BAYESIAN SPEECH SYNTHESIZERS CAN LEARN FROM MULTIPLE TEACHERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-to-speech (TTS) is a "one-to-many" problem: one text can be spoken in many valid ways, so its realizations form a distribution whose spread depends on the input. Most TTS models do not model this spread explicitly: they leave it implicit in the sampling process or fix it in advance. To model this distribution explicitly, we propose **BELLE** (**B**ayesian **e**vidential **l**earning with **l**anguag**e** modelling)[1], which places a Bayesian prior over the mean and variance of each mel-spectrogram frame and predicts this prior from the input, while keeping the architecture and single-pass sampling of common continuous-valued autoregressive TTS models such as MELLE. Estimating this variance requires several recordings of each text, whereas standard TTS corpora provide one. We therefore construct two parallel-speech corpora: **ParallelVox** from real human recordings and **LS+6TTS**, which adds speech from six pre-trained TTS teachers to LibriSpeech (LS). Trained on these corpora, BELLE outperforms MELLE under the same training conditions, is competitive with larger TTS systems on the objective metrics of LibriSpeech test-clean, and supports streaming generation. Analyses of repeated human readings show that real speech has the properties the Bayesian prior assumes and that the variance predicted by BELLE follows the variation across these readings. Parallel speech improves BELLE, and synthetic parallel speech performs comparably to real parallel speech at the same volume and is far easier to scale.

1 INTRODUCTION

Text-to-Speech (TTS) synthesis is fundamentally a "one-to-many" mapping problem, because one text can be spoken in many valid ways. Different speakers read the same sentence with different timbre, prosody, and rhythm, and the same speaker also reads it differently from one repetition to the next. The valid realizations of one text therefore form a distribution, and how widely this distribution spreads depends on the text and its context. A TTS model should thus learn this distribution, including its spread for each input, rather than a single output.

Existing TTS models fall into three main families. Discrete codec-based autoregressive (AR) models (Borsos et al., 2023; Wang et al., 2023) predict codec tokens, but tokenization introduces quantization error. Diffusion TTS models (Eskimez et al., 2024; Chen et al., 2024b; Lee et al., 2025) generate continuous features from noise through iterative denoising. Continuous-valued AR models predict mel-spectrogram frames one by one, either with diffusion or flow-matching heads (Liu et al., 2024; Wang et al., 2025a; Jia et al., 2025) or, as in MELLE (Meng et al., 2024), with a Gaussian whose Kullback-Leibler (KL) loss is minimized when the variance equals one for every frame, regardless of the input. In these models, the spread of the "one-to-many" distribution is either left implicit in the sampling process or fixed to a constant, rather than modeled explicitly.

To model this distribution explicitly, we propose BELLE (**B**ayesian **e**vidential **l**earning with **l**anguag**e** modelling), a continuous-valued AR TTS model built on evidential deep learning (Amini et al., 2020). Instead of fixing the variance, BELLE places a Bayesian prior over the mean and variance of each frame and predicts this prior from the input, so the spread of the distribution is estimated for each input rather than set in advance. This change only affects the layer that outputs each frame, so BELLE keeps the architecture and single-pass sampling of MELLE.

[1]Audio samples and code are publicly available at `https://anonymous-pages229.github.io/BELLE-demo/` and `https://github.com/anonymous-pages229/BELLE`.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Implementing this Bayesian framework, however, introduces a statistical challenge: accurate variance estimation theoretically requires multiple observations per input, whereas standard TTS datasets typically provide only a single recording per text. Two complementary sources of "one-to-many" supervision are therefore explored. On the real side, we construct **ParallelVox**, a human parallel-speech corpus assembled from five public datasets, in which every text prompt is paired with multiple recordings. On the synthetic side, LibriSpeech (LS; Panayotov et al., 2015) is augmented with parallel speech from six pre-trained TTS teachers (denoted **LS+6TTS**), which gives each text multiple realizations.

Trained on these parallel-speech corpora, BELLE outperforms MELLE under the same training conditions, is competitive with larger TTS systems on the objective metrics of LibriSpeech test-clean, and supports streaming generation. Analyses of repeated human readings of the same text show that real speech has the properties this Bayesian prior assumes, and that the variance predicted by BELLE changes with the input and follows the variation across these readings. On the data side, parallel speech improves BELLE over single-recording data of similar volume, and synthetic parallel speech performs comparably to real parallel speech at the same volume and is far easier to scale.

