

Qijun Xie^a^aThe Chinese University of Hong Kong, Hong Kong, China

ARTICLE INFO

Keywords:

Text-to-Speech
Direct Preference Optimization
Diffusion Model
Elderly-Facing Supervised Learning

ABSTRACT

Text-to-Speech systems play a crucial role in improving communication accessibility for individuals with hearing impairments, particularly among elderly Cantonese speakers. However, two major challenges persist: (1) limited availability of high-quality Cantonese resources in comparison to Mandarin or English, and (2) insufficient personalization for elderly users, whose auditory preferences and cultural nuances often differ from the broader population. To address these challenges, we propose *SPEECH*, a specialized TTS framework that integrates a latent diffusion architecture to operate effectively with relatively small datasets and employs Direct Preference Optimization to incorporate direct user feedback from elderly speakers. Our framework features two core components: Elderly-Centric Acoustic Supervisor and Direct Preference Optimization. ECAS focuses on collecting and refining a curated, high-quality audio-text pair dataset tailored for the elderly Cantonese demographic, leveraging noise reduction, precise alignment, and diverse augmentation strategies to capture the linguistic richness of Cantonese. And, DPO aligns the synthesized speech with the tonal balance, pacing, and clarity preferences articulated by elderly users. By uniting a robust data processing pipeline with feedback-driven optimization, *SPEECH* delivers culturally relevant, personalized, and intelligible speech that addresses the unique communication needs of the elderly Cantonese-speaking community. Our dataset and implementation details are available at: <https://github.com/Viicte/SPEECH-Text-to-Speech-for-Cantonese-Elderly>

1. Introduction

According to a 2019/20 survey by the Census and Statistics Department of Hong Kong, approximately 39,500 individuals aged 65 or above experience hearing difficulties, accounting for 82.4% of all persons with hearing impairment [1, 2, 3]. This widespread prevalence highlights the urgent need for tools that improve communication for the elderly population. Customized audio technologies, such as a specialized Cantonese Text-to-Speech (TTS) system, can help address these challenges by producing clearer and more culturally relevant speech outputs suited to the auditory needs and preferences of older listeners.

Standard text-to-speech systems, while increasingly natural-sounding, often neglect the perceptual challenges faced by older adults, such as reduced sensitivity in high-frequency bands and increased difficulty in parsing fast or noisy speech. Recent advances in latent diffusion models [4, 5] have dramatically improved TTS naturalness and flexibility. However, these systems are not specifically optimized for elderly comprehension and listener preference.

To address this gap, we propose the *SPEECH*, a diffusion-based TTS framework designed explicitly for elderly Cantonese listeners. As shown in Figure 1, existing methods typically follow a sequence-to-sequence (Seq2Seq) pipeline that predicts mel-spectrograms and feeds them into a neural vocoder. In contrast, our system operates in the latent space via a latent diffusion model and incorporates two novel components: 1) Elderly-Centric Acoustic Supervising (ECAS) is introduced to improve the intelligibility of synthesized speech for older listeners. ECAS functions as a specialized loss module that penalizes frequency mismatches in the


1–4 kHz range, where elderly hearing is often impaired, and detects intelligibility issues by comparing ASR transcripts of generated speech to ground truth text. This directs the model to emphasize acoustically critical regions and clarity-enhancing prosodic patterns during training. 2) Direct Preference Optimization (DPO) is employed to align synthesized speech with real-world user preferences. This method fine-tunes the model using human-labeled comparisons of generated audio samples, where listeners, including elderly users, which indicate which version is clearer or more pleasant. The model is then optimized using a preference-based objective to increase the likelihood of generating favored outputs.

We evaluate our approach on both a curated Elderly Cantonese Speech dataset and the public CommonVoice-Yue dataset. Experiments show that *SPEECH* significantly improves both objective metrics (FAD, IS, KL) and subjective scores (OVL, REL) over conventional systems. Ablation results further validate the independent and complementary roles of ECAS and DPO in enhancing output quality and listener satisfaction.

The contributions of this paper are as follows:

- We propose *SPEECH*, an Elderly centric TTS framework, which leveraging a latent diffusion model optimized specifically for elderly Cantonese listeners.
- To capture nuanced acoustic characteristics essential for elderly comprehension in Cantonese speech, we introduce the ECAS mechanism, which incorporates a specialized loss term guided by transcription errors and frequency-weighted feedback.
- To align the synthesized outputs with elderly auditory preferences regarding clarity, pace, and pleasantness, we employ Direct Preference Optimization, fine-tuning the model based on pairwise human feedback.

