Thesis for the degree
Master of Science

עבודת גמר (תזה) לתואר
מוסמך למדעים

Submitted to the Scientific Council of the
Weizmann Institute of Science
Rehovot, Israel

מוגשת למועצה המדעית של
מכון ויצמן למדע
רחובות, ישראל

By
Yadin Benyamin

מאת
ידין בנימין

Emotional Speech Synthesis (ESS) using a large pre-trained Text-to-Speech (TTS) Deep Learning Model

סינתזת דיבור רגשי באמצעות מודל טקסט לדיבור (TTS) מבוסס למידה עמוקה

Professor David Harel

פרופסור דוד הראל

October 2025

חשוון התשפ"ו

Abstract

Emotional Speech Synthesis (ESS), the process of generating speech with specific emotional qualities, is a foundational challenge at the intersection of artificial intelligence, speech and language processing, and affective computing. Despite significant progress, it remains constrained by the scarcity of large, naturalistic corpora and by reliance on closed models. We present Whisper Emotional Speech Synthesis (WESS), an open, controllable system that produces affective speech, obtained by fine-tuning the open-source WhisperSpeech (WS) backbone on MSP-Podcast v1.12., a dataset containing ∼250 hours of spontaneous, emotion-annotated speech. WESS follows the WS's two-stage token hierarchy: a text-to-semantic (T2S) Transformer that predicts discrete semantic tokens (content and coarse prosody) and a semantic-to-acoustic (S2A) Transformer that renders acoustic tokens conditioned on speaker embeddings. A pre-trained vocoder then synthesizes the waveform. Control is introduced via two learned prefix tokens that encode emotion category and dominance, yielding a simple, reproducible path to convert a TTS backbone into an ESS system.

We evaluate WESS against gpt-4o-mini-tts with blinded human listening tests across eight emotions, measuring naturalness, perceived intensity, and emotion identification. WESS matches the baseline on naturalness for most emotions with no significant differences in perceived intensity. Identification accuracy is competitive overall, with remaining gaps in anger and sadness synthesis. These results indicate that a compact, open model, trained on a few hundred hours of natural conversational speech, can deliver state-of-the-art perceptual quality with controllable affect. We discuss affective cues across semantic vs. acoustic token streams, limitations (e.g., fine-grained intensity control and English-only evaluation), and avenues for improvement via richer supervision, cross-corpus testing, and multilingual adaptation.

Project page & audio samples: click here.

Contents

1 List of abbreviations

ASR – Automatic Speech Recognition
CNN – Convolutional Neural Network
DL – Deep Learning
ECAPA-TDNN – Emphasized Channel Attention, Propagation and Aggregation Time-Delay Neural Network
ESC – Environmental Sound Classification
ESS – Emotional Speech Synthesis
FDR – False Discovery Rate
GAN – Generative Adversarial Network
HCI – Human–Computer Interaction
IU – Intonation Unit
KL – Kullback–Leibler (divergence)
LLM – Large Language Model
LUFS – Loudness Units relative to Full Scale
MOS – Mean Opinion Score
MSP-Podcast – Multimodal Signal Processing Podcast Corpus
MTurk – Amazon Mechanical Turk
S2A – Semantic-to-Acoustic (module)
SBC – Santa Barbara Corpus
SOTA – State of the Art
STT – Speech-to-Text
T2S – Text-to-Semantic (module)
TTS – Text-to-Speech
VQ – Vector Quantization
WESS – Whisper Emotional Speech Synthesis
Whisper – Web-Scale Supervised Pretraining for Speech Recognition
WS – WhisperSpeech

2 Introduction

Human–Computer Interaction (HCI) shapes how individuals engage with devices and software on a daily basis. As these systems continue to evolve - from personal mobile assistants to sophisticated conversational agents - recent research focuses on creating more natural and human-like modalities of communication [1]. One of the most intuitive ways to achieve this goal is through speech - a primary channel for expressing not only content but also emotion and intent.

As a practical foundation for voice interfaces, text-to-speech (TTS) technologies have drastically advanced in the past decade, primarily due to deep learning (DL) breakthroughs [2]. Representative approaches that drove this progress include convolutional neural networks (CNNs) [3], sequence-to-sequence (seq2seq) models [4], adversarial generative (GAN) methods [5], and Transformer-based architectures [6]. Contemporary TTS systems can produce remarkably natural-sounding synthetic speech, nearing, if not matching, the clarity and ease of comprehension found in real human speech.

Despite these gains, standard TTS systems are often limited to neutral prosody. Prosody, the music of speech, encompasses variations in fundamental frequency (also known as pitch), duration, intensity, and voice quality, and plays a crucial role in the expression and interpretation of non-verbal messages: emotions, intentions, and linguistic meaning [7]. A central aspect of prosody with particular importance for TTS is emotion: it shapes listeners' perception of speaker intent, empathy, and conversational appropriateness. Psychological theoretical models, such as Plutchik's wheel, organize emotions into primary categories and their more nuanced manifestations **(Fig. 1)**— providing a useful conceptual target for controllable synthesis [8]. Such models also quantify different qualities of emotion manifestation, most typically valence, dominance, and arousal [9, 10, 11, 12]. These are the considerations that motivate explicit modeling of emotion within TTS, methods that are known collectively as Emotional Speech Synthesis (ESS) methods.

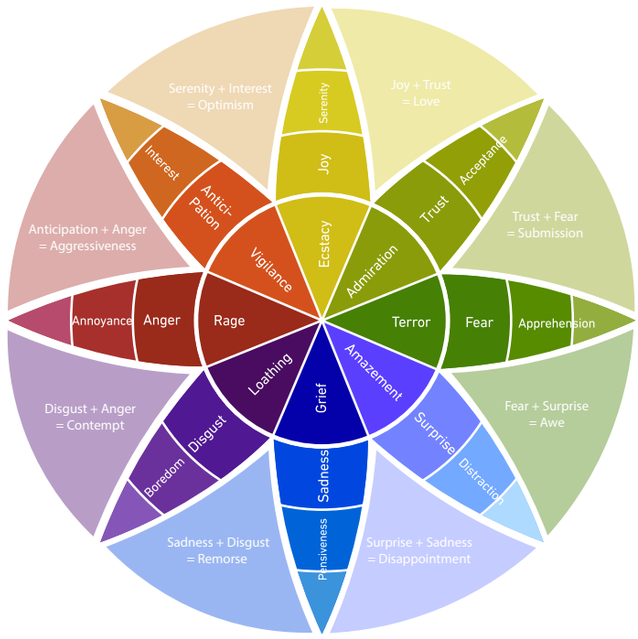

Figure 1: **Plutchik's emotion wheel.** Emotions arise as mixtures or derivatives of eight primary emotions [8].

ESS extends TTS by modulating prosodic cues to render specific emotional states [7]. Over time, ESS has evolved from early DL networks such as seq2seq with explicit emotion labels to generative approaches (e.g., GAN) and, most recently, pipelines guided by large language models (LLMs)- that plan prosody and speech style from explicit prompts. Recent models, based on such architectures, have achieved striking results leveraging vast training corpora and sophisticated conditioning schemes [13, 14, 15]. Yet, two obstacles still delay widespread adoption. First, publicly available emotional-speech datasets are typically recorded by professional actors who over-articulate their affect. Widely used corpora [16, 17, 18, 19, 20, 21] therefore capture stylized, exaggerated emotions that differ from everyday expression, and are often of inconsistent recording quality. Secondly, these datasets are orders of magnitude smaller than those used in neutral TTS. Whereas leading TTS datasets contain hundreds of thousands of hours of speech, emotional corpora rarely exceed a few dozen hours. Since state-of-the-art (SOTA) deep learning models depend on large, diverse training sets to generalize well, this data scarcity has become a critical bottleneck in developing high-quality expressive synthesizers. It is this two-fold challenge, of scale and of naturalness, that we aim to meet.

It may be argued that the recent development of LLMs and their application to TTS have already somewhat resolved these problems. For example, OpenAI recently introduced gpt-4o-mini-tts, an LLM-driven framework that synthesizes highly expressive speech directly from textual prompts, providing fine-grained control over style, accent, and affect. However, it currently exposes only a small, curated set of preset voices (eleven at the time of writing) [22]. Although the resulting audio reaches state-of-the-art naturalness, the model's scale, amounting to billions of parameters [23], its proprietary training data, and the absence of open weights limit its transparency, reproducibility, and broad adoption. Thus, the problem of developing an open-source ESS model that was trained on public datasets and would be accessible and reproducible by all, stands.

As mentioned above, our approach to solving this problem is two-fold. First, we leverage the recently introduced MSP-Podcast corpus. By harvesting naturally occurring conversations from public podcasts and annotating them for emotion, the corpus supplies a relatively extensive repository of spontaneous speech, serving as a promising basis for ESS research [24, 25]. Although the corpus is the largest open-source labeled emotional dataset currently available, it is still minuscule compared to the massive corpora that neutral TTS models are retrained on, which motivates the second, and more fundamental, part of our proposal.

Unlike ESS, the neutral TTS community already maintains large, open-source, pre-trained backbones. If one can transform such a backbone into an open, controllable ESS system using today's small, naturalistic emotion datasets, this would simultaneously provide a reproducible alternative to closed models and a practical solution to the data-scarcity barrier.

