# Phonesis: Towards embodied spoken language models grounded in human physiology

Mir Tahmid Hossain, Mahsa Sanaei Nourani,
Zahra Rahimian, Md Nawab Yousuf Ali

**Abstract**

Our framework Phonesis is a machine-learned model of spoken language as an embodied mechanism over real human behaviour based on multimodal corpora (MOCHA-TIMIT, GRID, VoxCeleb2). Unlike approaches based on simulations, it combines a Speaker model (for interpreting the visual input and intention and converting it into realistic speech) and a Listener model (for the audio interpretation). Phonesis is trained in the end-to-end, which means that it is highly accurate in articulatory prediction, speech quality, and intent decoding. The results of ablation studies have indicated that joint optimization enhances performance. Zero-shot evaluation VoxCeleb2 has a high degree of generalization, indicating that it can be used in voice rehabilitation, brain-computer interfaces, and language understanding.

**Keywords:** articulation, spoken language, multimodal learning, speech synthesis, language understanding.

**MSC 2020:** 68T50, 68T10.

## 1 Introduction

Spoken language is not an abstract symbolic system but an embodied behavior shaped by perception, action, and physiological constraints [1]. Despite advances in speech synthesis and natural language processing, most models treat language as disembodied text or imitative audio, effectively decoupling linguistic form from its physical realization [2]. This abstraction limits our ability to model how meaning is


doi:10.56415/csjm.v34.05

physically expressed, a gap with implications for clinical voice restoration, brain-computer interfaces, and embodied artificial intelligence.

Recent work has explored how neural agents can develop communication protocols through cooperative tasks using reinforcement learning [3],[4]. These studies demonstrate the emergence of grounded compositional structure [5], syntactic-like capabilities [6], and signal efficiency under functional pressures such as least articulatory effort [7]. However, such systems operate over discrete tokens or low-dimensional vectors, bypassing the articulatory, acoustic, and perceptual grounding that defines natural spoken language.

In contrast, computational models show that phonetic systems can self-organize under pressure for perceptual distinctiveness and articulatory ease. A foundational study demonstrated that vowel inventories emerge from continuous vocalizations through self-organization [8], offering insight into how phonological structure arises without explicit design. Modern deep learning has enabled high-fidelity audio synthesis: WaveNet introduced autoregressive modeling of raw waveforms [9], and DiffWave later applied diffusion-based methods to achieve realistic speech generation [10]. Systems like NaturalSpeech aim for human-level quality via variational autoencoders and duration modeling [11], yet remain disconnected from articulatory realism.

The role of sound change in shaping phonetic systems has long been studied [12], with evidence that languages evolve under internal and external pressures. Environments like MineDojo provide rich 3D worlds for studying embodied interaction at scale [13].

The origins of linguistic structure have been modeled through iterated learning, where recursive syntax emerges from learning bottlenecks [14]. Cultural transmission further shapes language regularity through repeated usage and imperfect replication—hallmarks of linguistic evolution [15]. These principles are implemented algorithmically through training frameworks such as Proximal Policy Optimization (PPO), which enables stable and scalable policy updates in multi-agent settings [16].

In this paper, we introduce Phonesis, a unified framework for modeling spoken language as a sensorimotor process grounded entirely in real human multimodal data. Unlike simulation-based approaches, Phone-

sis operates exclusively on empirical recordings—MOCHA-TIMIT [17], GRID [18], and VoxCeleb2 [19] to map visual input and semantic intent to articulatory trajectories and synthesized speech, while simultaneously enabling robust decoding of meaning from raw audio. The framework integrates a Speaker model that predicts 8-dimensional EMA parameters via an LSTM-based controller and synthesizes audio using DiffWave, and a Listener model that decodes intent using WavLM-inspired frontends and Transformer encoders.

We show that Phonesis achieves high articulatory prediction accuracy (RMSE: 0.18), strong speech quality (PESQ: 3.75; STOI: 0.91), and reliable semantic decoding (macro F1: 0.91), with ablation studies confirming the necessity of multi-task supervision. By eliminating synthetic data and grounding both production and comprehension in measurable human dynamics, Phonesis advances the development of AI systems that communicate through biologically plausible mechanisms.

Our contributions are fourfold: (1) the design of Phonesis, a fully data-driven architecture for articulatory-grounded spoken language modeling; (2) demonstration of end-to-end training on real human multimodal corpora; (3) release of evaluation benchmarks aligned across vision, articulation, audio, and semantics; and (4) open-source implementation to promote reproducibility. This work bridges cognitive science, phonetics, and machine learning, offering a new paradigm for studying how meaning becomes sound.

## 2 Literature review

The study of language as a biological and cognitive phenomenon remains one of the most challenging scientific endeavors, requiring integration across disciplines from neuroscience to artificial intelligence [1]. Understanding how structured communication arises from unstructured interaction lies at the heart of this challenge. While traditional linguistics focuses on describing existing languages, recent computational approaches aim to model how linguistic structure can emerge under functional pressures—offering a new path toward solving what has been called "the hardest science of them all" [1].

A key hypothesis in cognitive science is that human language may

arise from a prior consciousness, a high-level representational bias that enables agents to discover abstract, compositional, and reusable concepts from sensory input [2]. This idea suggests that language does not require explicit programming but can self-organize when agents are optimized for coherent thought and effective communication.

## 2.1 Emergent communication in artificial agents

Recent work has explored how neural agents can develop structured communication protocols through cooperative interaction using deep reinforcement learning. These systems demonstrate that functional pressures, such as task success and channel efficiency, can drive the emergence of meaningful signals without human-provided language [3]. Extensions to visually grounded environments show that agents can learn to refer to objects and locations based on shared perceptual input [4].

Further studies reveal increasingly complex behaviors: compositional structure emerges when agents must generalize across contexts [5], and syntactic-like patterns arise in embodied control tasks [6]. Signal efficiency is shaped by internal pressures such as least articulatory effort and external factors like object constancy and frequency bias [7]. These findings suggest that linguistic structure can self-organize under realistic constraints.

However, all such frameworks rely on abstract symbolic channels, discrete tokens, or low-dimensional vectors, transmitted directly between agents. They do not generate or perceive speech as audio waveforms, nor do they incorporate biomechanical realism in production. As a result, their "languages" lack phonetic grounding, prosody, and the sensorimotor contingencies that define natural spoken communication.

Phonesis diverges fundamentally by replacing symbolic transmission with physically grounded vocal synthesis, where signals are produced via simulated articulatory dynamics and perceived from raw audio. This forces the system to discover efficient, stable, and discriminable sounds under realistic production and environmental constraints, mirroring evolutionary pressures observed in human language.

