems are used: UTMOS (Saeki et al. 2022)[7], and WVMOS, which is based on a fine-

[3]https://github.com/coqui-ai/TTS (MPL-2.0 License)

[4]https://github.com/RVC-Boss/GPT-SoVITS (MIT License)

[5]https://github.com/lucidrains/mixture-of-experts (MIT License)

[6]https://github.com/facebookresearch/seamless_communication (MIT License)

[7]https://github.com/sarulab-speech/UTMOS22 (MIT License)

Table 1: Objective Evaluations on PATS test set. UTMOS, WVMOS, and AutoPCP are reported with 95% confidence intervals.

Method	Gesture Offset ↓	Mutual Info ↑	WER ↓	CER ↓	UTMOS ↑	WVMOS ↑	AutoPCP ↑
Same Text							
Ground Truth	1.0198	0.0362	35.61	25.20	3.34±0.16	3.32±0.23	–
Gesture2Speech (XTTS V2)	1.0386	0.0382	20.27	14.85	3.34±0.11	3.34±0.25	3.08±0.14
Gesture2Speech (GPT-SoVITS)	1.8656	0.0070	34.04	26.00	3.17±0.67	3.51±0.66	3.14±0.48
Gesture2Speech (unimodal MoE)	0.9794	0.0404	22.42	15.20	3.44±0.11	3.45±0.23	3.12±0.10
Gesture2Speech (H-MoE)	1.2008	0.0357	**16.93**	**11.74**	3.46±0.12	3.36±0.34	3.12±0.10
Gesture2Speech (multimodal MoE)	**0.9471**	**0.0559**	17.55	12.14	**3.70±0.09**	**3.65±0.16**	**3.19±0.06**
Different Text							
Gesture2Speech (XTTS V2)	2.0554	0.0433	19.22	12.80	3.25±0.11	3.18±0.26	2.65±0.12
Gesture2Speech (GPT-SoVITS)	4.9933	0.0047	34.29	24.17	3.42±1.10	2.75±1.53	2.33±0.70
Gesture2Speech (unimodal MoE)	2.5915	0.0411	19.89	12.69	3.40±0.10	**3.39±0.22**	2.65±0.08
Gesture2Speech (H-MoE)	3.2073	0.0265	20.56	13.53	3.55±0.12	3.32±0.26	2.61±0.09
Gesture2Speech (multimodal MoE)	**1.9434**	**0.0475**	**18.97**	**12.15**	**3.54±0.10**	**3.39±0.25**	**2.69±0.10**

Table 2: Subjective Evaluation on Speech Quality and Prosodic Similarity of Gesture2Speech Variants along with a margin of error corresponding to the 95% confidence interval.

	Gesture2Speech				
Metric	**XTTS v2**	**GPT-SoVITS**	**Unimodal MoE**	**H-MoE**	**Multimodal MoE**
Speech Quality ↑	75.79 ± 2.39	70.78 ± 2.87	72.22 ± 2.62	73.55 ± 2.63	$\mathbf{81.48 \pm 2.25}$
Prosodic Similarity ↑	72.78 ± 2.44	67.89 ± 3.20	71.59 ± 2.45	71.54 ± 2.91	$\mathbf{79.35 \pm 2.52}$

tuned Wave2Vec2.0 model (Baevski et al. 2020) (Andreev et al. 2023)[8]. These metrics are computed on the same text and different text scenarios. In the same text scenarios, the input text used for audio synthesis matches the text spoken in the reference video. On the other hand, in the different text scenarios, the synthesized audio is generated from text that differs from the content of the reference video.

Objective Evaluations

Table 1 presents the results of our objective evaluations. The proposed Gesture2Speech: Multimodal MoE model consistently outperforms all baselines across both alignment and perceptual metrics, under both same-text and different-text evaluation settings. To further enhance the style transfer network, we experimented with a hierarchical MoE (H-MoE) a hierarchical routing mechanism with top-k=2 for expert selection. Compared to H-MoE, the Multimodal MoE shows a gesture offset improvement of 39.3% and a gesture-audio mutual information gain of 79.9% under the different text scenarios, although in the same text scenarios, H-MoE shows 0.62% improvement in WER and 0.40% improvement in CER. We also report a margin of error corresponding to the 95% confidence intervals for UTMOS, WVMOS, and AutoPCP scores to assess the statistical reliability of our evaluation. These results demonstrate that the proposed multimodal MoE architecture provides consistent improvements in both alignment and speech quality metrics across evaluation conditions.