We thus created an ESS model by fine-tuning the TTS model WhisperSpeech (WS) [26] on the MSP-Podcast corpus, yielding Whisper Emotional Speech Synthesis (WESS). Despite being an order of magnitude smaller than OpenAI's gpt-4o-mini-tts, the resulting model, WESS, yields on par perceptual quality in our evaluations, indicating that carefully curated data and targeted adaptation can close much of the gap to state-of-the-art LLM-driven TTS.

3 Goals

Hypothesis

A pre-trained neutral TTS model can be transformed into an ESS system that provides controllable emotion and its intensity through targeted transfer learning using a relatively small amount of naturalistic emotion data. Moreover, a compact ESS model, roughly an order of magnitude smaller than current LLM-based systems, can achieve on-par perceptual quality when paired with careful data curation and lightweight conditioning.

Primary goals

- Demonstrate a practical pathway for converting a neutral TTS backbone into an ESS model through fine-tuning and minimal architectural changes with a minimal amount of data.
- Enable explicit control over emotion category and its level of intensity, and validate that the control is reliable and perceptually meaningful.
- Train and evaluate a compact ESS model and compare it to a state-of-the-art baseline.

Scientific importance

Establishing that expressive, controllable speech may be produced by a smaller model has direct implications for real-world HCI scenarios. Compact ESS systems lower compute and energy resources, improve latency, and enable on-device operation - benefiting privacy, accessibility, and sustainability - while enhancing reproducibility.

4 Methods

4.1 Models

As mentioned above, our approach leverages a pre-trained, open-source TTS model, WS. WS is based on a large open-source speech-to-text model called Whisper. The following two subsections describe these two models, providing some essential technical background for understanding our method.

4.1.1 Web-scale Supervised Pretraining for Speech Recognition (Whisper)

Speech-to-Text (STT, also called automatic speech recognition, ASR) maps an input waveform to a textual transcript and related metadata (e.g., timestamps and language ID). It is the inverse task of TTS, it provides complementary modeling tools and representations that can be repurposed for speech synthesis.

Figure 2: **Whisper pipeline.** An audio waveform is encoded by the Whisper encoder to a time-ordered sequence of latent vectors, which the Whisper decoder maps to text tokens to produce the transcript.

Within STT, OpenAI's Whisper is a multilingual, multitask encoder–decoder Transformer **(Fig. 2)** trained on roughly 680,000 hours of weakly supervised speech collected from the web [27]. This scale, together with joint training across tasks (transcription, speech translation, language identification, and timestamp prediction), yields strong zero-shot robustness to domain shifts and background noise. Whisper is released with open weights and code in multiple model sizes (from `tiny` ≈39M to `large` ≈1.5B parameters), making it a widely adopted baseline for STT and a practical foundation for downstream research.

Since Whisper already learns a powerful speech–text interface (log-Mel spectrogram front end, sequence modeling, and tokenization), a natural next step is to invert this interface for synthesis; that is, condition a generator on text to produce expressive speech using Whisper-derived representations. This idea is applied in WS [26], an open-source TTS project that "inverts" Whisper. We leverage this tool to create our ESS system, as described in the following subsection.

4.1.2 WhisperSpeech: "Inverting" Whisper for TTS

WS TTS system repurposes OpenAI's Whisper ASR by "inverting" its speech–text interface into a text–speech pipeline [26]. The design follows a two-stage speech generation pipeline: first, producing semantic tokens (content & high-level prosody). Secondly, rendering these into acoustic tokens that a vocoder converts to a waveform. With this design in view, we explain how speech is represented inside the model through these two types of tokens.

The above-mentioned semantic tokens are learned representations of the input text, which encodes the linguistic content and long-range prosodic structure (intonation contours, stress patterns, phrasing) in a speaker-independent way. They are obtained by encoding audio with a vector-quantized (VQ) version of Whisper's encoder, VQ [28, 29] being the process of learning discrete codebook elements instead of keeping continuous real-valued representations. This process yields semantic tokens that contain universal prosodic information, compactly represented, and thus easy to predict autoregressively. Similarly to semantic tokens, acoustic tokens are discrete codebook entries that represent crucial information for speech generation; however, unlike semantic tokens, they encode timbre and fine acoustic detail, governing the sound of the generated speech. A speaker embedding is added to the acoustic tokens to condition an audio rendering that clones the specific properties of a given speaker. In a way, semantic tokens represent "reading" (what is said and how it is said), whereas acoustic tokens represent "speaking" (how it sounds) [30, 31].

In more detail, the semantic tokens are derived from an implementation of Whisper's encoder paired with a VQ bottleneck, which are trained together to approximate the behavior of the full Whisper encoder. They operate at a fixed rate of 50 Hz (50 tokens/s). The acoustic tokens are trained to map discrete elements of an external, pre-trained VQ acoustic compression model— EnCodec [32]. These, on the other hand, operate at 24 kHz and at 1.5 kbps, yielding 150 discrete tokens per second. The differences in frequency between the two types of tokens stem from the differences in the nature of the two represented signals, linguistic information versus raw audio. By separating these levels of speech, WS can mix content with voices: semantic tokens ensure transcript/intonation fidelity, while acoustic tokens (guided by a speaker embedding) ensure the voice and audio texture are correct. This two-stage tokenization was popularized by AudioLM and SPEAR-TTS, which showed that long, coherent speech generation is easier when first generating a high-level semantic plan, then refining it acoustically. WS follows this paradigm, using Whisper for the semantic phase [31, 30, 26].

With the representations in place, the system architecture is as follows: the text-to-semantic (T2S) module is a Transformer model that maps text to semantic tokens, and the semantic-to-acoustic (S2A) module is a Transformer model which then uses them and the speaker embeddings to predict acoustic tokens. Speaker embeddings, a representation of the characteristics of the desired specific voice, are derived using another externally pre-trained model, ECAPA-TDNN [33], designed for speaker recognition. The speaker embedding is added to the acoustic tokens, which are then passed on to a vocoder, a pre-trained model named Vocos [34], to generate the final audio waveform at 24 kHz. The full pipeline is described in Fig. 3.

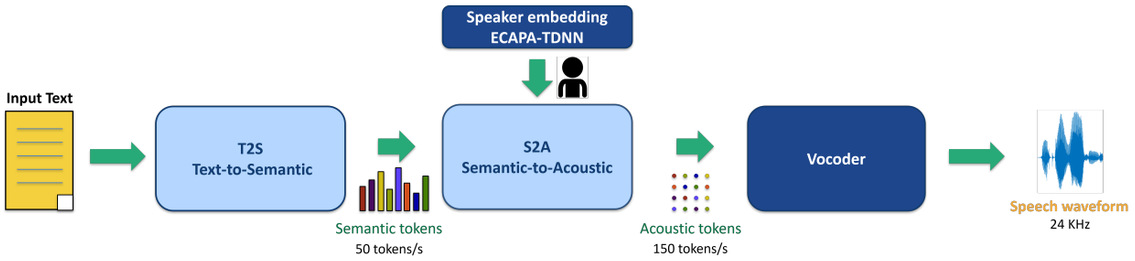

Figure 3: **Modular text-to-speech pipeline.** T2S maps text to semantic tokens. S2A predicts EnCodec acoustic codes from semantic tokens conditioned on a speaker embedding. A pre-trained vocoder synthesizes a 24 kHz waveform.

WS is trained primarily on LibriLight, an extensive public dataset comprising 60,000 hours of English audiobooks, providing paired samples of text and speech [35]. Training is implemented in three separate and consecutive stages [26]: **(Stage 1)** Training the VQ Whisper encoder using knowledge distillation from the original Whisper encoder to the VQ version created by WS [36]. **(Stage 2)** Training the T2S to produce semantic tokens from the input text that match the ones generated by the VQ encoder on the paired speech of the same samples **(Stage 3)** Training the S2A to predict acoustic tokens from the semantic tokens, where the ground truth acoustic tokens are obtained by passing the speech data samples through EnCodec [32].

4.2 WESS: Fine-tuning WS on the MSP-Podcast

We adapt WS to ESS by fine-tuning all trainable modules on MSP-Podcast release 1.12 to create WESS (see Section 5.1 for details regarding data and preprocessing). As mentioned above, WS was originally trained on LibriLight audiobooks [35, 26]. MSP and LibriLight share long English speech and diverse recording conditions, which ease transfer and make MSP a suitable corpus for adaptation.

In addition to training on the MSP dataset instead of LibriLight, our fine-tuning procedure introduces another important modification. To encode emotional information, we concatenate special tokens [37] to the input text that encodes both emotion class and dominance, thus directing the generation of emotional semantic tokens. Whereas the optimization of the two other modules, the VQ Whisper encoder and the S2A module, remains similar to the original WS training procedure, the special tokens modify the T2S module traning, as described below. Key training settings and hyperparameters for all modules are summarized in Table 1.

Stage 1: VQ Whisper encoder on MSP. We fine-tune the VQ Whisper encoder on MSP in order to produce 50 Hz semantic tokens that are robust to speaker and channel variation. The objective is a composite student–teacher loss that includes token cross entropy to the ground truth transcript, a KL divergence term that distills the output distribution of the original unquantized Whisper decoder into the VQ-augmented student, and a standard VQ commitment penalty - to stabilize codebook usage [29, 36, 38]. This stage mirrors the WS procedure while adapting the encoder to MSP speech characteristics (Fig. 4a).