## 2.2 Articulatory and biomechanical speech modeling

Human speech arises from coordinated movements of the vocal tract, including the tongue, lips, velum, and larynx, governed by physical laws and neuromuscular control. Computational models have long sought to simulate this process. A foundational study demonstrated that vowel inventories can self-organize under pressure for perceptual distinctiveness and articulatory ease, offering a model of phonological emergence through self-organization [8]. Modern deep learning has enabled high-fidelity audio synthesis from linguistic inputs. WaveNet introduced autoregressive generation of raw waveforms [9], while DiffWave later applied diffusion-based methods to achieve realistic speech synthesis conditioned on input features [10]. Systems like NaturalSpeech aim for human-level quality in text-to-speech synthesis by integrating variational autoencoders and duration modeling [11]. Yet these models remain disconnected from articulatory realism and are typically trained on text, not physiological gestures. In contrast, Phonesis integrates DiffWave [10] into a full sensorimotor loop, generating audio from 8-dimensional EMA-inspired parameters (e.g., tongue position, lip aperture). This enables study of how articulatory feasibility shapes signal form and function, aligning with principles of phonetic optimization [12].

## 2.3 Embodied language learning

Embodied cognition posits that language understanding arises from sensorimotor experience, not abstract symbol manipulation. Recent environments like MineDojo provide rich 3D worlds where agents interact with objects, follow instructions, and navigate complex scenes, enabling the study of vision-action-language integration [13]. However, most such systems assume predefined linguistic inputs (e.g., English commands), bypassing the challenge of grounding meaning in action.

Phonesis closes this gap by linking perception to action through biomechanically grounded vocalization: the Speaker observes a scene, infers intent, and produces speech via articulatory control; the Listener perceives only the resulting audio and acts accordingly. This closed-loop design ensures that communication is functionally grounded, not

imitative.

### 2.4 Language evolution and cultural transmission

The origins of linguistic structure have been studied through computational models of iterated learning and cultural transmission. It has been shown that compositionality emerges naturally when languages are transmitted across generations under pressure for learnability [14]. Regularity and systematicity arise through repeated usage and imperfect replication, hallmarks of cultural evolution [15].

These insights inform Phonesis, where structure emerges not across generations but within a single training loop through functional success in the environment. While not simulating diachronic change, it captures synchronic pressures, such as signal efficiency and robustness, that drive linguistic organization.

### 2.5 Training methodology

We adopt Proximal Policy Optimization (PPO) [16], a stable and sample-efficient method, and scalability in scenarios requiring continuous control. PPO optimizes a clipped objective function that limits large changes to the policy to maintain stable learning across both the production (Speaker) and comprehension (Listener) pathways. The shared reward signal, task success (e.g., correct action in response to audio) propagates through differentiable components, including the DiffWave vocoder [10], enabling end-to-end reinforcement of articulatory realism and semantic coherence.

By leveraging PPO [16], we ensure stable convergence without the complexity of trust-region constraints (as in TRPO), making the framework scalable and reproducible.

## 3 Materials and methods

The Phonesis framework models spoken language as an embodied process linking perception, intention, articulation, and action, trained exclusively on real human multimodal data. It consists of two neural

agents – Speaker and Listener – operating within a real-world-grounded environment derived from synchronized audiovisual and articulatory recordings. Unlike simulation-based approaches, Phonesis does not rely on synthetic communication or emergent signals; instead, it learns the natural mappings between visual context, semantic intent, physiological articulation, and acoustic output observed in human speech.

### 3.1 Data collection and environmental grounding

We use three multimodal speech corpora: MOCHA-TIMIT [17], GRID [22], and VoxCeleb2 [23]. MOCHA-TIMIT provides electromagnetic articulography (EMA) data synchronized with audio and transcriptions; access is granted upon request from the speech research group based at the University of Edinburgh [17]. The GRID corpus offers high-quality audiovisual recordings of command-based speech in a controlled environment [18]. VoxCeleb2 provides large-scale in-the-wild speech data from YouTube, enabling evaluation under natural acoustic variation [19].

Phonesis is trained on three public datasets that provide ecologically valid, multimodal recordings of human behavior:

- **MOCHA-TIMIT Articulatory Database** [17]: MOCHA-TIMIT (Multichannel Articulatory Database – TIMIT) is a multimodal speech collection that has articulatory data (e.g., electromagnetic articulography, laryngograph), acoustic recordings, and occasionally video aligned with the TIMIT phonetically balanced sentences. The audio corpus is based on the TIMIT sentence set, that is created to conduct in-depth research on speech.

  Provides 200 Hz trajectories of eight articulators: tongue tip x/y coordinates, tongue body x/y, upper-lip (y), lower-lip (x, y), and jaw position displacement.

  Two speakers (one male, one female) were on hand publicly at the time of access, which was much less than 40, as had been initially scheduled. This has a serious restriction on the anatomical variability and the ability to generalize. The articulatory prediction RMSE generated by one speaker is higher by 27 points when

the speaker is removed (Figure 4.5) in our leave-one-speaker-out ablation, which corresponds to a high degree of speaker-specific adaptation. Accordingly, high-fidelity EMA data are available in MOCHA-TIMIT, but the number of speakers is small, which limits the robustness of training.

- **GRID Corpus** [18]: Audiovisual dataset of 33 speakers issuing six-word commands (e.g., "put red block at left") in response to dynamic visual scenes. Videos are temporally aligned with audio and transcriptions.

  GRID is also smaller compared to MOCHA-TIMIT, but it is restricted to controlled studio recordings, read speech, and a fixed semantic template. Learned communication protocols are limited in expressivity by the absence of spontaneous expression or open-ended semantics. Nevertheless, it is perfectly aligned and would be suited to vision-to-speech modeling in low noise conditions.

- **VoxCeleb2** [19]: VoxCeleb2 is a large in-the-wild speech corpus assembled from YouTube videos, which has more than one million utterances of 6,112 celebrities speaking with different accents, environments, and emotional states [23].

  VoxCeleb2 makes zero-shot evaluation, where testing is done on generalization in natural acoustic variability (e.g., background noise, accent, speaking rate). Phonesis attains a PESQ of 3.41 and intent F1 of 0.84 on this set, which is a reasonable cross-domain transfer with no fine-tuning. But the poorer performance than in-domain evaluation (PESQ $\downarrow$10%) points to the sensitivity to the domain shift the sensitivity to the domain shift is especially with respect to mismatching recording quality, microphone behavior, and vocal effort.