*Corresponding author

 viicte@outlook.com (Q. Xie)

ORCID(s): 0000-0000-0000-0000 (Q. Xie)

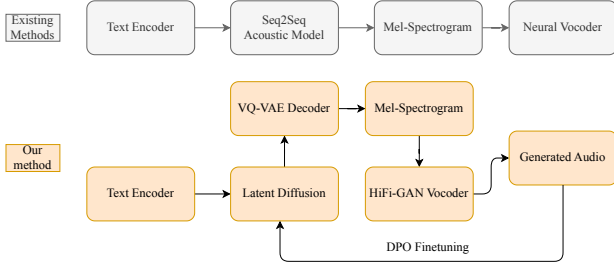


Figure 1: Comparison between existing TTS methods and our proposed elderly-centric *SPEECH* system. Traditional models employ a Seq2Seq acoustic predictor followed by a neural vocoder. Our system replaces this pipeline with a latent diffusion framework, which generates mel-spectrogram latents via a text-conditioned diffusion model, decoded by a VQ-VAE and rendered through HiFi-GAN. The addition of ECAS (training-time constraint) and DPO (preference-driven finetuning) further aligns the system with elderly users' clarity and preference needs.

Our dataset and implementation details are available at: <https://github.com/Viicte/SPEECH-Text-to-Speech-for-Cantonese-Elderly>

2. Related Work

2.1. Neural Text-to-Speech Synthesis

Neural Text-to-Speech (TTS) systems have seen remarkable progress through end-to-end architectures like Tacotron [6] and Tacotron2 [7]. Tacotron employs a sequence-to-sequence model with attention mechanisms to translate an input text sequence \mathbf{y} into a mel-spectrogram \mathbf{M} :

$$\mathbf{M} = f_{\text{Tacotron}}(\mathbf{y}), \quad (1)$$

where f_{Tacotron} denotes the Tacotron model. Tacotron2 improves upon this by using location-sensitive attention, leading to better alignment between text and speech. The attention weights $\alpha_{i,j}$, which align decoder timestep i with encoder timestep j , are computed as:

$$\alpha_{i,j} = \frac{\exp(e_{i,j})}{\sum_k \exp(e_{i,k})}, \quad e_{i,j} = \mathbf{v}^\top \tanh(\mathbf{W}s_{i-1} + \mathbf{V}h_j + \mathbf{U}\mathbf{F}_{i,j}), \quad (2)$$

where s_{i-1} is the decoder state at time $i-1$, h_j is the encoder output at position j , and $\mathbf{F}_{i,j}$ encodes location-based features based on the attention weights.

Neural vocoders such as WaveNet [8] generate raw waveforms \mathbf{x} from mel-spectrograms \mathbf{M} by modeling the waveform as a conditional autoregressive distribution:

$$p(\mathbf{x}) = \prod_{t=1}^T p(x_t | x_1, \dots, x_{t-1}, \mathbf{M}), \quad (3)$$

where x_t denotes the audio sample at time t and T is the total number of samples.

WaveGlow [9], a flow-based vocoder, generates audio by learning an invertible transformation g that maps audio \mathbf{x} to a latent variable \mathbf{z} and optimizes the exact log-likelihood via:

$$p(\mathbf{x}) = p(\mathbf{z}) \left| \det \left(\frac{\partial g^{-1}(\mathbf{x})}{\partial \mathbf{z}} \right) \right|, \quad (4)$$

where g^{-1} is the inverse mapping defined by WaveGlow and $p(\mathbf{z})$ is a Gaussian prior over the latent space.

GAN-based models such as MelGAN [10] generate audio using adversarial training. The generator G aims to produce realistic waveforms from mel-spectrograms \mathbf{M} , while the discriminator D attempts to distinguish real audio \mathbf{x} from generated audio $G(\mathbf{M})$. The optimization follows a min-max adversarial loss:

$$\min_G \max_D \mathbb{E}[\log D(\mathbf{x})] + \mathbb{E}[\log(1 - D(G(\mathbf{M})))] \quad (5)$$

2.2. Diffusion Models for Speech

Diffusion probabilistic models have recently emerged as powerful generative models for high-quality audio synthesis, with AudioLDM [4] and NaturalSpeech 2 [5] employing latent diffusion approaches. The core idea involves a diffusion process that progressively injects Gaussian noise into the data over a series of steps, forming intermediate noisy distributions:

$$q(\mathbf{x}_t | \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1 - \beta_t} \mathbf{x}_{t-1}, \beta_t \mathbf{I}), \quad (6)$$

where \mathbf{x}_t is the noisy version of the data at timestep t , β_t is a pre-defined noise schedule controlling the variance of the added noise, and \mathbf{I} is the identity matrix.

The generative process involves reversing the diffusion via a learned denoising model:

$$p_\theta(\mathbf{x}_t - 1 | \mathbf{x}_t, \mathbf{c}) = \mathcal{N}(\mathbf{x}_t - 1; \mu_\theta(\mathbf{x}_t, t, \mathbf{c}), \Sigma_\theta(\mathbf{x}_t, t, \mathbf{c})), \quad (7)$$

where μ_θ and Σ_θ are the predicted mean and variance parameterized by neural networks with parameters θ , conditioned on the noisy input \mathbf{x}_t , the timestep t , and the conditioning information \mathbf{c} . The conditioning variable \mathbf{c} typically encodes external semantic information, such as text embeddings derived from models like CLAP [11].