Stage 2: T2S with emotion and dominance tokens. The T2S model is a Transformer encoder–decoder trained to map text to semantic token sequences with a cross-entropy loss. We prefix two learned special tokens to every transcript (an instance of input text). One special token encodes the categorical emotion, and another encodes the discrete dominance level (as labeled in the MSP corpus). This follows the common practice of prefix tokens for

conditioning, as in classification tokens in Transformer language models and task or language tokens in Whisper [27, 37]. The model learns internal representations for each emotion and dominance setting and thus produces distinct prosodic plans conditioned on these tokens (Fig. 4b). Note that although this is a supervised learning procedure, the internal representation of each emotion is learned by the model and does not include any hand-crafted features or prosodic information regarding the various emotions.

Stage 3: S2A adaptation on MSP. The S2A model is trained to predict EnCodec acoustic tokens from semantic tokens and an ECAPA-TDNN speaker embedding using cross-entropy over the discrete codebooks [32, 33]. This stage follows WS but uses MSP audio to adjust to the distribution shift from audiobooks to podcasts and to improve naturalness (Fig. 4c). This training stage enables WESS to generate any chosen voice.

Stage	Pre-trained model	Params	Batch	LR
VQ Whisper encoder	Medium (EN & PL)	22.7 M	32	1.84×10^{-7}
T2S	Small (EN & PL)	214.1 M	64	6.31×10^{-7}
S2A	Tiny (EN & PL)	20.1 M	32	4.23×10^{-5}
Total trainable parameters: 256.9 M				
Hardware: single NVIDIA H100 NVL (96 GB)				

Table 1: Key training settings and module sizes for WESS fine-tuning on MSP-Podcast v1.12.

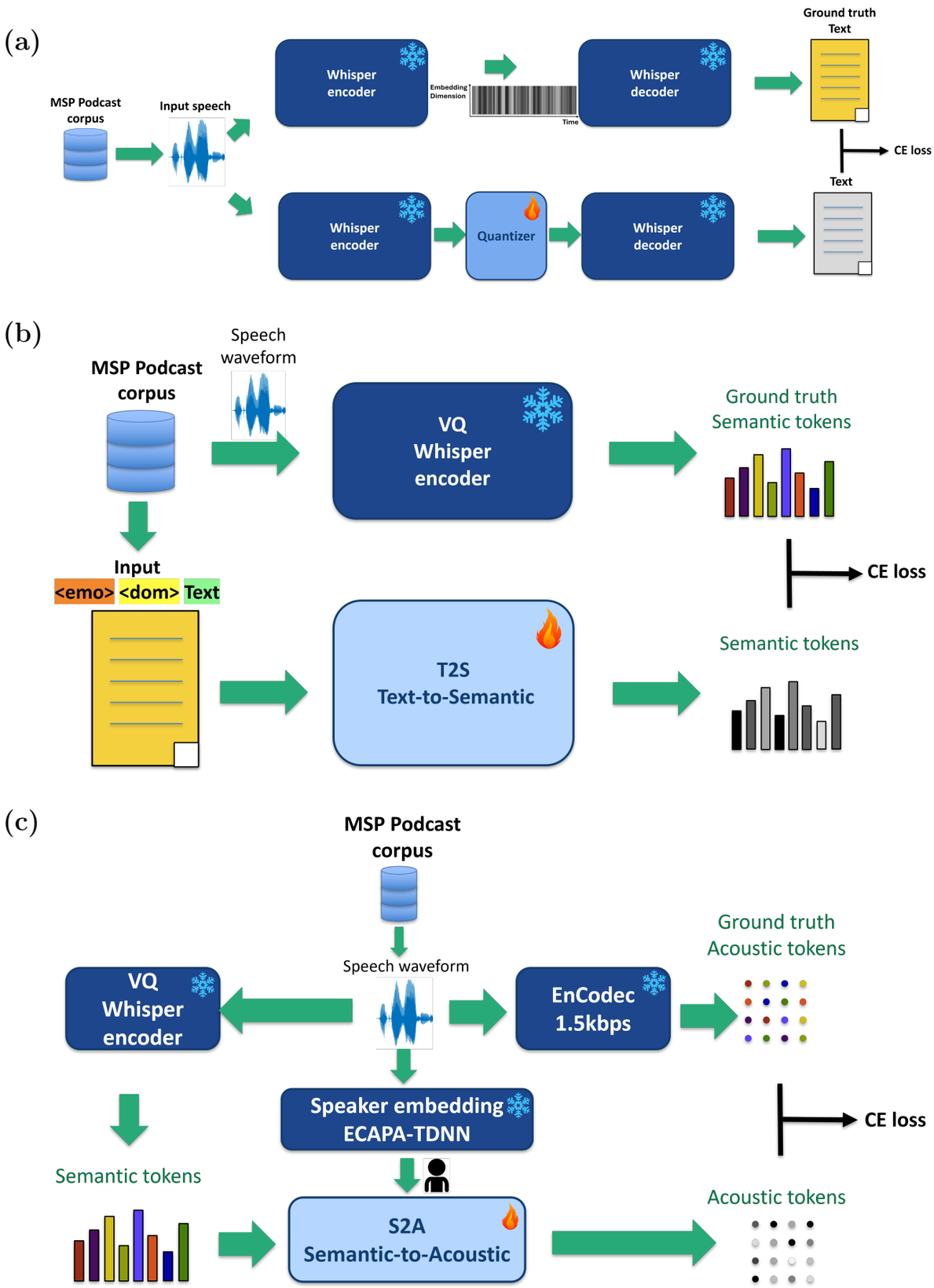

Figure 4: **The fine-tuning process of WESS.** (a) VQ Whisper encoder training on MSP. The student VQ-augmented encoder–decoder is trained with cross-entropy, KL distillation to the original Whisper decoder, and a VQ commitment term. (b) T2S training with special tokens. Two learned tokens encode emotion class and dominance and are prepended to the transcript before predicting semantic tokens. (c) S2A training on MSP. The model predicts EnCodec acoustic codes from semantic tokens and an ECAPA-TDNN speaker embedding. In all sub-figures, flame icons represent trainable modules, and snowflake icons represent frozen modules.

4.3 Using WESS for ESS and Voice Cloning

Following the WS fine-tuning described in Section 4.2, the final WESS model is able to generate audio via two different pipelines. The first, which was the main focus of the work presented above, is **emotional speech synthesis** (ESS). Given a reference text and emotional information (both category and dominance), the model generates speech that combines the text in a given emotion. This synthesis pipeline is shown in Fig. 5 and adheres to the two-stage WS design [26] described earlier. Results and comparisons to a SOTA model are presented in Section 6.

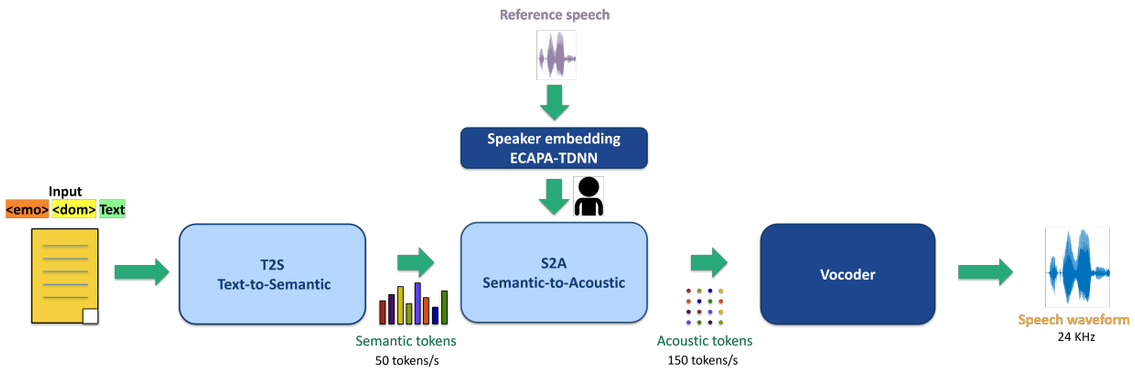

Figure 5: **WESS pipeline.** Text is mapped to semantic tokens by T2S. S2A produces acoustic tokens conditioned on a speaker embedding. A pre-trained vocoder synthesizes the waveform.

Importantly, this structure enables a small extension for **voice cloning** that incorporates speaker style transfer: at inference, the semantic and prosodic information of a source audio can be paired with a target speaker embedding to restore their timbre, while transferring prosody and emotion from the source. To do so, we feed a reference audio clip instead of text, modifying the standard pipeline: the T2S stage is replaced by the VQ-Whisper encoder that encodes the audio into semantic tokens [26], similarly to the T2S module's encoding of text. Crucially, using voice as input preserves the reference's fine-grained prosody-timing, emphasis, and pauses, which cannot be determined when only text is provided. By combining this preserved prosodic trajectory with a target speaker embedding, WESS produces the same style in a new emotion, enabling cross-speaker style transfer while retaining the source prosody (Fig. 6). Since cloning was not the primary focus of this research, we did not conduct a dedicated human listening evaluation of cloning fidelity or emotion preservation. We thus leave a targeted cloning study to future work.