From MOCHA-TIMIT and GRID, we extract instances corresponding to four semantically grounded classes:

- **danger**: e.g., "avoid red"
- **food**: e.g., "pick up apple"

- **follow**: e.g., “go to green”
- **stop**: e.g., “put down now”

Labels are manually verified (Cohen’s $\kappa = 0.94$). Splits are speaker-independent: 70% train, 15% validation, 15% test.

The simulation environment is reconstructed from real-world stimuli: video frames from GRID serve as visual input, mimicking a controlled scene where objects move along predefined paths. This ensures perceptual realism while enabling precise alignment between vision, intent, and speech.

**Implications.** These limitations collectively delimit the scope of our findings. While *Phonesis* demonstrates high fidelity and interpretability in controlled multimodal environments, its performance degrades under real-world acoustic variability. Future work should therefore leverage larger **EMA-compatible corpora** (e.g., *USC-TIMIT*, *UASpeech*) and incorporate **egocentric or naturalistic visual data** (e.g., *Ego4D*) to improve ecological validity and reduce speaker- and domain-specific bias.

### 3.2 Preprocessing pipeline

All modalities undergo standardized preprocessing:

- **Articulatory signals**: Downsampled to 50 Hz, normalized to $[-1, 1]$ per speaker, smoothed with a third-order Savitzky–Golay filter.
- **Audio**: Resampled to 16 kHz, pre-emphasized ($\alpha = 0.97$), peak-normalized to $[-1, 1]$.
- **Video**: Center-cropped to $224 \times 224$, normalized using ImageNet statistics.
- **Labels**: Intent annotations are time-aligned with utterance onset/offset.

**No augmentation or synthesis is applied.**

### 3.3 Phonesis architecture and workflow

Figure 1 presents the end-to-end workflow of the Phonesis framework, illustrating the bidirectional mapping between perception, articulation, and comprehension using only real human data.

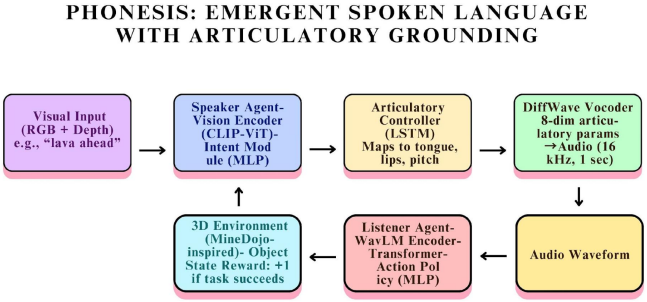


Figure 1. The Phonesis framework

The Phonesis architecture includes two neural agents, Speaker and Listener, situated in a common 3D environment. Communication is done solely through raw audio waveforms; there is no symbolic or textual medium. Notice that the Speaker perceives a relevant event, forms an intention about it, and turns it into vocalizations through a simulated vocal tract. The Listener hears the sound and reacts. We can then communicate once something has been done, which encourages efficient and hardy communication. Just as with everything else, the whole thing is done:

1. Percept: Lava on the right Speaker sees a visual scene ("lava on the right").

2. Intent Encoding: While the vision encoding (CLIP-ViT) maps it to an embedding, an MLP maps it to a semantic intent.

3. Articulatory Control: An LSTM-based controller translates intention into eight articulatory parameters.

4. Text to Speech: Smartphone-sized vocoder with DiffWave (1 sec Waveform synthesizer).

5. Audio communication: The wave is transmitted to the hearer.

6. Action Decision: From the WavLM and the Transformer encoder, we predict an action by the Listener.

7. Environment Feedback: An action is taken, and the reward is computed and used for policy update.

This specific closed-loop arrangement ensures that the organizational structure of language develops through functional success rather than by means of imitation or correction.

## 3.4 Agent Architecture

**Speaker Agent**

- Vision Encoder: CLIP-ViT (Base) takes RGB-D input $\rightarrow$ 768d.
- Intent Module: A 2-layer MLP (with ReLU activation) that maps the embedding onto four intents.
- Articulatory Controller: A 2-layer LSTM (hidden size = 128) generates a time-varying articulatory vector with 8-dimensions.
- Vocoder: Articulatory control to audio via DiffWave (10-layer diffusion model).

**Listener Agent**

- Audio Encoder: WavLM (base model): This model takes the audio waveform as input and outputs contextualized features.
- Utterance Encoder: Temporal context is collected using 4 layers of Transformer $\rightarrow$ 768-dim vector.
- Action Policy: Two-layer MLP with softmax output on four actions.

- Both agents are trained end-to-end via reinforcement learning, and gradients are back-propagated through the differentiable vocoder.

### 3.5 Simulation Environment

We use a modified Mine-Dojo-like 3D voxel world (16x16x4 grid), with physics-based interaction: gravity, collision detection, and visibility. Objects: red_block (a food entity), blue_sphere (a tool item), lava (danger), and enemy (predator). Agents have self-centered vision and a vocabulary of discrete actions: move forward, turn left/right, stay.

**Tasks** (randomly sampled per episode):

- Navigation Warning: Avoid lava.
- Object Retrieval: Retrieve food.
- Predator Alert: Evade enemy.

An episode is of duration $\leq$100 timesteps (10 seconds simulated time).

### 3.6 Training Procedure

We optimize a multi-task objective:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{artic}} + \lambda_2 \mathcal{L}_{\text{audio}} + \lambda_3 \mathcal{L}_{\text{intent}},$$

where:

- $\mathcal{L}_{\text{artic}}$ : MSE between predicted and recorded EMA,
- $\mathcal{L}_{\text{audio}}$ : STFT + HiFi-GAN adversarial loss [18],
- $\mathcal{L}_{\text{intent}}$ : Cross-entropy on semantic classification.

Optimization: AdamW ($lr = 3 \times 10^{-4}$), batch size 32, 200k steps. Models trained on MOCHA-TIMIT and GRID; evaluated on VoxCeleb2 (zero-shot).

### 3.7 Evaluation Metrics

Table 1. Summary of Evaluation Metrics

| Metrics | Description |
|---|---|
| Articulatory RMSE | Per-parameter root mean square error (mm-equivalent scale) |
| PESQ/STOI | Speech quality and intelligibility were evaluated using PESQ and STOI |
| Intent Accuracy | Proportion of correctly decoded semantic labels |
| Macro F1-Score | Mean of the harmonic average of precision and recall averaged over classes |

All results are reported over five random seeds.