Summary of Contributions.

- We propose BELLE, a TTS model that explicitly models the "one-to-many" distribution with evidential deep learning, predicting a Bayesian prior over the mean and variance of each frame.
- We construct two parallel-speech corpora with multiple realizations per text: ParallelVox from real human recordings and LS+6TTS from six pre-trained TTS teachers. At the same volume, synthetic parallel speech performs comparably to real parallel speech and is far easier to scale.
- Trained on these corpora, BELLE outperforms MELLE under the same training conditions, is competitive with larger TTS systems on the objective metrics of LibriSpeech test-clean, and supports streaming generation.

2 Related Work

Continuous-valued TTS. Continuous-valued TTS models generate acoustic features such as mel-spectrograms or variational autoencoder (VAE) latents without a discrete tokenizer. Non-autoregressive (NAR) models such as E2 TTS (Eskimez et al., 2024) and F5-TTS (Chen et al., 2024b) generate the whole utterance through iterative flow-matching refinement. AR models generate one frame at a time and differ in how each frame is sampled: MELLE (Meng et al., 2024) samples from a Gaussian trained with a KL loss toward a unit-variance target, several models attach diffusion or flow-matching heads (Liu et al., 2024; Wang et al., 2025a; Jia et al., 2025), and others predict a Gaussian or Gaussian-mixture distribution over VAE latents (Lin & He, 2025; Zhu et al., 2024). These VAE-based models learn the distribution of each frame from one recording per text, with a likelihood loss or a KL loss toward the VAE posterior of that recording. BELLE keeps single-pass sampling of each frame, but places a Bayesian prior over the mean and variance of the frame, and when parallel speech is available, trains this prior on all realizations of a text.

Streaming TTS. Streaming TTS systems output speech while the input text is still arriving. Several systems pair a language model over discrete speech tokens with a flow-matching decoder (Du et al., 2024; Sheng et al., 2025) or interleave text and speech tokens (Yang et al., 2024), while SMLLE (Sun et al., 2025) and StreamMel (Wang et al., 2025b) generate continuous features directly. BELLE-stream also generates mel-spectrograms within one continuous-valued AR model by interleaving text and speech chunks, and it uses the same evidential sampling as BELLE, so the Bayesian prior is predicted for every frame during streaming.

Evidential Learning and Bayesian Neural Networks. Bayesian deep learning places priors either on network weights, as in Bayesian Neural Networks (BNNs) (MacKay, 1992; Neal, 1996), or on the parameters of the likelihood, as in Evidential Deep Learning (EDL) (Sensoy et al., 2018; Amini et al., 2020). BNNs approximate the weight posterior with variational inference (Graves, 2011; Blundell et al., 2015) or Monte Carlo dropout (Gal & Ghahramani, 2016), so a prediction typically requires several forward passes, and in AR TTS this extra cost is paid at every frame. EDL instead predicts the parameters of a conjugate prior, such as the Normal-Inverse-Gamma (NIG) prior over the mean and

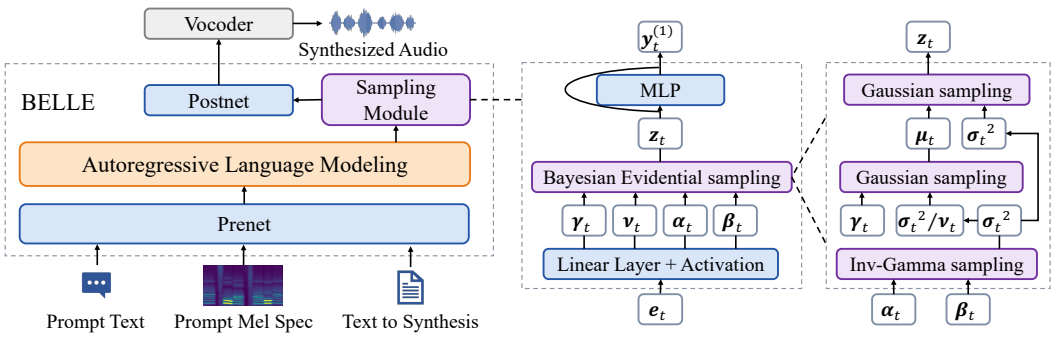

Figure 1: Overview of BELLE. Left: the mel-based AR TTS architecture (Sec. 3.2). Right: the Sampling Module, which predicts a Bayesian prior over the mean and variance of each frame and samples the frame from it (Sec. 4).