Subjective Evaluations

To assess the perceptual quality and prosodic naturalness of the generated speech, we conducted a subjective evaluation study involving 30 participants all with no known hearing impairments, aged between 25 and 37 years. Participants were instructed to rate each audio sample on a scale from 0 to 100, where higher scores reflect better quality and naturalness. Each subject evaluated a randomized set of 720 samples, for all five Gesture2Speech model variants: XTTS-V2, GPT-SoVITS, Unimodal MoE, Hierarchical MoE, and the proposed Multimodal MoE. The evaluation focused on two key metrics: overall speech quality and prosodic similarity. The scores were aggregated for all subjects and and we report the Mean Opinion Scores (MOS) along with 95% confidence intervals in Table 2. Compared to the XTTS v2 baseline, the proposed Multimodal MoE achieved an improvement of approximately 7.5% in speech quality and 9.1% in prosodic similarity. While the H-MoE model also showed improvements over the GPT-SoVITS and Unimodal MoE baselines, its scores remained approximately 10.8% lower in speech quality and 10.9% lower in prosodic similarity compared to the proposed Multimodal MoE. These results confirm that the integration of multimodal information via Mixture of Experts enhances both the perceived quality and expressiveness of the generated speech.

Ablation Experiments

We performed unimodal experiments by adding modality specific MoE's in architecture, first we experimented by including one Speech-unimodal MoE taking audio features (no other MoE), similarly we did for Video-unimodal MoE. The evaluation results presented in Table 3 demonstrate the performance of various Gesture2Speech models under same and different text conditions. The multimodal MoE model consistently outperforms other models in terms of prosody and naturalness as reflected in the AutoPCP, UTMOS and WVMOS scores in both conditions. As compared to the speech-only multimodal MoE reflects an approximate 9% improvement,

[8]https://github.com/AndreevP/wvmos

Table 3: Ablation Evaluations using different MoE configurations. UTMOS, WVMOS, and AutoPCP are reported with a margin of error corresponding to the 95% confidence intervals.

Method	Gesture Offset ↓	Mutual Info ↑	WER ↓	CER ↓	UTMOS ↑	WVMOS ↑	AutoPCP ↑
Same Text							
Gesture2Speech (Speech-Unimodal MoE)	0.9663	0.0424	20.78	15.64	3.40±0.11	3.35±0.27	3.16±0.12
Gesture2Speech (Video-Unimodal MoE)	1.0324	0.0191	31.43	25.01	3.41±0.11	3.48±0.26	3.12±0.06
Gesture2Speech (Multimodal MoE)	**0.9471**	**0.0559**	**17.55**	**12.14**	**3.70±0.09**	**3.65±0.16**	**3.19±0.06**
Different Text							
Gesture2Speech (Speech-Unimodal MoE)	2.2088	0.0340	26.87	15.93	3.47±0.11	3.27±0.24	**2.71**±0.10
Gesture2Speech (Video-Unimodal MoE)	2.1835	**0.0479**	27.42	14.74	3.52±0.10	3.36±0.24	2.67±0.07
Gesture2Speech (Multimodal MoE)	**1.9434**	0.0475	**18.97**	**12.15**	**3.54±0.10**	**3.39±0.25**	2.69±0.10

Table 4: Ablation with respect to Fusion Strategies.

Method	Gesture Offset ↓	Mutual Info ↑	UTMOS ↑	WVMOS ↑
Cross-Attention	0.8410	0.0223	3.36±0.25	3.42±0.33
Concatenation	1.0295	0.0134	3.04±0.36	3.32±0.46
MoE Fusion	**0.7576**	0.0606	**3.64±0.22**	**3.67±0.30**

similarly, the WVMOS score shows about a 9% gain in perceptual quality. The AutoPCP metric, representing a relative increase of around 2%–6% over the unimodal variants. Under the different text condition, the multimodal MoE still maintains strong performance, maintaining competitive mutual information, only video-unimodal MoE showed improved 0.84% higher mutual information score.