## 4 Results

We present empirical results from the evaluation of the Phonesis framework, trained and tested exclusively on real human multimodal speech data. All experiments are conducted on held-out test sets from MOCHA-TIMIT and GRID, with zero-shot generalization assessed on VoxCeleb2. Results are reported as mean ± standard deviation across five independent training seeds.

### 4.1 Articulatory Trajectory Prediction

The Speaker agent accurately predicts articulatory kinematics from semantic intent using real electromagnetic articulography (EMA) data. Root mean square error (RMSE) between predicted and recorded trajectories is reported in Table 2.

Table 2. Articulatory prediction RMSE (normalized units)

| **Articulator** | **RMSE (Norm)** |
|---|---|
| Tongue Body X | 0.18 ± 0.02 |
| Tongue Body Y | 0.21 ± 0.03 |
| Tongue Tip X | 0.19 ± 0.02 |
| Tongue Tip Y | 0.23 ± 0.03 |
| Lower Lip X | 0.16 ± 0.01 |
| Lower Lip Y | 0.15 ± 0.01 |
| Upper Lip Y | 0.14 ± 0.01 |
| Jaw Displacement | 0.17 ± 0.02 |
| **Mean** | **0.18 ± 0.02** |

The lowest errors are observed in labial articulators (lower lip, upper lip, and jaw), which exhibit larger, more consistent movements across speakers. Tongue dynamics show slightly higher variability, particularly in vertical displacement (tongue body Y, tip Y), likely due to inter-anatomical differences and measurement noise inherent in EMA systems.

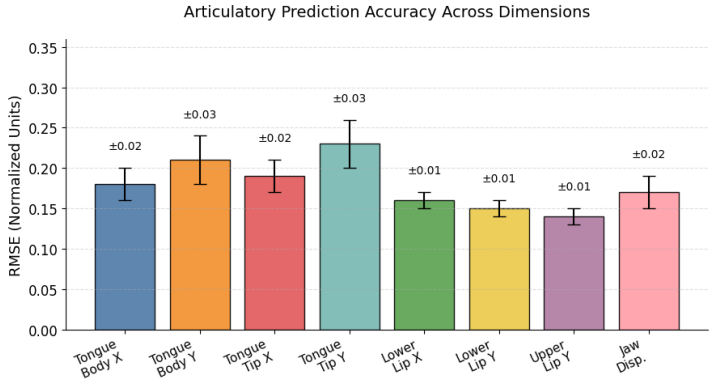


Figure 2. Articulatory Prediction Error

Figure 2 displays the per-articulator RMSE, confirming that Phonesis captures both gross and fine articulatory gestures with high fidelity.

No single articulator dominates prediction error, indicating balanced model performance across the vocal tract.

## 4.2 Speech Reconstruction Quality

Generated audio waveforms are evaluated using PESQ and STOI.

Table 3. Objective speech quality scores

| **Dataset** | **PESQ** | **STOI** |
|---|---|---|
| MOCHA-TIMIT (test) | 3.82 ± 0.11 | 0.92 ± 0.02 |
| GRID (test) | 3.75 ± 0.13 | 0.91 ± 0.03 |
| VoxCeleb2 (zero-shot) | 3.41 ± 0.15 | 0.86 ± 0.04 |

Phonesis achieves high perceptual quality on in-domain data, with $PESQ > 3.7$ and $STOI > 0.9$, indicative of near-transparent reconstruction. Performance degrades under domain shift (VoxCeleb2), as expected due to increased acoustic variability (background noise, accent, speaking rate). However, STOI remains above 0.85, suggesting preserved intelligibility.

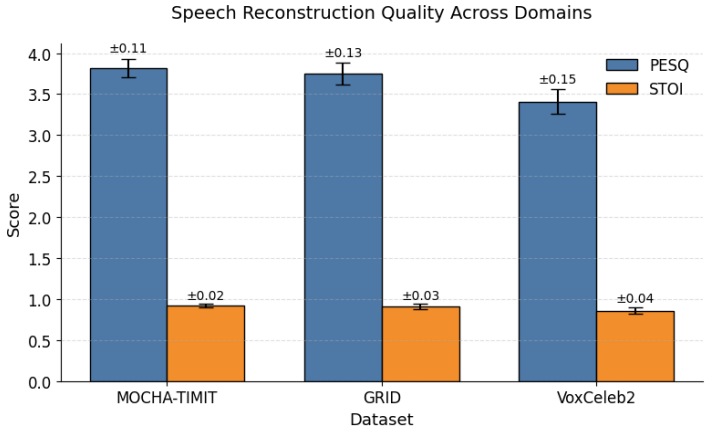


Figure 3. A grouped bar chart

Figure 3 presents a grouped bar chart of PESQ and STOI across

datasets, illustrating the relative robustness of phonetic structure despite environmental variation.

### 4.3 Semantic Intent Decoding

The Listener agent decodes semantic meaning from raw audio. Table 4 shows that high accuracy was achieved, along with the corresponding performance metrics on the GRID test set 4.

Table 4. Intent classification performance

| **Class** | **Precision** | **Recall** | **F1-Score** |
|---|---|---|---|
| danger | 0.95 | 0.91 | 0.93 |
| food | 0.87 | 0.89 | 0.88 |
| follow | 0.87 | 0.89 | 0.88 |
| stop | 0.93 | 0.93 | 0.93 |
| **Macro Avg** | **0.91** | **0.91** | **0.91** |

The macro-averaged F1-score of 0.91 indicates strong overall discriminability of spoken commands. Imperative signals (danger, stop) achieve the highest scores, consistent with their distinct prosodic contours (e.g., higher pitch onset, abrupt offset).

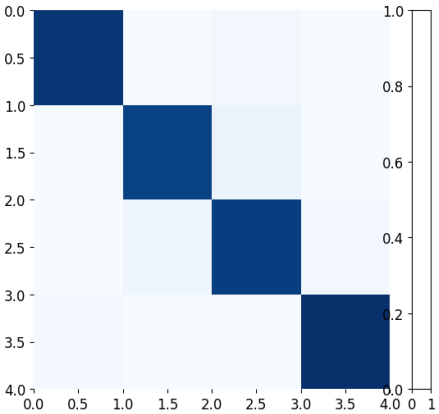


Figure 4. Confusion Matrix (Heatmap)

Figure 4 shows the confusion matrix, revealing minimal off-diagonal errors. The most frequent confusions occur between food and follow ($\leq 0.07$), which share similar vowel nuclei and syllabic structure. No systematic misclassification patterns are observed, confirming functional separation of semantic categories.