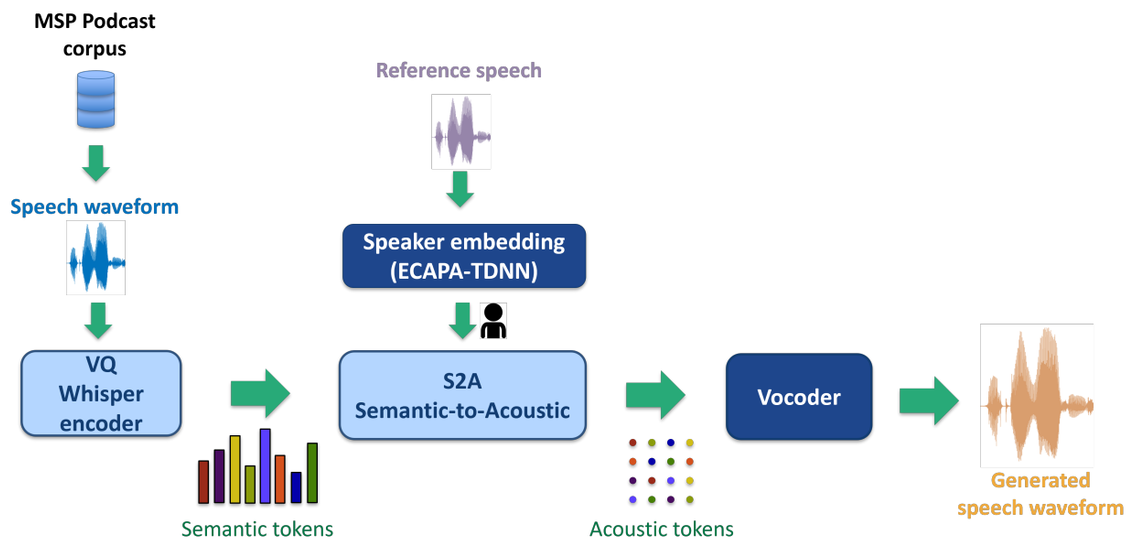

Figure 6: **WESS voice cloning with emotion pipeline.** Left: re-voicing of a reference clip by swapping the speaker embedding while preserving the semantic tokens. Right: generation of a new text with the same emotion by setting the emotion and dominance prefix tokens and conditioning on the target speaker embedding.

5 Experimental Setting

5.1 MSP-Podcast Dataset

The MSP-Podcast corpus contains spontaneous speech retrieved from Creative Commons podcasts and segmented into speaker turns for affective research [24]. The Odyssey 2024 challenge used a subset of release 1.11 and provides a public description of the splits and labeling protocols [25]. We work with release 1.12, which comprises approximately 320 hours of emotion-labelled speech [39]. From this set, we exclude items that were labeled as "no agreement" (X) and "other" (O), leaving ˜250 hours that were used for our experiments. Each utterance includes categorical labels over eight primary emotions: anger, happiness, sadness, fear, surprise, contempt, disgust, and neutral. Additionally, it contains dimensional annotations for arousal, valence, and dominance on a seven-point scale [24, 25]. In our experiments, we use the eight-class primary emotion set and the dominance attribute (Fig. 7). Dominance captures the perceived degree of speaker control and social power in an interaction. It corresponds to arousal and valence in MSP-Podcast, where it is annotated on a seven-point scale and used as a target attribute in the Odyssey challenge. Dominance helps disambiguate emotions that share similar arousal or valence patterns, for example, anger and fear [24, 25].

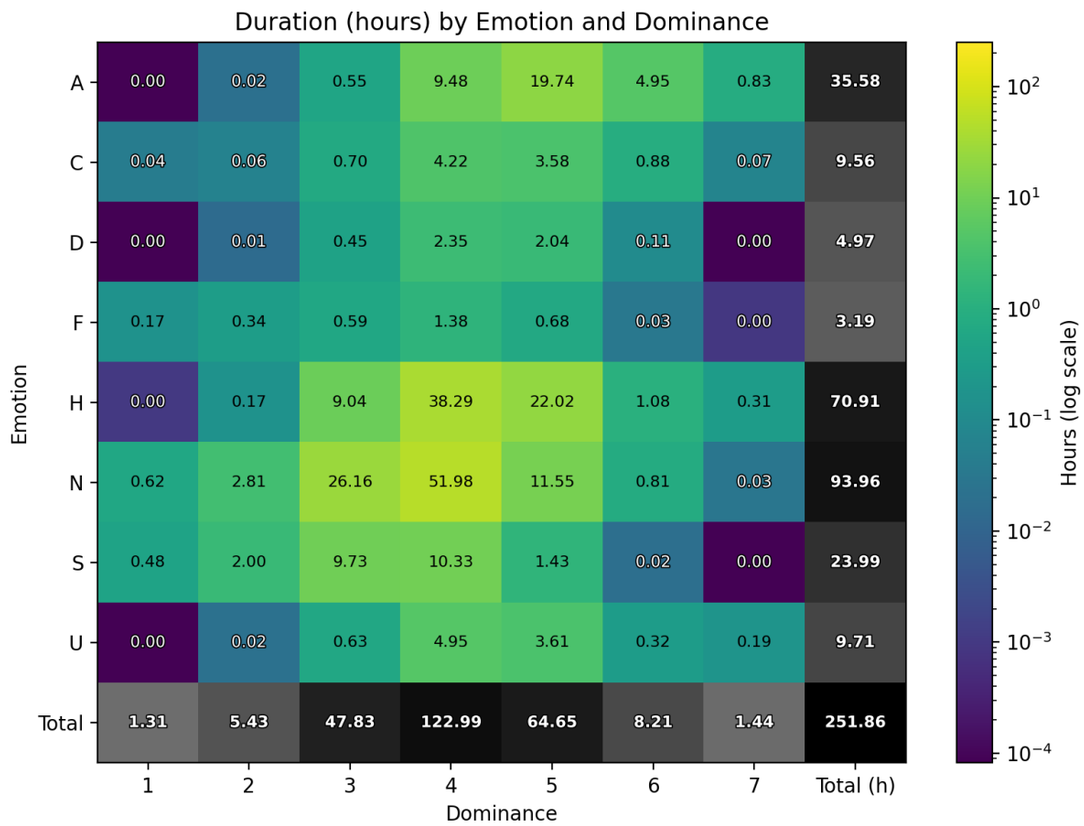

Figure 7: **Duration in hours by primary emotion and dominance level for the MSP-Podcast v1.12 subset used in this work.** Letter codes: A - Angry, S - Sad, H - Happy, U - Surprise, F - Fear, D - Disgust, C - Contempt, N - Neutral. Values are shown on a logarithmic color scale.

For preprocessing purposes, the data were cleaned and standardized before modeling. Audio was used at its native sampling rate and loudness-normalized per utterance to a target of −14 Loudness Units relative to Full Scale (LUFS). From the normalized signal, we derived three representations for downstream models. First, discrete semantic tokens at 50 Hz from a VQ Whisper encoder after resampling to 16 kHz. Second, EnCodec acoustic tokens after resampling to 24 kHz at the 1.5 kbps setting with sequences limited to a maximum of 30 s [32]. Third, a 192-dimensional speaker embedding extracted with the ECAPA-TDNN pipeline for conditioning [33]. This setup follows the WS formulation that separates semantic and acoustic streams for controllable synthesis [26].

5.2 Human Evaluation

To evaluate the performance of our method, we conducted human evaluation surveys, allowing independent listeners to judge various qualities of the generated speech. The surveys evaluated the performance of our method with that of an SOTA method on two standard ESS criteria: audio naturalness and perceived emotion intensity [1, 40], collected as Mean Opinion Scores (MOS) on five-point Likert scales [41, 42]. In addition, we measured emotion identification using a blind assessment of the target label, a test that is stricter vis-à-vis affect control and which is less common in the ESS literature.

As mentioned above, OpenAI's gpt-4o-mini-tts [22] is currently considered to be the SOTA model for TTS. Although not designed specifically for ESS, it shows very strong capabilities as a result of its free-form prompt formulation that allows its users to condition the generated speech on any natural language instructions [22]. Recent works of ESS have been repeatedly evaluated against this model and have shown it is also on par with SOTA in the ESS task [43, 44, 45], thus our evaluation survey includes generated speech from our method and gpt-4o-mini-tts.

Each survey version contained 32 clips that were created from a single sentence, repeated and produced with eight emotions, in two speaker genders, by the two models,. There were four survey versions in total, one per sentence. To make as reliable a comparison as possible, we set the emotions of the two models in a similar fashion. For WESS, we used the learned prefix tokens, setting the emotion token to the target category and the dominance token to that category's most frequent dominance level in the MSP-Podcast dataset (see Fig. 7). For gpt-4o-mini-tts, which has no explicit emotional category or intensity control, we supplied a simple style prompt such as "Emotion: Disgust. Read naturally with that feeling." As gpt-4o-mini-tts provides a fixed set of voices, we used Alloy for the female voice and Ash for the male. To obtain comparable voices for WESS, we extracted a speaker embedding from a reference utterance of each of these two voices using an ECAPA-TDNN encoder [33], searched the MSP-Podcast dataset for the nearest speaker to this embedding in the speaker embedding space, and conditioned WESS on that speaker. This matching reduces timbral biases that might make clips easier to tell apart across models [33].