Table 4 presents the quantitative fusion strageties evaluations. In style transfer network, we compared multimodal Mixture of Experts, cross-attention and concatenation fusion strategies. The proposed MoE Fusion strategy achieves substantial improvements over baseline methods. Compared to Cross-Attention, it reduces gesture offset by 9.9% and increases mutual information by 171.7%. In terms of perceptual metrics, MoE Fusion improves UTMOS by 8.3% and WVMOS by 7.3%. Relative to the Concatenation baseline, it reduces gesture offset by 26.4%, , and increases UTMOS and WVMOS by 19.7% and 10.5%, respectively.

t-SNE Analysis of Expert Specialization

To better understand the behavior of the individual MoE modules, we visualize their output embeddings using t-SNE, as shown in Figure 3. The key objective is to assess how well the different expert pathways specialize across modalities and how effectively the system integrates them. The Multimodal MoE, which processes the global style tokens from the perceiver module, shows a clear separation in the t-SNE space, indicating a strong expert specialization. This suggests that the learned representation captures distinct prosodic and stylistic features across different inputs. For the Speech MoE and Video MoE, we observe partial segregation of clusters. While not as clearly separated as the Multimodal MoE, these modules exhibit an emergent structure, indicating that the experts are beginning to specialize with some overlap. This is expected given that these components are processing modality specific features, such as spectrogram embeddings and motion embeddings that may share some temporal correlations. This supports the idea that combining complementary modalities in a controlled MoE framework leads to richer and more informative latent space representations. These findings align with the qualitative performance of the system, where gesture-conditioned speech outputs exhibit better alignment and prosodic richness.

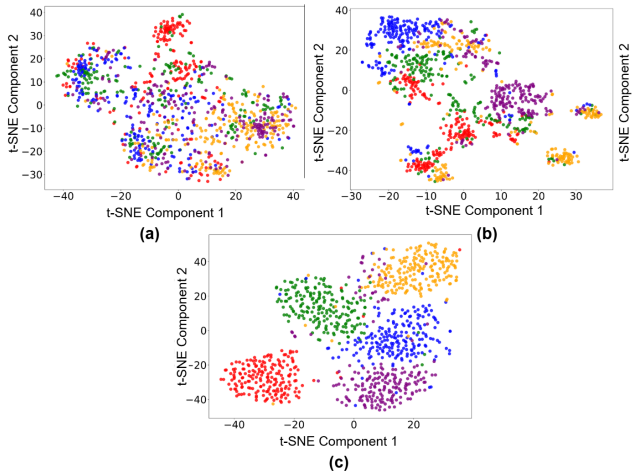

Figure 3: Speaker-wise t-SNE plots analysis of proposed style transfer network. The t-SNE plot of (a) Speech MoE embeddings, (b) video MoE embeddings and (c) Multimodal MoE embeddings.

Conclusion

In this work, we introduced Gesture2Speech, a gesture-conditioned text-to-speech (TTS) system that synthesizes expressive speech by integrating multimodal cues, such as, text, audio, and video-based hand gesture features through a cross-attention mechanism. Our framework employs modality-specific Mixture-of-Experts (MoE) modules for adaptive fusion and incorporates a gesture-speech alignment loss to achieve fine-grained temporal synchrony between gestures and prosodic contours. Experiments on the PATS dataset demonstrate consistent improvements in prosody, alignment, and naturalness across objective and subjective evaluations. This study underscores how bodily cues, particularly hand gestures, can enhance prosodic expressivity and emotional grounding in neural speech synthesis. Future work will extend this framework to full-body motion cues and explore lightweight routing strategies for expert selection and more nuanced gesture-speech synchronization in real-world scenarios.

References

Ahuja, C.; Lee, D. W.; Ishii, R.; and Morency, L.-P. 2020a. No Gestures Left Behind: Learning Relationships between Spoken Language and Freeform Gestures. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, 1884–1895.