2.3. Speech for Elderly Listeners

Prior research primarily focuses on automatic speech recognition and hearing aid design aimed at enhancing speech intelligibility for older adults. This is often achieved by adjusting acoustic parameters such as speech rate, frequency emphasis, and clarity [12]. One such enhancement technique is spectral shaping:

$$Y(f) = X(f) \times H(f), \quad (8)$$

where $X(f)$ represents the original speech spectrum, and $H(f)$ is a frequency-dependent amplification function. $H(f)$ is designed based on audiological profiles typical of elderly listeners, often emphasizing frequencies between 1-4 kHz to improve comprehension.

However, there is limited research in TTS tailored for elderly listeners, especially in tonal languages like Cantonese. This gap motivates our ECAS pipeline, which integrates auditory preferences of older adults directly into the TTS training process, aiming for improved accessibility and listening comfort by the proper constraints.

2.4. Preference-Based Model Fine-tuning

Recent advances in natural language processing have shown the effectiveness of leveraging human feedback for refining generative models. Direct Preference Optimization [13] introduces a framework that incorporates comparative human preferences directly into the model training process, bypassing the need for explicit reward modeling.

Given two generated outputs y_1 and y_2 , human annotators indicate a preference, and the model estimates the preference probability as:

$$p(y_1 > y_2) = \frac{\exp(s_\theta(y_1))}{\exp(s_\theta(y_1)) + \exp(s_\theta(y_2))}, \quad (9)$$

where $s_\theta(y)$ is a scalar-valued scoring function predicted by the model with parameters θ for a given output y .

The model parameters θ are then optimized by maximizing the log-likelihood of the preferred outputs using the following objective:

$$\mathcal{LDPO} = - \sum_{(y_1, y_2) \in \mathcal{D}} \log \frac{\exp(s_\theta(y_{\text{preferred}}))}{\exp(s_\theta(y_1)) + \exp(s_\theta(y_2))}, \quad (10)$$

where \mathcal{D} is the dataset of paired outputs annotated with preference labels. Each pair (y_1, y_2) consists of two candidate outputs generated by the model for the same prompt.

Our work is among the first to apply DPO to speech synthesis, directly optimizing TTS model outputs based on elderly listener preferences. This approach enhances subjective satisfaction, and lowers computational cost.

3. Methodology

In this section, we describe the architecture of our TTS system and its key components. Our methodology includes three parts: (a) data collection and processing, (b) integrating the Elderly-Centric Acoustic Supervising mechanism into model training, and (c) a Direct Preference Optimization fine-tuning stage as shown in Figure 1.

3.1. Overview of SPEECH

SPEECH is built on a latent diffusion model that generates speech audio from text input. The system takes a text

prompt as input and processes it through a text encoder to obtain a text embedding, which serves as a conditioning signal for the latent diffusion model. During training, the corresponding ground-truth waveform is transformed into a compact latent representation via a variational autoencoder. The latent diffusion model, denoted D_θ , is trained to generate this latent audio representation given the text embedding.

As shown in Figure 2, y is the input text, e.g., a sequence of characters and x be the corresponding ground-truth audio waveform. We first convert x into a time-frequency representation using a Short-Time Fourier Transform (STFT) with a Mel filter bank, yielding a mel-spectrogram M . A pre-trained encoder E from a VAE then compresses M into a latent code $z = E(M)$. The diffusion model takes as input the conditioning text embedding $c = f_{\text{text}}(y)$ which is obtained from a transformer-based text encoder f_{text} and iteratively denoises a random latent toward z .

After training, at inference time the model can generate a new latent z' from an arbitrary text prompt y' by running the diffusion process conditioned on $c = f_{\text{text}}(y')$. The decoder portion of the VAE, D , converts the generated latent z' into a mel-spectrogram \hat{M} , which is then transformed into a waveform \hat{x} using a vocoder.

We incorporate additional modules to enhance audio quality and clarity. As shown in Fig. 2, a post-processing chain including a denoising model [14] for noise reduction, as well as equalization and dynamic range compression, is applied to the output waveform \hat{x} to further improve clarity for elderly listeners. These modules are applied only during inference to refine the synthesized speech.

Components of the system with trainable parameters such as the latent diffusion model are used during both training and inference, while certain pre-trained or fixed modules, such as, the post-processing filters are used in inference only.

3.2. Elderly-Centric Acoustic Supervising

A key innovation of our approach is the ECAS mechanism, which guides the model training with objectives designed to improve the clarity of output speech for elderly listeners. ECAS is implemented as an additional loss term and training strategy integrated into the diffusion model's learning process.

We identified several acoustic factors important for elderly comprehension: clear pronunciation of consonants (which often have energy in higher frequencies), appropriate speaking rate (not too fast), and low background noise. During training, we introduce a specialized loss L_{ECAS} that penalizes the model if the synthesized speech is likely to violate these criteria. For example, we utilize a pre-trained speech recognition model to transcribe the model's output during training; if the transcription has errors compared to the input text (indicating poor intelligibility), it contributes to L_{ECAS} . Similarly, we apply a frequency-weighted error measure that gives more weight to deviations in the 1-4 kHz range of the output mel-spectrogram (since insufficient energy in this band can make speech harder to understand for