The sentences used for generating the audio clips are as follows (n denotes the number of participants who completed the 32-item survey for that sentence):

- S1 "A train passed beyond the distant fields." (n=14)
- S2 "A bicycle rolled past the cafe window." (n=17)
- S3 "I can't believe it. This changes everything." (n=20)
- S4 "It finally happened, he came back." (n=24)

The setup of the survey required that after listening to each audio clip (a generated speech of a certain model, gender and emotion), the listener answer three questions. The questions and their response formats are as follows:

- **Q1 - Audio naturalness:** "How human-like and natural was the audio?" Responses on a five-point Likert scale, where 1 means "bad - completely synthetic" and 5 means "excellent - completely natural" [46].
- **Q2 - Speaker emotion:** "Which emotion best describes the clip?" Responses were one of the eight categories used in the study. Namely: anger, happiness, sadness, surprise, contempt, disgust, fear, and neutral.
- **Q3 - Emotion dominance:** "How strongly is that emotion expressed?" Responses on a five-point Likert scale, where 1 means very weak and 5 means very strong.

We recruited participants on the Amazon Mechanical Turk (MTurk) web platform [47]. Each worker completed exactly 32 judgments. We excluded submissions with total correct label counts below 8 or above 28 out of 32, in order to remove near-random and likely automated responses, respectively. Deduplication and completeness checks were enforced before analysis. See Appendix A for rater-screening details and the accuracy-distribution histogram.

5.3 Statistical analysis

Naturalness MOS and intensity MOS were summarized per model using the sample mean with two-sided 95% confidence intervals based on the t distribution. Paired comparisons between models used two-sided paired t-tests on matched clips. We adjusted p-values using the Benjamini–Hochberg procedure to control the false discovery rate (FDR) [48]. Emotion identification accuracy was computed as the proportion correct with Wilson 95% confidence intervals [49]. Model comparison for accuracy used McNemar's test for paired binary outcomes [50]. All tests used $\alpha = 0.05$ and follow recommendations for reproducible crowdsourced speech quality evaluation [41, 42, 51].

6 Results

Overall, WESS performs on par with gpt-4o-mini-tts across most emotions. For naturalness, the systems are indistinguishable for seven of the eight emotions, with a significant advantage for gpt-4o-mini-tts on happiness (Fig. 8). Emotion dominance MOS shows no significant differences for any emotion (Fig. 9). For emotion identification accuracy, gpt-4o-mini-tts performed better on anger and sadness, while the remaining emotions show no reliable differences (Fig. 10). As mentioned above, differences are judged with paired tests and 95% confidence intervals with FDR control at $\alpha = 0.05$.

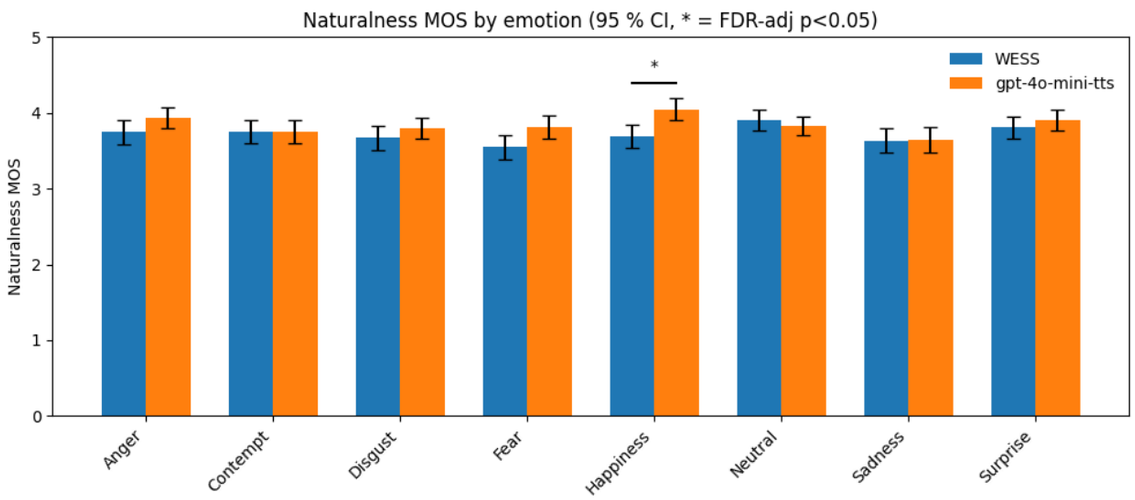

Figure 8: **Naturalness MOS by emotion.** gpt-4o-mini-tts exceeds WESS only for Happiness. Bars show MOS with 95% confidence intervals. Stars mark emotions where the paired t-test indicates a significant difference after FDR adjustment ($p < 0.05$).

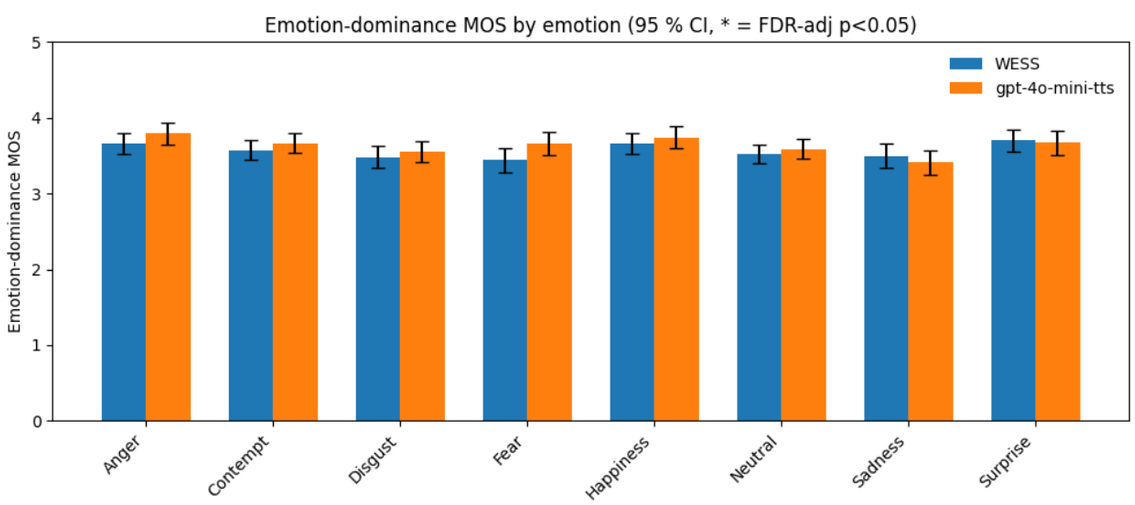

Figure 9: **Emotion dominance MOS by emotion.** Mean scores with 95% confidence intervals. Following FDR adjustment, no emotions show significant differences between models.

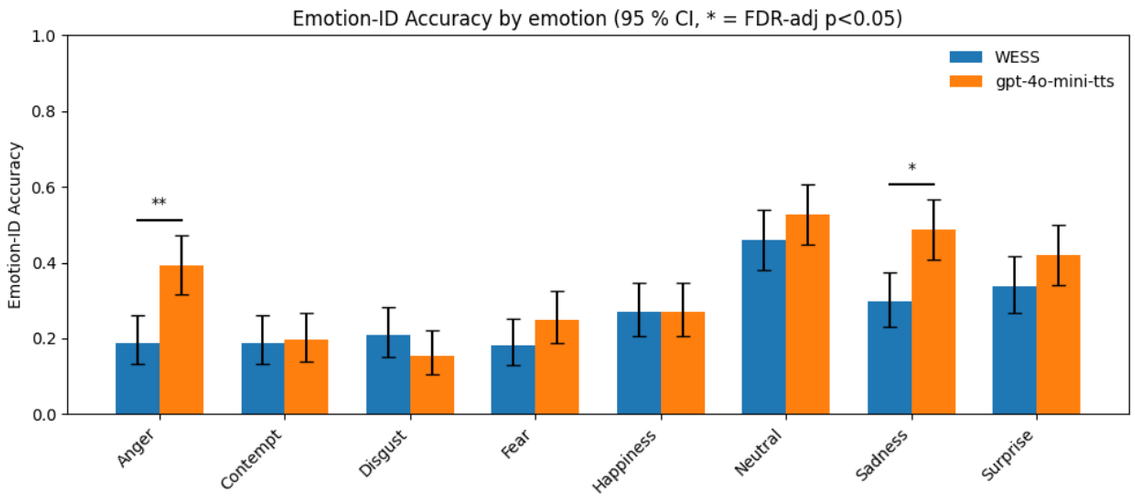

Figure 10: **Emotion identification accuracy by emotion.** Proportions correct with Wilson 95% confidence intervals. Stars mark emotions with a significant difference by McNemar's test after FDR adjustment ($p < 0.05$). gpt-4o-mini-tts is higher for Anger and Sadness. The other emotions are statistically comparable.