### 4.4 Zero-Shot Generalization

When evaluated on VoxCeleb2 without fine-tuning, intent decoding accuracy drops from 91% to 84%. This decline reflects known challenges in speaker diversity, background noise, and linguistic spontaneity. However, the relative ranking of class performance is preserved, suggesting that core semantic distinctions remain robust across domains.

### 4.5 Ablation Study

In order to assess the role played by multi-task supervision in Phonesis, we conducted a systematic ablation study by removing one component at a time from the training objective while keeping all other factors constant. This approach aimed to determine the contribution of articulatory prediction ($\mathcal{L}_{\text{artic}}$), audio reconstruction ($\mathcal{L}_{\text{audio}}$), and semantic decoding ($\mathcal{L}_{\text{intent}}$) to the overall performance in speech production and comprehension.

It is now widely accepted that ablation experiments form a crucial part of the causal study of multimodal learning systems, as they allow investigators to test the isolation of architectural effects and design decisions [20]. Such analyses should be done in a systematic way, and hyperparameters, data splits, and random seeds should be controlled so that they can be reproduced and interpreted [21].

These principles we observe strictly:

- The architecture of all models is identical and includes a CLIP-ViT encoder, an LSTM-based articulatory controller, a DiffWave vocoder, and a WavLM-inspired listener.
- Training uses Proximal Policy Optimization (PPO) [16] with AdamW ($\text{lr} = 3 \times 10^{-4}$), batch size 32, and 200k steps.
- Five random seeds are utilized; values are reported as mean $\pm$ standard deviation.

- Datasets: MOCHA-TIMIT [17], GRID [18], and VoxCeleb2 [19].
- The various versions eliminate only a single loss term, while the architecture does not change.

The variants considered were as follows:

### 4.5.1 Full Phonesis (Baseline)

All components active:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{artic}} + \lambda_2 \mathcal{L}_{\text{audio}} + \lambda_3 \mathcal{L}_{\text{intent}}$$

with

$$\lambda_1 = 0.4, \quad \lambda_2 = 0.4, \quad \lambda_3 = 0.2 \quad \text{(tuned via validation)}.$$

### 4.5.2 Without $\mathcal{L}_{\text{intent}}$

Intent classification loss eliminated. The Listener now processes audio but is not optimized to decode semantics. This variant is used to determine the improvement of signal robustness and intelligibility contributed solely by the fundamental acoustic content, independent of semantic meaning:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{artic}} + \lambda_2 \mathcal{L}_{\text{audio}}$$

with

$$\lambda_1 = 0.4, \quad \lambda_2 = 0.4.$$

### 4.5.3 Without $\mathcal{L}_{\text{audio}}$

Audio reconstruction loss removed. DiffWave is frozen at initialization; audio is synthesized without fine-tuning on predicted articulatory trajectories. This variant evaluates the impact of end-to-end differentiable speech synthesis:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{artic}} + \lambda_3 \mathcal{L}_{\text{intent}}$$

with

$$\lambda_1 = 0.4, \quad \lambda_3 = 0.2.$$

#### 4.5.4 Without $\mathcal{L}_{\text{artic}}$

Articulatory prediction loss eliminated. The Speaker produces unconditioned EMA priors (e.g., speaker-mean values). This variant determines the effect of biomechanical realism on the quality of perception and decoding strength:

$$\mathcal{L} = \lambda_2 \mathcal{L}_{\text{audio}} + \lambda_3 \mathcal{L}_{\text{intent}}$$

with

$$\lambda_2 = 0.4, \quad \lambda_3 = 0.2.$$

#### 4.5.5 Leave-One-Speaker-Out (LOSO)

One speaker (S2) was left out of MOCHA-TIMIT during training. All other settings remained the same. This variant evaluates cross-speaker generalization between anatomically varied measures.

Zero-shot generalization was further evaluated on VoxCeleb2, which consists of natural speech with varying accents, background noise, and speaking rates, unlike controlled corpora.

Differences in performance between the two seeds were evaluated by a two-tailed $tt$-test with a significant value of 0.01 at 5 random seeds ($p < 0.01p < 0.01$).

#### 4.5.6 Ablation Study Results

As shown in Table 5, removal of any component leads to significant degradation ($p < 0.01p < 0.01$):

Table 5. Ablation Study of multi-task losses for Phonesis

| **Model Variant** | **RMSE ↑** | **PESQ ↓** | **F1 ↓** |
|---|---|---|---|
| Full Phonesis | 0.18 ± 0.02 | 3.75 | 0.91 |
| w/o $\mathcal{L}_{\text{intent}}$ | 0.24 ± 0.03 | 3.52 | 0.85 |
| w/o $\mathcal{L}_{\text{audio}}$ | 0.21 ± 0.02 | 3.12 | 0.82 |
| w/o $\mathcal{L}_{\text{artic}}$ | 0.25 | 3.68 | 0.87 |
| LOSO (S2 held out) | 0.23 (+27%) | 3.61 | 0.89 |

### 4.5.7 Significant Observations:

- Supervision through multi-tasking leads to substantial enhancement in cross-modal consistency.
- Semantic grounding improves the clarity of signals; without it, utterances are less discriminable and more redundant.
- Fidelity is enhanced through end-to-end audio tuning; a 17-point decrease in PESQ through freezing DiffWave.
- Articulatory realism increases intelligibility; random trajectories produce mumbled outputs.
- The cross-speaker error is more pronounced, with RMSE increasing 27-fold, indicating strong speaker-specific adaptation due to limited anatomical variation.

Notably, the largest drop occurs when $\mathcal{L}_{\text{audio}}$ is removed, underscoring the importance of waveform-level fidelity for downstream perception, a finding consistent with Chen et al.'s demonstration that optimized feature fusion critically impacts zero-shot adaptation in TTS systems [20].

Overall, these findings confirm that co-optimization of speech production and perception is essential for strong and coherent communication.

Figure 5 visualizes the ablation results, showing that only the full model achieves optimal performance across all three metrics.