Ahuja, C.; Lee, D. W.; Nakano, Y. I.; and Morency, L.-P. 2020b. Style Transfer for Co-speech Gesture Animation: A Multi-speaker Conditional-Mixture Approach. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII*, 248–265. Berlin, Heidelberg: Springer-Verlag. ISBN 978-3-030-58522-8.

Alexanderson, S.; et al. 2020. Generating coherent spontaneous speech and gesture from text. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*. New York, NY, USA: Association for Computing Machinery.

Andreev, P.; Alanov, A.; Ivanov, O.; and Vetrov, D. 2023. HIFI++: A Unified Framework for Bandwidth Extension and Speech Enhancement. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Baevski, A.; Zhou, Y.; Mohamed, A.; and Auli, M. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems (Neurips)*, volume 33, 12449–12460. Curran Associates, Inc.

Barrault, L.; et al. 2023. Seamless: Multilingual Expressive and Streaming Speech Translation. arXiv:2312.05187.

Brannon, W.; Virkar, Y.; and Thompson, B. 2023. Dubbing in practice: A large scale study of human localization with insights for automatic dubbing. *Transactions of the Association for Computational Linguistics*, 11: 419–435.

Cao, Z.; Hidalgo, G.; Simon, T.; Wei, S.-E.; and Sheikh, Y. 2021. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(1): 172–186.

Casanova, E.; Davis, K.; Gölge, E.; Göknar, G.; Gulea, I.; Hart, L.; Aljafari, A.; Meyer, J.; Morais, R.; Olayemi, S.; et al. 2024. Xtts: a massively multilingual zero-shot text-to-speech model. In *INTERSPEECH*. Kos Island, Greece.

Chu, Y.; Shim, Y.; and Park, U. 2024. Facial Expression-Enhanced TTS: Combining Face Representation and Emotion Intensity for Adaptive Speech. *arXiv preprint arXiv:2409.16203*.

Fedus, W.; Zoph, B.; and Shazeer, N. 2022. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. arXiv:2101.03961.

Feyereisen, P.; and De Lannoy, J.-D. 1991. *Gestures and speech: Psychological investigations*. Cambridge University Press.

Ginosar, S.; Bar, A.; Kohavi, G.; Chan, C.; Owens, A.; and Malik, J. 2019. Learning individual styles of conversational gesture. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3497–3506.

Han, W.; Kang, M.; Kim, C.; and Yang, E. 2025. StableTTS: Stable Speaker-Adaptive Text-to-Speech Synthesis via Prosody Prompting. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

He, X.; Huang, Q.; Zhang, Z.; Lin, Z.; WU, Z.; Yang, S.; Li, M.; Chen, Z.; Xu, S.; and Wu, X. 2024. Co-Speech Gesture Video Generation via Motion-Decoupled Diffusion Model. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2263–2273.

Hu, C.; Tian, Q.; Li, T.; Yuping, W.; Wang, Y.; and Zhao, H. 2021. Neural dubber: Dubbing for videos according to scripts. *Advances in Neural Information Processing Systems (Neurips))*, 34: 16582–16595.

Jacobs, R. A.; Jordan, M. I.; Nowlan, S. J.; and Hinton, G. E. 1991. Adaptive Mixtures of Local Experts. *Neural Computation*, 3(1): 79–87.

Jawaid, A.; Chandra, S. S.; Lu, J.; and SISMAN, B. 2024. Style Mixture of Experts for Expressive Text-To-Speech Synthesis. In *Audio Imagination: NeurIPS 2024 Workshop AI-Driven Speech, Music, and Sound Generation*.

Jiménez-Bravo, M.; and Marrero-Aguiar, V. 2024. Multimodal prosody: gestures and speech in the perception of prominence in Spanish. *Frontiers in Communication*, Volume 9 - 2024.

Kong, J.; Kim, J.; and Bae, J. 2020. HiFi-GAN: generative adversarial networks for efficient and high fidelity speech synthesis. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (Neurips))*, NIPS '20. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781713829546.