In summary, despite the modest size of the model as well as the training set of publicly available data, WESS matches the baseline on most emotions for naturalness, exhibits comparable perceived intensity, and is competitive on emotion identification, with gaps limited to anger and sadness. These results support the feasibility of converting an open neutral TTS backbone into an expressive system with strong perceptual performance and controlled emotion.

7 Discussion

Our results support the central hypothesis that a pre-trained neutral TTS backbone can be transformed into an effective ESS system with limited naturalistic supervision. Despite using only ∼250 hours of emotional speech (MSP-Podcast v1.12) for fine-tuning, roughly 0.4% of the ∼60,000 hours used to train WS and about 0.04% of Whisper's original pretraining scale, WESS matches a SOTA reference (gpt-4o-mini-tts) on most perceptual criteria [35, 26, 27]. These findings suggest that separating the representation of speech into semantic and acoustic token streams is a strong inductive bias for affect control, and that lightweight conditioning with emotion and dominance prefix tokens can compensate for orders of magnitude less data.

Compared to large proprietary systems, WESS is compact (approximately two orders of magnitude smaller), open, and reproducible. It can synthesize arbitrary, individual voices by conditioning S2A on an ECAPA-TDNN, as embedding extracted from a reference utterance, which enables broad voice coverage without retraining [33]. In contrast, gpt-4o-mini-tts -although it reaches very high naturalness - makes available only a small, curated set of preset voices [22]. The open-source nature and size of WESS lower compute costs and enable on-device or private deployments wherever feasible.

Adapting an open TTS backbone with a few hundred hours of spontaneous podcast speech offers a practical alternative to large, closed models. Our attempts with substantially smaller, acted-speech corpora produced unstable results and weaker control, which aligns with reports that acted datasets can misrepresent everyday emotion and are too small for present-day pipelines. Continued growth of naturalistic corpora like MSP-Podcast will likely close the remaining gaps.

Regarding evaluation of models in the field of ESS, there is no widely accepted automatic metric that quantifies human perception of expressive speech across emotions. In current literature, many report automatic scores from a speech emotion recognizer, sometimes even trained on the same dataset as that of the synthesizer. This risks circularity and misalignment with human judgments [52, 53]. We therefore relied on blinded human listening tests with MOS for naturalness and intensity and with 8-way identification accuracy, following current recommendations for subjective speech evaluation [41, 42]. This stricter design assesses actual affect control instead of recognition against a disclosed target.

Crowdsourcing introduces variability in attention, language proficiency, and device quality. We mitigated this with completeness checks and exclusion thresholds that removed both near-random and suspiciously perfect responders. The distribution of per-user accuracy and the applied inclusion thresholds are shown in Fig. A1. Details of the screening procedure appear in Appendix A. A sizable fraction of submissions was excluded, which reduces the nominal sample size but improves the reliability of retained judgments.

Fine-grained control over intensity and micro-prosody remains limited. Our dominance prompt provides coarse intensity shifts, yet continuous and reliably calibrated control needs richer supervision or preference-learning objectives [1]. Absolute identification accuracies leave room for improvement, with the best categories near 60%, and with clear gaps for Anger and Sadness. English-only evaluation is another limitation that calls for multilingual adaptation.

Regarding further elaboration of our proposed methods, we see three immediate future directions. (i) **Richer supervision**: continuous affect trajectories and preference learning for calibrated intensity of emotion [1]. (ii) **Generalization**: cross-corpus testing and multilingual adaptation, along with evaluator-agnostic reporting that combines human MOS with independent automatic judges. (iii) **Ablations and analysis**: token-stream probing for affect localization and systematic studies of prefix-token design and label granularity.

In summary, we present a simple and reproducible method that converts a neutral, open-source TTS backbone into a controllable ESS system using learned special tokens for emotion category and dominance. With careful data curation and lightweight conditioning, a compact open model can approach the perceptual performance of much larger proprietary systems while remaining transparent and broadly accessible.

Appendix A: Statistical analysis of results

This section presents the rater-screening procedure and presents the per-user accuracy distribution obtained from MTurk (Fig. A1).

Each participant answered $n = 32$ independent 8-way forced–choice items for emotion identification. Under the null hypothesis (i.e., that a respondent is guesses uniformly at random), the probability of a correct choice on any single item is $p_0 = 1/8$. Let X be the number of correct responses for a given rater. Then

$$X \sim \text{Binomial}(n = 32,\ p_0 = 1/8), \qquad \Pr(X = k) = 32k\, p_0^k (1 - p_0)^{32-k}.$$

We test

$$H_0 : p = p_0 \quad versus \quad H_1 : p > p_0$$

with a one–sided binomial test [54, 55]. The critical region is $X \geq c$ where c is the smallest integer such that $\Pr_{H_0}(X \geq c) \leq \alpha$ for a chosen significance level α.

Evaluating the binomial tail under H_0 gives

$$\Pr_{H_0}(X \geq 8) = \sum_{k=8}^{32} 32k p_0^k (1 - p_0)^{32-k} \approx 0.0395 < 0.05,$$

while

$$\Pr_{H_0}(X \geq 7) \approx 0.0965.$$

Thus $c = 8$ controls the type I error at $\alpha \approx 0.04$, whereas $c = 7$ would not meet a 5% threshold. Equivalently, $X = 8$ is about 2.14 standard deviations above the null mean $\mathbb{E}[X] = np_0 = 4$ with $\text{sd}(X) = \sqrt{np_0(1 - p_0)} \approx 1.87$. We therefore retain raters with $X \geq 8$ and exclude those with $X \leq 7$, since achieving at least 8 correct out of 32 is unlikely under random guessing at the 5% level.

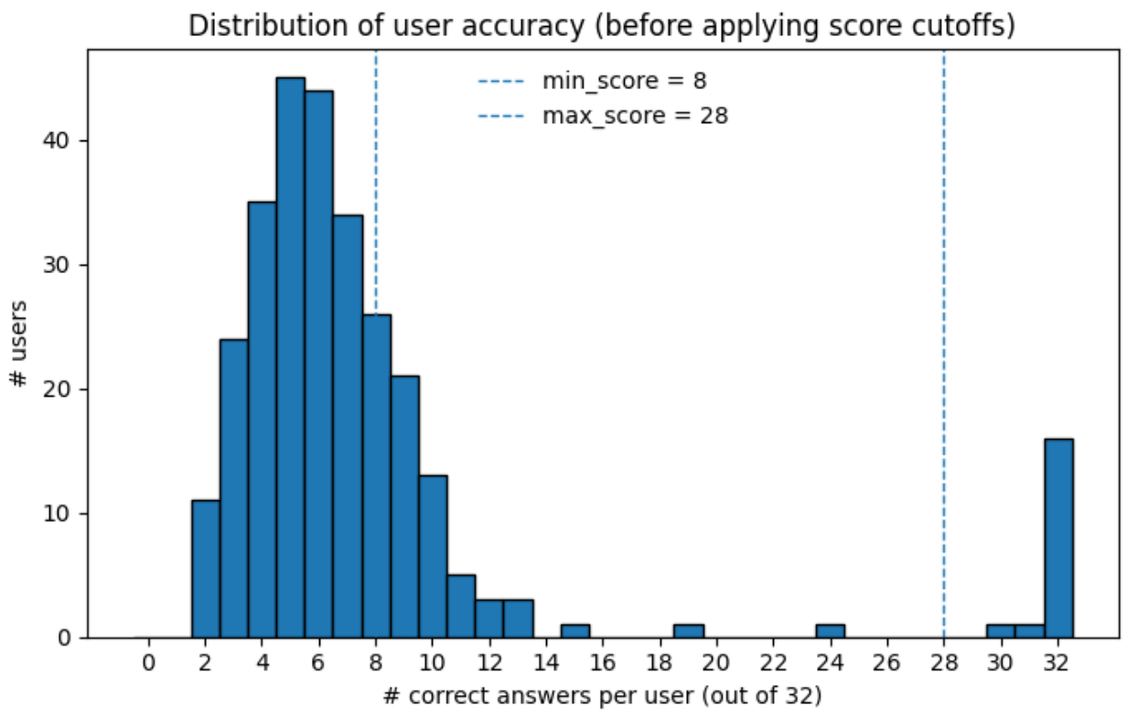

Figure A1: **Distribution of per-user accuracy on the 8-way emotion identification task before applying score cutoffs.** The histogram shows the number of correct responses out of 32 samples vs. the number of raters. Vertical dashed lines mark the inclusion thresholds at 8 and 28 correct labels that was used to discard near-random and likely automated submissions.

Appendix B: Additional Project - Data Augmentation for Prosody Analysis and Recognition

In addition to the main research project of this thesis, we carried out a small-scale project - an initial exploration of prosody analysis using DL models . We provide below the full report of the methods and the results of that exploration.