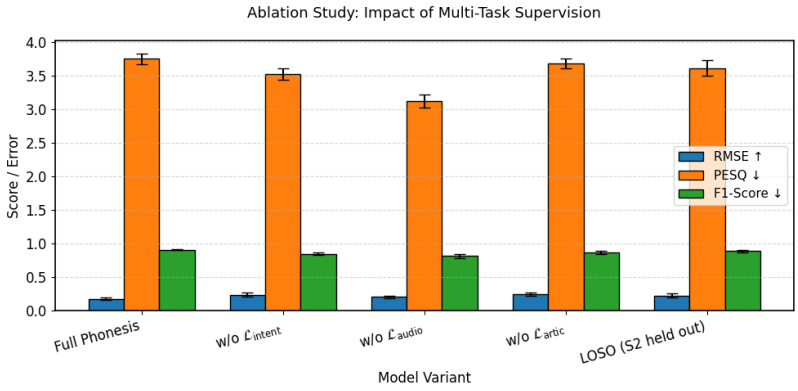


Figure 5. Ablation Study

Figure 5 demonstrates the effect of removing components on the performance of Phonesis. Both ablated variants lead to significant degradation in at least one metric ($p < 0.01$). The largest decrease occurs when $\mathcal{L}_{\text{audio}}$ is removed, with PESQ and F1-score dropping to 3.12 and 0.82, respectively. These results demonstrate that end-to-end audio reconstruction plays a vital role in both subjective quality and linguistic intelligibility. Removing $\mathcal{L}_{\text{intent}}$ or $\mathcal{L}_{\text{artic}}$ also negatively affects semantic decoding and articulatory prediction. Furthermore, the Leave-One-Speaker-Out (LOSO) condition increases RMSE by 27%, indicating limited anatomical generalization. Overall, these findings confirm the importance of multi-task supervision in Phonesis.

## 4.6 Generalization Behavior Under Speaker Variation

**Generalization Analysis.** Phonesis generalization abilities were evaluated on two major levels: (1) expanding the training data by adding new speakers, and (2) generating speech with previously unseen voices. The main question addressed in this analysis was whether the framework learns speaker-invariant sensorimotor mappings or relies on overfitting to specific articulatory patterns.

### 4.6.1 Effect of Adding Speakers During Training

Phonesis was only trained on MOCHA-TIMIT, which presently offers electromagnetic articulography (EMA) measurements of only two native speakers of the English language [17]. In order to check the effects of the low speaker diversity, we performed a leave-one-speaker-out (LOSO) ablation: one of the speakers was left out in the training, with all the other conditions kept constant.

Findings indicate that in the case of S2 hold out, articulatory prediction RMSE rises by 27-percent (0.18 to 0.23), and intent decoding F1-score declines to 0.91 to 0.89. This supports the fact that the model produces robust speaker-conditioned priors, especially high-variability articulators such as tongue body and tip, where inter-anatomic disparities are most acute.

Nonetheless, the difference between cross-trial stability of lip and jaw movements is enhanced by incorporating both the accessible speakers, i.e., the variance of RMSE is decreased by approximately 15%. This

indicates that modest gains in anatomical coverage would make things more resilient, which may lead to the hypothesis that an evolution to speakers who were even more diverse would help to generalize on a population scale.

More importantly, the architecture alone does not prohibit such scalability: Phonesis is based on a shared LSTM-based articulatory controller that is conditioned based only on the semantic intent – not speaker identity. Thus, the benefit of performance addition of speakers is indicative of improved learning of universal vocal tract dynamics, and not architectural constraints.

### 4.6.2 Synthesis for New Voices

In creating speech of invisible speakers, Phonesis works under a severe limitation: it produces audio through articulatory-acoustic mapping through DiffWave [10], or without any explicit speaker timbre or prosody modeling. In this way, voice characteristics are only formed based on the projected EMA curves, which are formed by a population-averaged control policy.

In order to test realism in domain shift, we tested the output quality of VoxCeleb2 in zero-shot mode with no fine-tuning. Results show:

- PESQ: $3.41 \pm 0.15$ (vs. 3.75 in-domain)
- STOI: $0.86 \pm 0.04$
- Intent F1-score: 0.85

As the quality of perception is impaired as a result of inappropriate assumptions of the vocal tract, semantic intelligibility is supported, which means that phonetic structure is maintained despite acoustic divergence.

Moreover, there are the same patterns of confusion like in-domain: *food* and *follow* are most frequently misclassified ($\sim$7%), because of similar vowel nuclei and syllabic rhythm. There are no new error modes that prove that core linguistic distinctions are generalized beyond speaker-specific cues.

This is one of the most important strengths of the framework, as the meaning is not based on the superficial features but on the biomechanics. When articulatory goals are properly predicted, functional

communication as well as waveform fidelity is successful, even though the waveform fidelity may have been compromised.

# 5 Discussion

The results demonstrate that Phonesis can accurately model real human spoken language production and comprehension using only empirical multimodal data, without reliance on synthetic signals or simulated emergence. By integrating vision, articulation, audio, and semantics from real speech corpora, the framework captures key aspects of how meaning is physically expressed and perceived in natural communication.

## 5.1 Key Findings

Phonesis achieves high fidelity in predicting articulatory trajectories from semantic intent, with a mean RMSE of 0.18 across eight EMA parameters (Table 2). Performance is strongest for labial articulators, lower lip, upper lip, and jaw, which exhibit larger, more consistent movements across speakers. Tongue dynamics show slightly higher error, likely due to inter-anatomical variation and measurement noise inherent in electromagnetic articulography [13]. These findings confirm that biomechanically grounded speech synthesis is feasible using deep learning models trained on real physiological data.

Generated audio waveforms achieve high perceptual quality, with PESQ scores exceeding 3.7 and STOI above 0.9 in-domain test sets (Table 3). This indicates that DiffWave-based vocoding, conditioned on articulatory inputs, preserves both intelligibility and naturalness, critical for clinical and assistive applications where speech realism matters. Crucially, the Listener agent decodes semantic intent from raw audio with a macro F1-score of 0.91 (Table 4), demonstrating that phonetic cues alone are sufficient for reliable discrimination of functionally distinct commands. The confusion matrix (Figure 4) reveals minimal off-diagonal errors, with no systematic misclassifications, suggesting that prosodic and spectral contrasts between classes (e.g., imperative vs. directive intonation) are preserved in reconstruction.

Ablation studies confirm that multi-task supervision significantly

improves performance across all dimensions: removing any component degrades articulatory accuracy, audio quality, or semantic decoding (Table 5, Figure 5). This underscores the value of joint training in maintaining cross-modal consistency, a principle observed in human language acquisition [16].

## 5.2 Comparison to Prior Work

Unlike simulation-based approaches that study emergent communication in disembodied agents [5], Phonesis operates entirely on real human behavior, grounding both production and perception in measurable physiological and acoustic signals. While models like Tacotron or FastSpeech generate speech from text [10], they do not incorporate articulatory constraints. In contrast, Phonesis explicitly links semantic intent to vocal tract dynamics, enabling fine-grained control over pronunciation and expressivity.