Li, J.; Zhong, J.; and Wang, N. 2023. A multimodal human-robot sign language interaction framework applied in social robots. *Frontiers in Neuroscience*, Volume 17 - 2023.

Lu, J.; Sisman, B.; Liu, R.; Zhang, M.; and Li, H. 2022. Visualtts: Tts with accurate lip-speech synchronization for automatic voice over. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8032–8036. IEEE.

Madhiarasan, M.; and Roy, P. P. 2022. A Comprehensive Review of Sign Language Recognition: Different Types, Modalities, and Datasets. arXiv:2204.03328.

Mehta, S.; et al. 2023. Diff-TTSG: Denoising probabilistic integrated speech and gesture synthesis. *arXiv preprint arXiv:2306.09417*.

Mehta, S.; et al. 2024. Unified Speech and Gesture Synthesis Using Flow Matching. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Nyatsanga, S.; Kucherenko, T.; Ahuja, C.; Henter, G. E.; and Neff, M. 2023. A comprehensive review of data-driven co-speech gesture generation. In *Computer Graphics Forum*, 569–596.

Papastratis, I.; Chatzikonstantinou, C.; Konstantinidis, D.; Dimitropoulos, K.; and Daras, P. 2021. Artificial Intelligence Technologies for Sign Language. *Sensors (Basel)*, 21(17): 5843.

Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2022. Robust Speech Recognition via Large-Scale Weak Supervision. arXiv:2212.04356.

Ren, Y.; Ruan, Y.; Tan, X.; Qin, T.; Zhao, S.; Zhao, Z.; and Liu, T.-Y. 2019. FastSpeech: Fast, Robust and Controllable Text to Speech. arXiv:1905.09263.

Riquelme, C.; Puigcerver, J.; Mustafa, B.; Neumann, M.; Jenatton, R.; Pinto, A. S.; Keysers, D.; and Houlsby, N. 2021. Scaling Vision with Sparse Mixture of Experts. arXiv:2106.05974.

RVC-Boss. 2024. GPT-SoVITS. https://github.com/RVC-Boss/GPT-SoVITS. Accessed: 2025-05-12.

Saeki, T.; Xin, D.; Nakata, W.; Koriyama, T.; Takamichi, S.; and Saruwatari, H. 2022. UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022. arXiv:2204.02152.

Sahipjohn, N.; Gudmalwar, A.; Shah, N.; Wasnik, P.; and Shah, R. R. 2024. DubWise: Video-Guided Speech Duration Control in Multimodal LLM-based Text-to-Speech for Dubbing. In *INTERSPEECH*. Kos Island, Greece.

Shen, J.; Pang, R.; Weiss, R. J.; Schuster, M.; Jaitly, N.; Yang, Z.; Chen, Z.; Zhang, Y.; Wang, Y.; Skerry-Ryan, R.; Saurous, R. A.; Agiomyrgiannakis, Y.; and Wu, Y. 2018. Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions. arXiv:1712.05884.

Shimizu, R.; Yamamoto, R.; Kawamura, M.; Shirahata, Y.; Doi, H.; Komatsu, T.; and Tachibana, K. 2024. Prompttts++: Controlling speaker identity in prompt-based text-to-speech using natural language descriptions. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 12672–12676. IEEE.

Teh, T. H.; Hu, V.; Ram Mohan, D. S.; Hodari, Z.; Wallis, C. G. R.; Gómez Ibarrondo, T.; Torresquintero, A.; Leoni, J.; Gales, M.; and King, S. 2023. Ensemble Prosody Prediction For Expressive Speech Synthesis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5.

Wagner, P.; Malisz, Z.; and Kopp, S. 2014. Gesture and speech in interaction: An overview.

Wang, S.; et al. 2021. Integrated speech and gesture synthesis. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, 177–185.

Wang, Y.; Stanton, D.; Zhang, Y.; Skerry-Ryan, R.; Battenberg, E.; Shor, J.; Xiao, Y.; Ren, F.; Jia, Y.; and Saurous, R. A. 2018. Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis. arXiv:1803.09017.