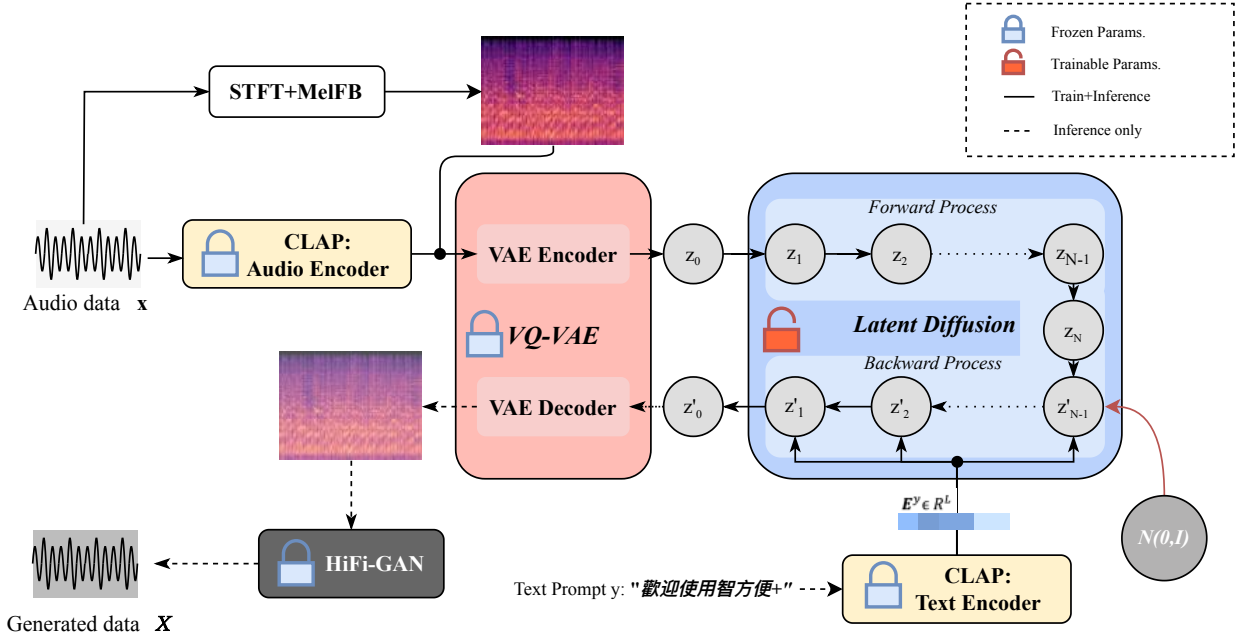


Figure 2: Overview of *SPEECH*. The input text prompt is encoded into a text embedding via a frozen CLAP Text Encoder. A latent diffusion model generates mel-spectrogram latent representations conditioned on this embedding. These are decoded by a VQ-VAE into full spectrograms and subsequently converted into waveforms using a frozen HiFi-GAN vocoder. The system operates in the mel-spectrogram latent space, leveraging both audio and text encoders from CLAP. Orange locks indicate trainable components, while blue locks and dashed lines indicate frozen parameters and inference-only modules.

the elderly). This encourages the model to produce adequate energy in that important frequency range.

The ECAS loss is combined with the standard diffusion training loss. Let L_{diff} denote the normal denoising score-matching loss for the diffusion model. For example, if ϵ_θ is the diffusion model's predicted noise and ϵ is the actual noise added at step t , we can define:

$$L_{\text{diff}} = \mathbb{E}_{z_0, y, \epsilon, t} [\|\epsilon - \epsilon_\theta(z_t, t, c)\|_2^2]. \quad (11)$$

where $z_0 = E(M)$ is the ground-truth latent, z_t is the latent at diffusion step t obtained by adding noise ϵ to z_0 , and $c = f_{\text{text}}(y)$ is the text conditioning embedding. Our total training objective becomes:

$$L_{\text{total}} = L_{\text{diff}} + \lambda_{\text{ECAS}} L_{\text{ECAS}}, \quad (12)$$

where λ_{ECAS} is a weight that controls the influence of the ECAS term. In practice, we set a relatively small λ_{ECAS} (e.g., 0.1) at the beginning of training to avoid destabilizing the diffusion model optimization, and gradually increase it once the base model produces reasonable speech. This scheduling allows the model to first learn general speech synthesis and then increasingly focus on the elderly-specific clarity aspects.

By incorporating ECAS, the diffusion model not only learns to match the training data distribution but also gains a bias toward producing clearer, more intelligible speech. This is effectively injecting domain-specific knowledge (the needs of elderly listeners) into the model training, which we find leads to outputs that yield fewer errors in transcription and better subjective clarity.

3.3. Direct Preference Optimization

Direct Preference Optimization fine-tunes the model using actual human feedback on the generated speech. After training the model with ECAS, we conduct a DPO stage to further align the speech output with listener preferences.

To collect preference data, we generated multiple candidate speech samples for a set of validation text prompts (using different random seeds or variations of our model). These samples were presented in pairs to human evaluators (including both typical listeners and older adults), who were asked which sample in each pair is clearer and more pleasant. From these tests, we gathered a set of comparisons with preference labels. We then applied the DPO algorithm [13] to our model using this data.