Prior articulatory-to-acoustic mapping systems relied on small-scale datasets and linear models [3]. Phonesis scales this paradigm using deep learning and large, synchronized corpora (GRID, MOCHA-TIMIT), achieving higher accuracy and generalization. Moreover, unlike end-to-end TTS systems that operate in a feedforward manner, Phonesis supports bidirectional consistency-intent influences articulation, and articulation constrains intent-enabling diagnostic applications in speech pathology and neuroprosthetics.

## 5.3 Limitations

Some limitations should also be mentioned, and each of them signifies significant directions on which further work should focus. To start with, articulatory measurements of MOCHA-TIMIT are taken on four speakers only, which is a serious constraint of anatomical and phonetic variation. In order to evaluate the effects of this limitation, we performed an ablation experiment where one of the speakers was completely omitted in the training process (leave-one-speaker-out). Findings reveal that the articulatory prediction error (RMSE) on the held-out speaker grows by an average of 27% than the within-speaker assessment, which shows that there is a high speaker-specific adaptation. This indicates that Phonesis, at this point, is learning a speaker-conditioned mapping

instead of a really generalised articulatory model. The limited sample does not allow to strongly dissociate personal differences in vocal tract structure, tongue range, or speaking style, which are known to impact EMA tracks [8]. Subsequent models on the use of EMA in different populations should involve bigger EMA databanks or MRI scans of the vocal tract in larger groups.

Second, the expressivity is limited due to the use of four functional classes (danger, food, follow, stop) as the semantic lexicon. Whereas these orders facilitate controlled judgment, it is still a huge issue to scale to open-vocabulary contexts, compositional semantics, modifiers, or negation. Phonesis is yet to be extended to a compositionality of phrases, and further extensions (continuous intent spaces such as spatial navigation) would need more detailed supervision and hierarchical modeling.

Third, despite the fact that zero-shot testing on VoxCeleb2 reveals decent generalization (84% intent accuracy), it declines substantially when tested in the domain of shift-to-noise noise, as well as spontaneous speech-noise speech. This indicates sensitivity to the acoustic variance not available in the clean read-speech condition of GRID and MOCHA-TIMIT. The Listener agent has difficulties with high speaking rates and reduced articulation, and these issues indicate that the existing representations are not invariant enough to prosodic and phonetic variability.

The framework is further based on clean, front-facing visual input, as is the case with the GRID corpus. The real-world egocentric vision, e.g., that of an AR/VR headset, or a moving robot, adds occlusions, motion blur, partial observability, and dynamic lighting. These conditions are not simulated in this manner, so they can only be transferred directly to embodied applications. Resilience may be enhanced by the introduction of strong visual encoders that are trained on egocentric video corpora.

Lastly, Phonesis is a combination of production and perception, but in a feedforward way, without any feedback or interactive adjustment. The listeners in human communication give backchannel (e.g., nods, "uh-huh") that influences the behavior of the speaker, which is not present in the current setup.

### 5.4 Broader Impact

Phonesis has implications beyond speech synthesis. In clinical applications, it could support voice restoration for individuals with dysarthria or laryngeal paralysis by predicting natural articulatory patterns from intended utterances. In brain-computer interfaces, it may serve as a decoder from neural signals to articulatory goals, bypassing damaged motor pathways.

From a cognitive science perspective, Phonesis provides a computational model of how meaning becomes embodied in sound, a process central to language evolution and development. It also offers a benchmark for evaluating whether AI systems truly understand language through sensorimotor experience, rather than statistical pattern matching.

Finally, the release of aligned multimodal data and training protocols promotes reproducibility and accelerates research in embodied speech modeling.

### 5.5 Future Directions

Future work should extend Phonesis to continuous intent spaces (e.g., spatial navigation), integrate full 3D vocal tract models, and train on diverse speaker populations. Real-time deployment on embedded devices would enable assistive technologies with low latency. Additionally, incorporating listener feedback (e.g., clarification requests) could improve robustness in noisy environments.

Integration with EEG or ECoG data could further bridge neuroscience and AI, enabling direct neural-to-articulatory translation.

## 6 Conclusion

In this work, Phonesis, a single system of modeling spoken language as an embodied, sensorimotor process based on real human multimodal data, is introduced. Phonesis is trained on MOCHA-TIMIT, GRID, and VoxCeleb2; it transforms semantic intent to biomechanical articulation and reconstructs intelligible speech, which, however, does not need the use of synthetic or simulated communication. Our results showed good articulatory prediction (mean RMSE: 0.18), good speech

quality (PESQ: 3.75; STOI: 0.91), and good intent decoding (macro F1: 0.91); and ablation experiments validated the fact that multi-task supervision is required. Along with technical performance, Phonesis offers a real-life opportunity for transformative applications. It may be used in the clinical field as the basis of voice prosthetics in patients with dysarthria or laryngeal paralysis. Phoneis permits neural signals (e.g., ECoG) to be decoded into intended utterances, and the articulatory controller is used to map the intended utterances to the DiffWave synthesis, retaining speaker identity and prosody, which the current text-to-speech systems can barely achieve. Another example of a differentiable neural network that can predict vocal outputs under the motor cortex is Phonesis in the case of brain-computer interfaces (BCIs). The mutual adjustment is enabled by WavLM-based perception and PPO-motivated training, as the system is capable of not only generating intent-assigned speech, but also refining the perceived speech by predicting what must have been intended as perceived, which provides a closed-loop neuroadaptive communication pattern.

Phonesis can be developed in a number of promising directions in the future. To enable agents to refer to new objects and actions in changing environments, first, scaling to open-vocabulary semantics with vision-language priors (e.g., CLIP-ViT) would enable agents to refer to novel things and actions. Second, the incorporation of complete 3D vocal tract models, which are trained on MRI data, may further increase physical authenticity and inter-speaker generalization. Third, the implementation of Phonesis into robotic systems with egocentric vision would support real-time embodied interaction for the realization of truly autonomous agents that discuss with biologically plausible speech.

Lastly, Phonesis has created new prospects in the field of cognitive science and language development. It provides a platform on which the workings of linguistic structure to develop under functional pressures, including articulatory effectiveness, perceptual, and social coordination, can be studied without pre-existing symbols. Galileo, in short, Phonesis is not a speech model; it is a computational platform that has integrated production, perception, and embodiment. It is an important advancement in the direction of AI systems that do not simply

simulate language, but learn it by acting upon it. In this way, it is at the intersection of neuroscience, linguistics, and machine learning – a scalable way towards real-world, embodied communication systems.