Xu, B.; Wang, N.; Chen, T.; and Li, M. 2015. Empirical Evaluation of Rectified Activations in Convolutional Network. arXiv:1505.00853.

Yan, Y.; Tan, X.; Li, B.; Zhang, G.; Qin, T.; Zhao, S.; Shen, Y.; Zhang, W.; and Liu, T. 2021. AdaSpeech 3: Adaptive Text to Speech for Spontaneous Style. *CoRR*, abs/2107.02530.

Zhang, J.; Guo, Z.; He, M.; and Yoshie, O. 2025. FastTalker: An unified framework for generating speech and conversational gestures from text. *Neurocomputing*, 638: 130074.

Zhang, X.; Tie, Y.; and Qi, L. 2021. SlowFast Convolution LSTM Networks for Dynamic Gesture Recognition. In *Proceedings of the 2021 3rd Asia Pacific Information Technology Conference*, APIT '21, 59–63. New York, NY, USA: Association for Computing Machinery. ISBN 9781450388108.

Zhang, Y.; Frassinelli, D.; Tuomainen, J.; Skipper, J. I.; and Vigliocco, G. 2021. More than words: word predictability, prosody, gesture and mouth movements in natural language comprehension. *Proceedings of the Royal Society B: Biological Sciences*, 288(1955): 20210500. Epub 2021 Jul 21.

Reproducibility Checklist

1. General Paper Structure

1.1. Includes a conceptual outline and/or pseudocode description of AI methods introduced (yes/partial/no/NA) Yes

1.2. Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results (yes/no) Yes

1.3. Provides well-marked pedagogical references for less-familiar readers to gain background necessary to replicate the paper (yes/no) Yes

2. Theoretical Contributions

2.1. Does this paper make theoretical contributions? (yes/no) No

If yes, please address the following points:

2.2. All assumptions and restrictions are stated clearly and formally (yes/partial/no) Type your response here

2.3. All novel claims are stated formally (e.g., in theorem statements) (yes/partial/no) Type your response here

2.4. Proofs of all novel claims are included (yes/partial/no) Type your response here

2.5. Proof sketches or intuitions are given for complex and/or novel results (yes/partial/no) Type your response here

2.6. Appropriate citations to theoretical tools used are given (yes/partial/no) Type your response here

2.7. All theoretical claims are demonstrated empirically to hold (yes/partial/no/NA) Type your response here

2.8. All experimental code used to eliminate or disprove claims is included (yes/no/NA) Type your response here

3. Dataset Usage

3.1. Does this paper rely on one or more datasets? (yes/no) Yes

If yes, please address the following points:

3.2. A motivation is given for why the experiments are conducted on the selected datasets (yes/partial/no/NA) Yes

3.3. All novel datasets introduced in this paper are included in a data appendix (yes/partial/no/NA) yes

3.4. All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no/NA) NA

3.5. All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations (yes/no/NA) Yes

3.6. All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available (yes/partial/no/NA) Yes

3.7. All datasets that are not publicly available are described in detail, with explanation why publicly available alternatives are not scientifically satisficing (yes/partial/no/NA) NA

4. Computational Experiments

4.1. Does this paper include computational experiments? (yes/no) Yes

If yes, please address the following points:

4.2. This paper states the number and range of values tried per (hyper-) parameter during development of the paper, along with the criterion used for selecting the final parameter setting (yes/partial/no/NA) Yes

4.3. Any code required for pre-processing data is included in the appendix (yes/partial/no) No (Subject to internal Approval)

4.4. All source code required for conducting and analyzing the experiments is included in a code appendix (yes/partial/no) No (Subject to internal Approval)

4.5. All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no) Partial

4.6. All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from (yes/partial/no) Yes

4.7. If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results (yes/partial/no/NA) NA

4.8. This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks (yes/partial/no) Partial

4.9. This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics (yes/partial/no) Yes

4.10. This paper states the number of algorithm runs used

to compute each reported result (yes/no) Yes

4.11. Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information (yes/no) Yes

4.12. The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank) (yes/partial/no) Yes

4.13. This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments (yes/partial/no/NA) Yes