Concretely, for a given pair of outputs (\hat{x}_i, \hat{x}_j) for the same input text where the evaluator preferred \hat{x}_i over \hat{x}_j , we define a preference loss:

$$L_{\text{DPO}} = -\log \frac{\exp(s(\hat{x}_i))}{\exp(s(\hat{x}_i)) + \exp(s(\hat{x}_j))}, \quad (13)$$

where $s(\hat{x})$ is a scoring function representing the model's inclination to produce output \hat{x} . In practice, $s(\hat{x})$ can be derived from the TTS model's internal likelihood or a small auxiliary network trained to predict preferences. We then fine-tune the TTS model to minimize L_{DPO} across all collected comparisons (with a regularization term to prevent the model output from drifting too far from the original distribution, as recommended in [13]). This process directly increases the probability of the model generating the kind of outputs that listeners preferred during the comparison tests.

4. Experiment

4.1. Dataset Overview and Statistics

The raw data was collected by crawling the Information Technology and Innovation Section of the Hong Kong Government website¹, specifically targeting the Online Course section. A total of 97 Cantonese-language educational videos, accompanied by PDF scripts, were retrieved; these videos cover topics such as digital literacy, online safety, and technology use for elderly audiences.

Audio Preprocessing

The preprocessing pipeline is implemented in Python to optimize audio quality for elder-friendly listening. First, each input WAV file is loaded with `pydub` and processed to remove DC offset; If the recording is stereo, it is converted to mono. The audio is then filtered using `SoX`, applying a high-pass filter at 80 Hz and a low-pass filter at 8000 Hz to remove extraneous low, and high-frequency components. When available, a `SoX` noise profile is used to reduce background noise. Next, a light dynamic range compression is applied via `SoX`'s `compand` function, and the audio is finally normalized to a target loudness level (typically -12 dBFS) using `pydub`. The processed output is saved with a `_cleaned.wav` suffix in the designated output directory.

Transcription

Cleaned audio files are transcribed using OpenAI's Whisper model. The model is loaded on a GPU (device set to `cuda` when available) and configured for traditional Chinese (language code 'zh'). For each file, the transcription process generates a full caption and detailed segment-level timestamps. The complete transcriptions are stored in a CSV file (`captions.csv`), while a separate CSV file (`timestamps.csv`) records the segment indices along with the start and end times and the corresponding text. This structured output supports subsequent forced alignment and analysis tasks.

Audio files are further categorized by voice type (human or non-human), speaker composition (male, female, both, or unknown), and the presence of background music. Table 1 summarizes the key statistics of the dataset, including the number of files, total duration, average duration, and the minimum and maximum durations. All audio files are sampled uniformly at 44,100 Hz.

The category naming convention in Table 1 is as follows: Y or N at the start indicates human or non-human voice, the next letter denotes speaker gender composition (M for male, F for female, B for both, and N for unknown), and the letter after the hyphen indicates presence (B) or absence (N) of background music. For example, "YF-N" corresponds to a human female voice with no background music.

Data Augmentation

To enhance the diversity of the dataset and improve model robustness, three augmentation strategies are applied:

1. **Punctuation-based Segmentation:** Utilizes word-level timestamps from raw Whisper outputs to delineate segments at punctuation boundaries. Speaker

labels are extracted directly from filenames (e.g., parsing `speaker3` from `speaker3:...`). In cases lacking punctuation, at least one segment is generated.

2. **Silence-based Segmentation:** Employs `pydub`'s silence detection functionality to identify nonsilent regions. Each continuous speech segment is annotated with the corresponding speaker label and the full transcript.
3. **Mix-up Strategy:** Randomly combines segments from the same speaker by linearly mixing pairs of audio samples. Specifically, the augmented audio segment $x_{1,2}$ is generated by the following operation:

$$x_{1,2} = \lambda x_1 + (1 - \lambda)x_2, \quad (14)$$

where λ is a scaling factor sampled from a Beta distribution $B(5, 5)$. Captions from mixed audio segments are concatenated to form new paired samples. This augmentation strategy increases training diversity and helps the model generalize better to varied audio embeddings, thereby enhancing robustness and performance.

Additional Dataset (CommonVoice-Yue)

The Common Voice project, initiated by Mozilla, is a crowd-sourced effort to create a free, publicly available repository of voice recordings for speech recognition development. Volunteers contribute by recording specified sentences and validating others' recordings, resulting in a multilingual corpus with diverse languages and accents. As of the latest release, Common Voice 20 in December 2024, the dataset includes 33,150 hours of speech data across 133 languages, with 22,108 hours validated for quality assurance.

The Cantonese subset, known as CommonVoice-Yue, provides valuable resources for training and evaluating models for Cantonese speech recognition. While specific details about the number of hours, sentences, speakers, and sample rate for the Cantonese subset are not readily available, the dataset is continuously updated with new contributions. Researchers can access the latest data from the Common Voice website or platforms like Hugging Face.

4.2. Baselines

AudioLDM [4]: A latent diffusion model specialized in audio synthesis tasks. AudioLDM utilizes cross-modal alignment between audio and text embeddings in a latent space, generating high-quality audio conditioned on textual prompts through iterative denoising steps.