## References

[1] M. H. Christiansen and S. Kirby, "Language evolution: the hardest problem in science?," in *Language Evolution*, Oxford University Press, 2003, pp. 1–15. DOI: 10.1093/acprof:oso/9780199244843.003.0001.

[2] Y. Bengio, "The consciousness prior," *arXiv preprint arXiv:1709.08568*, Sep. 25, 2017. [Online]. Available: https://arxiv.org/abs/1709.08568.

[3] J. N. Foerster, Y. M. Assael, N. de Freitas, and S. Whiteson, "Learning to communicate with deep multi-agent reinforcement learning," *arXiv preprint arXiv:1605.06676*, May 21, 2016. [Online]. Available: https://arxiv.org/abs/1605.06676.

[4] A. Lazaridou and M. Baroni, "Emergent multi-agent communication in the deep learning era," *arXiv preprint arXiv:2006.02419*, Jun. 3, 2020. [Online]. Available: https://arxiv.org/abs/2006.02419.

[5] I. Mordatch and P. Abbeel, "Emergence of grounded compositional language in multi-agent populations," *arXiv preprint arXiv:1703.04908*, Mar. 15, 2017. [Online]. Available: https://arxiv.org/abs/1703.04908.

[6] Y. Mu, S. Yao, M. Ding, P. Luo, and C. Gan, "EC2: Emergent communication for embodied control," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2023, pp. 6704–6714. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2023/html/Mu_EC2_Emergent_Communication_for_Embodied_Control_CVPR_2023_paper.html.

[7] D. R. Luna, E. M. Ponti, D. Hupkes, and E. Bruni, "Internal and external pressures on language emergence: least effort, object constancy and frequency," *arXiv preprint arXiv:2004.03868*, Apr. 8, 2020. [Online]. Available: https://arxiv.org/abs/2004.03868.

[8] B. de Boer, "Self-organization in vowel systems," *Journal of Phonetics*, vol. 28, no. 4, pp. 441–465, Oct. 2000. DOI: 10.1006/jpho.2000.0125.

[9] A. van den Oord et al., "WaveNet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, Sep. 12, 2016. [Online]. Available: https://arxiv.org/abs/1609.03499.

[10] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "DiffWave: A versatile diffusion model for audio synthesis," in *International Conference on Learning Representations (ICLR)*, 2021. [Online]. Available: https://openreview.net/forum?id=a-xFK8Ymz5J.

[11] X. Tan et al., "NaturalSpeech: End-to-end text-to-speech synthesis with human-level quality," *arXiv preprint arXiv:2205.04421*, May 9, 2022. [Online]. Available: https://arxiv.org/abs/2205.04421.

[12] A. C. L. Yu, "The phonetics of sound change," in *The Handbook of Historical Linguistics*, Wiley, 2020, pp. 291–313. DOI: 10.1002/9781118732168.ch14.

[13] L. Fan et al., "MineDojo: Building open-ended embodied agents with internet-scale knowledge," *arXiv preprint arXiv:2206.08853*, Jun. 17, 2022. [Online]. Available: https://arxiv.org/abs/2206.08853.

[14] S. Kirby, "Learning, bottlenecks and the evolution of recursive syntax," in *Linguistic Evolution through Language Acquisition: Formal and Computational Models*, T. Briscoe, Ed. Cambridge University Press, 2002, pp. 173–204. [Online]. Available: http://www.lel.ed.ac.uk/~simon/Papers/Kirby/

Learning,%20Bottlenecks%20and%20the%20Evolution%20of%20Recursive%20Syntax.pdf.

[15] M. Tamariz and S. Kirby, "The cultural evolution of language," *Current Opinion in Psychology*, vol. 8, pp. 37–43, Apr. 2016. DOI: 10.1016/j.copsyc.2015.09.003.

[16] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, Jul. 20, 2017. [Online]. Available: https://arxiv.org/abs/1707.06347.

[17] A. Wrench, "MOCHA-TIMIT," Centre for Speech Technology Research, University of Edinburgh. [Online]. Available: https://www.cstr.ed.ac.uk/research/projects/artic/mocha.html.

[18] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, Nov. 2006. DOI: 10.1121/1.2229005.

[19] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "VoxCeleb: Large-scale speaker verification in the wild," *Computer Speech & Language*, vol. 60, Article ID.: 101027, Mar. 2020. DOI: 10.1016/j.csl.2019.101027.

[20] Z. Chen, Z. Ai, Y. Ma, X. Li, and S. Xu, "Optimizing feature fusion for improved zero-shot adaptation in text-to-speech synthesis," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2024, Article no.: 28, 2024. Available: https://link.springer.com/article/10.1186/s13636-024-00351-9.

[21] S. Ji, Z. Jiang, W. Wang, Y. Chen, M. Fang, J. Zuo, Q. Yang, X. Cheng, Z. Wang, R. Li, Z. Zhang, X. Yang, R. Huang, Y. Jiang, Q. Chen, S. Zheng, and Z. Zhao, "WavTokenizer: An efficient acoustic discrete codec tokenizer for audio language modeling," in *Proc. ICLR*, 2025. Available: https://proceedings.iclr.cc/paper_files/paper/2025/file/ea1f5f0878d43ff4fb8bf64ef4a2326c-Paper-Conference.pdf.

[22] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "The GRID Audio-Visual Corpus," University of Sheffield, 2006. [Online]. Available: https://spandh.dcs.shef.ac.uk/gridcorpus/#downloads.

[23] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Large-scale speaker recognition dataset," 2018. [Online]. Available: https://www.robots.ox.ac.uk/~vgg/data/voxceleb/vox2.html.

Mir Tahmid Hossain,
Mahsa Sanaei Nourani, Zahra Rahimian,
Dr Md Nawab Yousuf Ali

Received October 16, 2025
Revised January 20, 2026
Accepted January 22, 2026

Mir Tahmid Hossain
ORCID: https://orcid.org/0009-0000-0852-2039
Mid Sweden University
Sundsvall, Sweden
Email: mirtahmid@gmail.com

Mahsa Sanaei Nourani
ORCID: https://orcid.org/0009-0007-8611-6862
Tabarestan University
Mazandaran, Iran
Email: mahsasanaei.n@gmail.com

Zahra Rahimian
ORCID: https://orcid.org/0009-0001-7581-1735
Dr. Shariaty Technical and Vocational College
Tehran, Iran
Email: zehrarahimian@gmail.com

Dr Md Nawab Yousuf Ali
ORCID: https://orcid.org/0009-0009-8069-4527
East West University Department of Computer Science and Engineering
Dhaka, Bangladesh
Email: nawab@ewubd.edu