B.1 Abstract

Prosody analysis is a branch of linguistics that studies elements such as intonation and stress. A major challenge for the automated analysis of prosody using DNNs is the limited amount of labeled data.We thus explore two novel data augmentation methods that preserve the integrity of prosodic elements. The first method applies Room Impulse Response (RIR) to the recorded speech, and the second adds background noise. The study aims to increase the amount of labeled prosodic data with prosody-respecting augmentation and thus improve the performance of automated prosody classification. The Santa Barbara Corpus of Spoken American English (SBC) provides data labeled for prosodic boundaries and boundary type, while RIR and background noise augmentations are applied. The evaluation employs the Whisper automatic speech recognition (ASR) system, specifically its small architecture, for efficiency. Results reveal slight improvements on boundary detection (segmentation kappa) values with noise or combined noise-RIR augmentations compared to unaugmented baseline models. The combined augmentation yields the most promising enhancement, although gains are modest and must be balanced against increased training time.

B.2 Introduction

Prosody analysis, a subdomain within the broader field of linguistics, focuses on the study of elements such as intonation, stress, and rhythm that transcend individual phonetic segments (i.e., vowels and consonants). Despite its importance, the complexity of prosodic elements demands professional linguistic labeling. The lack of labeled data poses challenges to training artificial neural networks (ANNs) for this task.

A promising solution to the problem of data scarcity is data augmentation. However, traditional audio augmentation techniques, such as pitch alteration or cutting, do not lend themselves to prosody analysis due to the potential loss of essential prosodic information. Therefore, alternative augmentation techniques that respect the integrity of prosodic elements are required.

This project investigates two such methods: the application of Room Impulse Response (RIR), which to date has not been extensively used in machine learning audio tasks, and the addition of background noise. Both techniques offer the potential for enhancing the quantity of the limited amount of labeled data without disrupting the inherent prosodic features.

To quantify the benefits derived from these augmentations, a controlled experiment was conducted. In this experiment, a network was trained on the same data set several times: once with the augmented data and once without. By comparing the performance in these two scenarios, we established the effectiveness of data augmentation in improving the results for prosody analysis.

B.3 Goals

- Increase the amount of labeled prosodic data, using data augmentation techniques that preserve prosody.
- Improve the performance of prosody classification networks using augmented data.

B.4 Methods

B.4.1 Datasets

The Santa Barbara Corpus (SBC)
The Santa Barbara Corpus of Spoken American English (SBC) [56] is a voice dataset published by the linguistics department at the University of California, Santa Barbara (UCSB). The corpus consists of a set of 60 audio files that record the spontaneous speech of various genres, from multi-party kitchen conversations and couples' dialogues to child tutoring, guided tours, sermons, and university classes. The dataset contains basic prosodic labeling for boundaries and prototypes. Additional intonation units (IU) labeling has been added by the prosody team in Prof. David Harel's lab [57].

Room impulse response (RIR)
To perform a prosody-preserving augmentation, an RIR is applied to the speech. RIR adds the properties of a room, without changing the prosody. Various RIRs were collected for natural reverberation analysis [58].

Environmental sound classification (ESC)
Another prosody-preserving augmentation is background noise. Natural background noises can be found in the ESC-50 dataset [59]. The dataset is a labeled collection of 2,000 environmental audio recordings originally collected for benchmarking methods of environmental sound classification. The dataset consists of 5-second-long recordings organized into 50 semantic classes (40 examples per class), loosely arranged into 5 major categories: animals; natural soundscapes and water sounds; human, non-speech sounds; interior/domestic sounds; and exterior/urban noises.

B.4.2 Models

Web-scale Supervised Pretraining for Speech Recognition (Whisper)
OpenAI has developed Whisper, an automatic speech recognition (ASR) system that utilizes a vast dataset of 680,000 hours of multilingual and multitask supervised data sourced from the internet [27]. This extensive and varied dataset significantly enhances the system's resilience to challenges like accents, background noise, and technical language. Additionally, it empowers the system to transcribe various languages and translate them into English. OpenAI is making both the models and inference code open source, providing a valuable resource for creating practical applications and advancing research in robust speech processing. Comparisons with human performance reveal similar accuracy and robustness. The model I used for the augmentation project is a fine-tuned version of Whisper developed at our lab [57].

B.5 Results

Whisper has several architecture types: tiny, small, medium, and large-v2. The architecture types indicate the number of parameters. For this experiment, I used the small Whisper model architecture (244 million parameters). The small size makes it faster to train and evaluate the augmentation's effect.

In **Fig. B1** below are the results of Kappa segmentation values for different augmentation types and sizes. The exact numbers are shown in Table B1. Kappa segmentation is a measurement of Cohen's kappa on the IU segmentation.

I used pure noise augmentation, pure RIR augmentation, and a combined augmentation (both noise and RIR). Each augmented sample was created by random choice of the noise/RIR type out of the noise/RIR dataset. Augmentations were applied only to the training data, whereas the test data is the same for all models.

Inserting noise or a combination of noise and RIR shows a slight improvement compared to the baseline model. Inserting RIR did not show an improvement. Adding more augmentations,

thus increasing the amount of data, showed some improvement in all models.

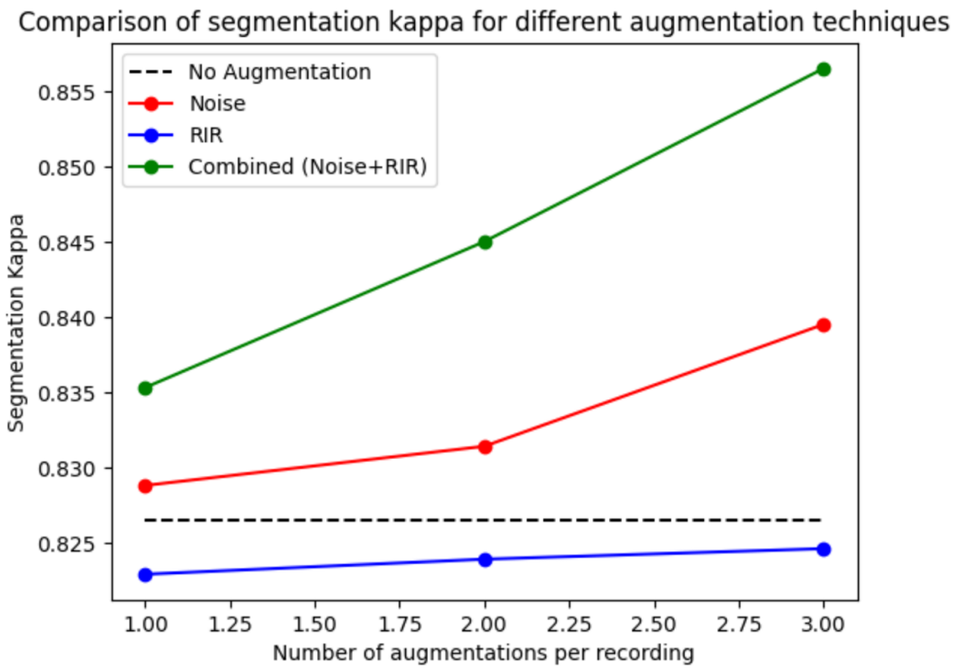

Figure B1: **Comparison of segmentation kappa for different augmentation techniques**
The x-axis is indicative of the change in the size of the data. No augmentation is the baseline model. Lines represent an addition of data with noise, RIR, or both.

	Number of augmentations per recording			
	0	1	2	3
No augmentation	82.65%			
Noise		82.88%	83.14%	83.95%
RIR		82.29%	82.39%	82.46%
Combined (noise + RIR)		83.53%	84.50%	85.65%

Table B1: Segmentation Kappa scores by number of augmentations per recording and augmentation technique

B.6 Discussion

The study's exploration of RIR and environmental noise as data augmentation methods for prosody analysis yielded only modest improvements in segmentation kappa values, suggesting these techniques offer limited enhancement to model performance.

The augmentation with noise, and particularly the combined noise-RIR approach, seems to provide minor performance gains (improvement of 3% from baseline), indicating that the added variability helps the model generalize better. This improvement might stem from the model learning to focus on key prosodic features amid diverse acoustic conditions, rather than overfitting to clean data.

Despite these slight benefits, the improvements might not justify the significant increase in computational resources and training time required for augmented data. The RIR-only augmentations did not demonstrate a clear positive impact on model performance, suggesting the need for further research to understand their role in prosody recognition.

Considering the complexity of prosodic features and the subtle improvements observed, future research should consider exploring more sophisticated and nuanced augmentation techniques, possibly in conjunction with expert linguistic knowledge, to craft methods that could yield more significant improvements without incurring prohibitive training costs.

Acknowledgments

I am sincerely grateful to my advisor, Professor David Harel, whose guidance and encouragement shaped this research journey. Special thanks to Dr. Tirza Biron for her endless support and practical advice. I appreciate Moshe Barboy, Alona Golubchik, Smadar Szekely, Yaron Winter, and Eran Ben-Artzy for their valuable insights and for fostering a collaborative research environment.

References

[1] A. Triantafyllopoulos, B. W. Schuller, G. İymen, M. Sezgin, X. He, Z. Yang, P. Tzirakis, S. Liu, S. Mertes, E. André *et al.*, "An overview of affective speech synthesis and conversion in the deep learning era," *Proceedings of the IEEE*, 2023.

[2] H. Barakat, O. Turk, and C. Demiroglu, "Deep learning-based expressive speech synthesis: a systematic review of approaches, challenges, and resources," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2024, no. 1, p. 11, 2024.

[3] H. Choi, S. Park, J. Park, and M. Hahn, "Multi-speaker emotional acoustic modeling for cnn-based speech synthesis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6950–6954.