Tacotron2 [7]: A sequence-to-sequence neural network architecture designed for text-to-speech synthesis. Tacotron2 employs an attention-based encoder-decoder mechanism and outputs mel-spectrograms, subsequently converted into audio waveforms through neural vocoders.

FastSpeech [15]: A non-autoregressive TTS model that significantly improves inference speed and stability over traditional autoregressive models like Tacotron. It utilizes duration prediction to align text and mel-spectrograms.

¹<https://www.it2.gov.hk/tc/index.php#ps1>

Table 1
Dataset Statistics for Audio Files

Category	Count	Total Duration (s)	Avg. Duration (s)	Min (s)	Max (s)	Sample Rate (Hz)
NB-B	4	1401.6	350.4	203.8	441.71	44100
NB-N	6	2489.41	414.9	174.99	901.1	44100
NF-B	4	1712.17	428.04	273.9	575.99	44100
NF-N	5	2190.36	438.07	209.17	841.12	44100
NM-B	4	1549.82	387.45	253.79	493.61	44100
NM-N	7	2298.57	328.37	230.46	551.87	44100
NN-B	1	315.23	315.23	315.23	315.23	44100
NN-N	3	954.83	318.28	251.57	421.07	44100
YB-B	5	1637.8	327.56	295.85	384.17	44100
YB-N	8	2256.28	282.04	147.12	478.91	44100
YF-B	6	2129.92	354.99	234.61	625.13	44100
YF-N	7	3145.68	449.38	215.99	745.15	44100
YM-B	5	1642.74	328.55	193.1	451.16	44100
YM-N	8	3139.46	392.43	241.93	620.62	44100
YN-B	6	1080.75	180.12	100.12	361.44	44100
YN-N	2	392.46	196.23	67.85	324.62	44100
Non-Human Total	34	12911.99	379.76	174.99	901.1	44100
Human Total	47	15425.09	328.19	67.85	745.15	44100
Overall Total	81	28337.08	349.84	67.85	901.1	44100

VITS [16]: A unified framework that integrates variational autoencoders, normalizing flows, and adversarial training to achieve high-quality, end-to-end speech synthesis with improved naturalness and prosody.

4.3. Evaluation Metrics

The evaluation of the proposed TTS system involves both objective and subjective metrics to comprehensively assess the quality and relevance of the generated audio.

Objective Metrics

Frechet Audio Distance (FAD):

FAD measures the similarity between the distribution of generated audio and real audio. It is computed as:

$$FAD = \|\mu_r - \mu_g\|_2^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2}), \quad (15)$$

where (μ_r, Σ_r) and (μ_g, Σ_g) are the mean and covariance of embeddings for real and generated audio, respectively.

Inception Score (IS):

IS evaluates both the quality and diversity of generated audio samples:

$$IS = \exp \left(\mathbb{E}_x [\text{KL}(p(y|x) \| p(y))] \right), \quad (16)$$

where $p(y|x)$ is the conditional probability distribution over labels given a generated sample, and $p(y)$ is the marginal distribution over labels.

Kullback-Leibler Divergence (KL):

KL divergence quantifies the divergence between real and generated data distributions at the sample level, providing a fine-grained comparison.

Subjective Metrics

For subjective evaluation, six audio professionals rated the generated samples on a scale of 1 to 100 based on the

following criteria:

Overall Quality (OVL):

This metric assesses the naturalness and clarity of the generated audio, with higher scores indicating higher quality.

Relevance to Text (REL):

This metric evaluates the consistency between the generated audio and the input text description, ensuring that the synthesized speech aligns with user expectations.

The combination of objective and subjective metrics provides a holistic evaluation framework, ensuring the generated audio is both technically accurate and contextually relevant.

4.4. Implementation Details

Datasets: We trained our model separately on two datasets to evaluate its performance:

1. **Self-Collected and Augmented Elderly Cantonese Speech Dataset:** This dataset comprises 12,020 audio clips, totaling approximately 26 GB. It was specifically curated to address the needs of elderly Cantonese listeners by incorporating clear pronunciation, reduced speaking rates, and enhanced audio clarity through the ECAS pipeline.
2. **Common Voice Cantonese Dataset (CommonVoice-Yue):** Sourced from Mozilla's Common Voice project, this dataset contains approximately 120,000 audio clips with a total size of about 3.8 GB. It provides a diverse range of Cantonese speech recordings contributed by volunteers.

Key parameters selected in our work include a sampling rate of 16 kHz, mel-spectrogram with 64 bins, and latent embedding dimensions set to 8. The latent diffusion model utilizes latent dimensions of 256 (time) and 16 (frequency).

Table 2

Objective Metrics on the Elderly Cantonese Speech Dataset and CommonVoice-Yue. Lower is better for FAD and KL; higher is better for IS.

Model	Elderly Cantonese			CommonVoice-Yue		
	FAD ↓	IS ↑	KL ↓	FAD ↓	IS ↑	KL ↓
FastSpeech	27.2	4.11	0.48	23.2	3.99	0.48
Tacotron2	25.8	4.00	0.49	22.1	3.85	0.47
AudioLDM	22.7	4.22	0.39	19.3	4.18	0.39
VITS	21.1	4.32	0.36	18.4	4.21	0.36
SPEECH	19.8	4.48	0.37	18.0	4.29	0.38

The UNet configuration employs 8 input channels and outputs latent embeddings with dimensions of 8, alongside model channels set at 128, and attention resolutions at 8, 4, and 2. The model training includes a warmup period of 2,000 steps, a batch size of 2, and a learning rate of 1.0×10^{-4} , using a linear diffusion noise schedule ranging from $\beta_1 = 0.0015$ to $\beta_N = 0.0195$. Training was conducted on a local NVIDIA RTX 3080 GPU and required approximately 50 hours to reach convergence. CLAP embeddings are used for conditioning. During evaluation, an unconditional guidance scale of 3.5 with 200 DDIM sampling steps is applied, generating multiple outputs per prompt for robust analysis.

4.5. Computation Complexity

The computational complexity of the proposed model arises primarily from the forward and reverse diffusion processes in the latent diffusion model and the attention mechanisms in the UNet backbone. The forward diffusion transforms the latent representation z_0 into noise through N timesteps, where each step operates with a complexity of $\mathcal{O}(N \cdot D^2)$, where D is the latent dimensionality. The reverse process, which iteratively reconstructs the original data, has the same order of complexity due to its iterative nature. The attention layers within the UNet encoder and decoder compute relationships between tokens, contributing a complexity of $\mathcal{O}(n^2 \cdot d)$, where n is the sequence length and d is the embedding dimension. The hierarchical structure of the UNet significantly enhances its ability to capture global dependencies in the latent representation. Combining the diffusion process and attention mechanisms, the total complexity of training is $\mathcal{O}(N \cdot (D^2 + n^2 \cdot d))$. This design ensures a balance between computational efficiency and model expressiveness, enabling high-quality audio synthesis with manageable hardware requirements.

5. Experimental Results

5.1. Objective Evaluation

As shown in Table 2, *SPEECH* consistently achieves the best performance in terms of Frechet Audio Distance (FAD) and Inception Score (IS) across both datasets, indicating that it produces more natural and diverse audio samples aligned with the target distribution. While VITS slightly outperforms *SPEECH* in KL divergence, suggesting marginally better alignment in localized statistical features, our model

outpaces it in the other two metrics. Notably, Tacotron2 performs the weakest in terms of FAD and IS, highlighting its limitations in handling prosodic variation and diversity for elderly-oriented speech. AudioLDM performs competitively in IS but suffers from higher KL divergence, implying that while it generates diverse outputs, the alignment to real data distributions is less precise. Compared to FastSpeech, which serves as a strong non-autoregressive baseline, all other models show clear advantages in distributional quality. Overall, the results validate that the combination of ECAS and DPO in *SPEECH* offers a balanced and effective enhancement to diffusion-based TTS for elderly Cantonese, outperforming both diffusion and non-diffusion baselines in terms of clarity, naturalness, and expressiveness.

5.2. Subjective Evaluation

A total of 32 samples were randomly selected from each model on both datasets: our Elderly-Centric dataset and the CommonVoice-Yue dataset. Ground-truth human recordings from both datasets were also rated under the same criteria to serve as upper-bound references. Six audio professionals independently rated the samples on a 1–100 scale using the following metrics: (1) *Overall Quality (OVL)*, which reflects the naturalness and clarity of the speech, and (2) *Relevance to Text (REL)*, which assesses the consistency between the generated audio and the ground truth.

As shown in Table 3, *SPEECH* achieves the best subjective performance among all synthetic models on both datasets. On the Elderly Cantonese dataset, *SPEECH* reaches an OVL of 86.1 and a REL of 86.5, closely approaching the ground truth (94.2 and 93.6, respectively). This reflects the model's strong ability to synthesize clear, intelligible, and faithful speech aligned with the auditory needs of elderly listeners. Although performance drops slightly on the more acoustically variable CommonVoice-Yue dataset, *SPEECH* still leads in OVL (61.1) and remains competitive in REL (61.7), outperforming other baselines such as VITS, FastSpeech, and Tacotron2. These findings support the benefit of our ECAS and DPO modules in enhancing perceptual speech quality and linguistic alignment, particularly when the model is trained on specialized elderly-oriented data.

5.3. Ablation Studies

We conducted an ablation study to examine the individual impacts of the *Elderly-Centric Acoustic Supervising* and

Table 3

Subjective evaluation results on the Elderly Cantonese Speech Dataset and CommonVoice-Yue. Ground truth refers to real human recordings.

Model	Elderly Cantonese		CommonVoice-Yue	
	OVL \uparrow	REL \uparrow	OVL \uparrow	REL \uparrow
Tacotron2	71.3	68.7	40.8	39.5
FastSpeech	76.5	74.2	43.0	42.1
AudioLDM	78.9	79.6	47.2	48.0
VITS	82.7	84.3	50.2	52.5
SPEECH	86.1	86.5	61.1	61.7
<i>Ground Truth</i>	94.2	93.6	81.4	83.1

Table 4

Ablation study results (OVL / REL). Ground truth human recordings are included as reference points.