[4] Y. Lee, A. Rabiee, and S.-Y. Lee, "Emotional end-to-end neural speech synthesizer," *arXiv preprint arXiv:1711.05447*, 2017.

[5] Z. Luo, J. Chen, T. Takiguchi, and Y. Ariki, "Emotional voice conversion using dual supervised adversarial networks with continuous wavelet transform f0 features," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 10, pp. 1535–1548, 2019.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[7] Y. Xu, "Speech prosody: A methodological review," *Journal of Speech Sciences*, vol. 1, no. 1, pp. 85–115, 2011.

[8] R. Plutchik and H. Kellerman, *Theories of Emotion*. New York, NY, USA: Academic Press, 2013, vol. 1.

[9] P. Ekman, "Are there basic emotions?" *Psychological Review*, vol. 99, no. 3, pp. 550–553, 1992.

[10] P. Ekman, W. V. Friesen, and P. Ellsworth, *Emotion in the Human Face*, 2nd ed. Cambridge, UK: Cambridge University Press, 1982.

[11] A. Mehrabian and J. A. Russell, *An Approach to Environmental Psychology*. Cambridge, MA: MIT Press, 1974.

[12] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[13] Y. Wang, H. Zhan, L. Liu, R. Zeng, H. Guo, J. Zheng, Q. Zhang, X. Zhang, S. Zhang, and Z. Wu, "Maskgct: Zero-shot text-to-speech with masked generative codec transformer," *arXiv preprint arXiv:2409.00750*, 2024.

[14] Z. Du, Q. Chen, S. Zhang, K. Hu, H. Lu, Y. Yang, H. Hu, S. Zheng, Y. Gu, Z. Ma *et al.*, "Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens," *arXiv preprint arXiv:2407.05407*, 2024.

[15] Y. Guo, C. Du, X. Chen, and K. Yu, "Emodiff: Intensity controllable emotional text-to-speech with soft-label guidance," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[16] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, pp. 335–359, 2008.

[17] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE transactions on affective computing*, vol. 5, no. 4, pp. 377–390, 2014.

[18] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PloS one*, vol. 13, no. 5, p. e0196391, 2018.

[19] (2015) Surrey audio-visual expressed emotion (savee). [Online]. Available: http://kahlan.eps.surrey.ac.uk/savee/

[20] K. Zhou, B. Sisman, R. Liu, and H. Li, "Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 920–924.

[21] (2010) Toronto emotional speech set (tess). [Online]. Available: https://tspace.library.utoronto.ca/handle/1807/24487

[22] (2025) Gpt-4o-mini-tts model card. Accessed May 2025. [Online]. Available: https://platform.openai.com/docs/models/gpt-4o-mini-tts

[23] A. B. Abacha, W.-w. Yim, Y. Fu, Z. Sun, M. Yetisgen, F. Xia, and T. Lin, "Medec: A benchmark for medical error detection and correction in clinical notes," *arXiv preprint arXiv:2412.19260*, 2024.

[24] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, 2017.

[25] L. Goncalves, A. N. Salman, A. R. Naini, L. Moro Velazquez, T. Thebaud, L. P. Garcia, N. Dehak, B. Sisman, and C. Busso, "Odyssey 2024 speech emotion recognition challenge: Dataset, baseline framework, and results," in *The Speaker and Language Recognition Workshop (Odyssey 2024)*, Quebec City, Canada, 2024.

[26] WhisperSpeech Contributors, "Whisperspeech: An open source text-to-speech system built by inverting whisper," https://github.com/WhisperSpeech/WhisperSpeech, 2025, gitHub repository; accessed 2025-08-14.

[27] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*. PMLR, 2023, pp. 28 492–28 518.

[28] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Transactions on Communications*, vol. 28, no. 1, pp. 84–95, 1980.

[29] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[30] E. Kharitonov *et al.*, "Spear-tts: Stable and efficient TTS with discrete semantics and acoustic tokens," in *NeurIPS*, 2023.

[31] Z. Borsos *et al.*, "Audiolm: a language modeling approach to audio generation," in *ICLR*, 2023.

[32] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," in *NeurIPS*, 2022, enCodec.

[33] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn-based speaker verification," in *Interspeech*, 2020.

[34] C. Inc., "Vocos: A neural vocoder for discrete audio tokens," https://github.com/charactr/vocos, 2023, gitHub repository.

[35] J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P.-E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, G. Synnaeve, and A. Joulin, "Libri-light: A benchmark for ASR with limited or no supervision," in *ICASSP*, 2020.

[36] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[37] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT*, 2019.

[38] N. Zeghidour, F. de Chaumont Quitry, O. Teboul, M. Tagliasacchi, A. Luebs, W. Han, J. Skoglund, and R. J. Weiss, "Soundstream: An end-to-end neural audio codec," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, 2021, pp. 300–311.

[39] Multimodal Signal Processing Lab, "Msp-podcast corpus, release 1.12," Online, 2025, access details provided by the authors. Release 1.12 contains approximately 320 hours of emotional speech. [Online]. Available: https://lab-msp.com/MSP-PODCAST-Publish-1.12/

[40] J. Shen, R. Pang *et al.*, "Natural tts synthesis by conditioning WaveNet on mel spectrogram predictions," https://google.github.io/tacotron/publications/tacotron2/, 2017, uses MOS listening tests for naturalness.

[41] "Subjective evaluation of speech quality with a crowdsourcing approach," International Telecommunication Union, ITU-T Rec. P.808, Geneva, Switzerland, Tech. Rep., 2021, crowdsourcing methodology for MOS.

[42] "Methods for subjective determination of transmission quality," International Telecommunication Union, ITU-T Rec. P.800, Geneva, Switzerland, Tech. Rep., 1996, telephony speech quality recommendation.

[43] G. Yang, C. Yang, Q. Chen, Z. Ma, W. Chen, W. Wang, T. Wang, Y. Yang, Z. Niu, W. Liu *et al.*, "Emovoice: Llm-based emotional text-to-speech model with freestyle text prompting," *arXiv preprint arXiv:2504.12867*, 2025.

[44] K. Huang, Q. Tu, L. Fan, C. Yang, D. Zhang, S. Li, Z. Fei, Q. Cheng, and X. Qiu, "Instructttseval: Benchmarking complex natural-language instruction following in text-to-speech systems," *arXiv preprint arXiv:2506.16381*, 2025.

[45] R. R. Manku, Y. Tang, X. Shi, M. Li, and A. Smola, "Emergenttts-eval: Evaluating tts models on complex prosodic, expressiveness, and linguistic challenges using model-as-a-judge," *arXiv preprint arXiv:2505.23009*, 2025.

[46] R. Likert, "A technique for the measurement of attitudes," *Archives of Psychology*, vol. 22, no. 140, pp. 1–55, 1932.

[47] G. Paolacci, J. Chandler, and P. G. Ipeirotis, "Running experiments on amazon mechanical turk," *Judgment and Decision Making*, vol. 5, no. 5, pp. 411–419, 2010.

[48] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society. Series B*, vol. 57, no. 1, pp. 289–300, 1995.

[49] E. B. Wilson, "Probable inference, the law of succession, and statistical inference," *Journal of the American Statistical Association*, vol. 22, no. 158, pp. 209–212, 1927.

[50] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, 1947.

[51] B. Naderi, R. Cutler, S. Braun, and I. Tashev, "An open source implementation of itu-t recommendation p.808 with validation," in *Interspeech*, 2020.

[52] R. Liu, B. Sisman, and H. Li, "Reinforcement learning for emotional text-to-speech synthesis with improved emotion discriminability," in *Proc. Interspeech*, 2021, uses an SER trained on ESD and reports SER-based accuracy on synthesized audio.

[53] J. Wang *et al.*, "Ed-tts: Multi-scale emotion modeling using cross-domain emotion diarization for emotional speech synthesis," *arXiv preprint arXiv:2401.08166*, 2024, reports Emotion Reclassification Accuracy (ERA) using a pretrained SED/SER as the evaluator.

[54] A. Agresti, *Categorical Data Analysis*, 3rd ed. Hoboken, NJ: Wiley, 2013.

[55] G. Casella and R. L. Berger, *Statistical Inference*, 2nd ed. Pacific Grove, CA: Duxbury, 2002.

[56] J. W. Du Bois, W. L. Chafe, C. Meyer, S. A. Thompson, and N. Martey, "Santa barbara corpus of spoken american english," *CD-ROM. Philadelphia: Linguistic Data Consortium*, 2000.

[57] T. Biron, M. Barboy, E. Ben-Artzy, A. Golubchik, Y. Marmor, A. Marron, S. Szekely, Y. Winter, and D. Harel, "Disentanglement of prosodic meaning: Toward a framework for the analysis of nonverbal information in speech," *Proceedings of the National Academy of Sciences*, vol. 122, no. 37, p. e2500510122, 2025.

[58] J. Traer and J. H. McDermott, "Statistics of natural reverberation enable perceptual separation of sound and space," *Proceedings of the National Academy of Sciences*, vol. 113, no. 48, pp. E7856–E7865, 2016.

[59] K. J. Piczak, "Esc: Dataset for environmental sound classification," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 1015–1018.