Model Configuration	Elderly Cantonese		CommonVoice-Yue	
	OVL \uparrow	REL \uparrow	OVL \uparrow	REL \uparrow
<i>Ground Truth</i>	94.5	96.2	88.0	89.5
Ours w/o ECAS & DPO	78.2	76.7	60.4	58.9
Ours w/o ECAS	82.6	84.1	66.3	63.8
Ours w/o DPO	84.0	85.5	69.0	65.2
SPEECH (Full)	86.1	86.5	71.1	67.7

Direct Preference Optimization components. Specifically, we evaluated three configurations of our model: (1) without both ECAS and DPO, (2) without ECAS, and (3) without DPO. Evaluation was performed on both the Elderly Cantonese Speech Dataset and the CommonVoice-Yue dataset using subjective metrics, Overall Quality and Relevance to Text, rated by human annotators on a 1–100 scale.

As shown in Table 4, the full *SPEECH* system achieves the highest OVL and REL scores across both datasets, approaching the subjective quality of natural human recordings, especially on the Elderly Cantonese dataset. Removing both ECAS and DPO results in the largest drop in perceived quality and relevance. Excluding only DPO leads to slight degradation in preference alignment, while excluding only ECAS causes a sharper decline in clarity-related perception. These results confirm that ECAS enhances acoustic clarity, while DPO improves subjective satisfaction, with both modules contributing uniquely to overall performance.

6. Conclusion

In this work, we present *SPEECH*, a diffusion-based text-to-speech system specifically designed to enhance clarity and intelligibility for elderly Cantonese listeners. Distinct from conventional TTS pipelines, *SPEECH* integrates two novel components: the *Elderly-Centric Acoustic Supervising* mechanism, which enforces frequency- and intelligibility-aware constraints during training to promote clearer articulation, and the *Direct Preference Optimization* module, which fine-tunes the model based on real listener feedback to align outputs with human auditory preferences. Through both

objective metrics and human evaluation, *SPEECH* demonstrates consistent improvements over strong baseline models in clarity, naturalness, and listener satisfaction. Our ablation studies further validate the complementary roles of ECAS and DPO in achieving these gains. Future work may expand the scope of human feedback, incorporate audiogram-based perceptual modeling, and explore multi-speaker adaptation to support broader personalization. This study lays a foundation for accessibility-oriented TTS systems that consider real-world perceptual needs and user preferences.

7. Limitation

While our proposed *SPEECH* system demonstrates improved clarity and preference alignment for elderly Cantonese listeners, it has several limitations. 1): the effectiveness of the Elderly-Centric Acoustic Supervising relies on heuristic approximations (e.g., frequency band emphasis and ASR-based intelligibility proxies) rather than direct perceptual modeling of hearing loss profiles. 2): the Direct Preference Optimization stage depends on a limited number of human annotations, which may constrain generalization across demographics and speaking styles. 3): the computational overhead of diffusion-based generation remains higher than that of autoregressive models, posing challenges for real-time deployment on resource-constrained devices. Lastly, the scope of our experimental study is constrained by limited human and computational resources, and the sample size for subjective evaluation and the scale of human feedback are relatively small due to the intensity of the pipeline, which was developed and executed as a single-researcher effort.

References

- [1] T. R. H. K. S. Government, "Census and statistics department. 2021. special topics report no. 63: Persons with disabilities and chronic diseases." 2021.
- [2] healthbureau, "Health bureau of hong kong. preventive care for older adults - core document," 2021.
- [3] chp.gov.hk., "Centre for health protection. (2016). non-communicable disease watch: Hearing difficulties." 2016.
- [4] H. Liu, K. Chen, Y. Yuan, S. Liu, Z. Zhang, Y. Wang, X. Qiu, and L. Xie, "Audioldm: Text-to-audio generation with latent diffusion models," *arXiv preprint arXiv:2301.12503*, 2023.
- [5] K. Shen, Z. Liu, Y. Shen, and et al., "Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers," *arXiv preprint arXiv:2304.09116*, 2023.
- [6] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *Proc. Interspeech*, 2017, pp. 4006–4010.
- [7] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan, R. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4779–4783.
- [8] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.
- [9] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 3617–3621.
- [10] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brebisson, Y. Bengio, and A. Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019, pp. 14 910–14 921.
- [11] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," *arXiv preprint arXiv:2209.14275*, 2022.
- [12] S. Li and H. Meng, "Personalized speech synthesis for the elderly," pp. 1–13, 2020.
- [13] R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, and C. Finn, "Direct preference optimization: Your language model is secretly a reward model," *arXiv preprint arXiv:2305.18290*, 2023.
- [14] A. Défossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," *arXiv preprint arXiv:2006.12847*, 2020.
- [15] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "FastSpeech: Fast, robust and controllable text to speech," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.
- [16] J. Kim, S. Kim, J. Jung, B.-J. Kim, and K. Yoo, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *International Conference on Machine Learning (ICML)*, 2021.