variance of a Gaussian, and marginalizes over them in closed form within one forward pass, which places it in the Empirical Bayes framework (Appendix B). BELLE applies EDL to every frame of a continuous-valued AR TTS model, so the prior over each frame is predicted at the cost of a single forward pass.

3 Mel-based Autoregressive TTS

3.1 Problem Formulation

Mel-based AR TTS models formulate synthesis as next-frame prediction. Given an input text sequence $\boldsymbol{x} = [x_1, \ldots, x_L]$, the model generates a mel-spectrogram $\boldsymbol{y} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T]$, where each frame $\boldsymbol{y}_t \in \mathbb{R}^D$ is the spectrum at time step t and D is the number of mel-frequency bands. Each frame $\boldsymbol{y}_t$ is generated conditioned on the text $\boldsymbol{x}$ and the previous frames $\boldsymbol{y}_{<t} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{t-1}]$, so the distribution of $\boldsymbol{y}$ factorizes as

$$p(\boldsymbol{y} \mid \boldsymbol{x}; \theta) = \prod_{t=1}^{T} p(\boldsymbol{y}_t \mid \boldsymbol{y}_{<t}, \boldsymbol{x}; \theta), \tag{1}$$

where θ denotes the parameters of the neural network that models each conditional distribution.

In **streaming generation**, speech is output while the input text is still arriving. Both the input text and the target mel-spectrogram are therefore split into K consecutive chunks, $\boldsymbol{X} = [\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(K)}]$ and $\boldsymbol{Y} = [\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(K)}]$. Unlike Eq. (1), where every frame is conditioned on the full text, the k-th audio chunk $\boldsymbol{y}^{(k)}$ is conditioned only on the previous audio chunks $\boldsymbol{y}^{(<k)}$ and on the text chunks received so far, $\boldsymbol{x}^{(\leq k)}$:

$$p(\boldsymbol{Y} \mid \boldsymbol{X}; \theta) = \prod_{k=1}^{K} p\big(\boldsymbol{y}^{(k)} \mid \boldsymbol{y}^{(<k)}, \boldsymbol{x}^{(\leq k)}; \theta\big). \tag{2}$$

3.2 Model Architecture

As shown in Fig. 1, a general mel-based AR TTS model consists of a **Prenet**, an **AR LM** (AR language model), a **Sampling Module**, a **Stop-Prediction Module**, a **Postnet**, and a **Vocoder**. The Prenet projects the text tokens $\boldsymbol{x}$ and the previous mel frames to the hidden dimension of the AR LM, a decoder-only Transformer. At each step t, the AR LM attends to the text and to the frames before t and outputs a hidden representation $\boldsymbol{e}_t$, from which frame t is generated. The Sampling Module projects $\boldsymbol{e}_t$ to the parameters of a distribution over the frame, which is an NIG distribution in BELLE (Sec. 4), and draws a sample $\boldsymbol{z}_t$ from it. Because $\boldsymbol{z}_t$ is a random sample, it carries sampling noise, so a denoising network, implemented as a multilayer perceptron (MLP), refines $\boldsymbol{z}_t$ into the frame $\boldsymbol{y}_t^{(1)}$ (Fig. 1, right). Because the model generates continuous mel frames rather than discrete tokens, it cannot end generation by emitting an end-of-sequence token as discrete AR models do. Instead, as in TransformerTTS and SpeechT5 (Li et al., 2019; Ao et al., 2022), the Stop-Prediction Module, a linear layer with sigmoid activation applied to $\boldsymbol{e}_t$, outputs a stop score in $[0, 1]$, and generation halts once the score exceeds a threshold. Following Tacotron2 and TransformerTTS (Shen et al., 2018;

Li et al., 2019), the Postnet then adds a convolutional residual to the generated mel-spectrogram $\boldsymbol{y}^{(1)}$, yielding the refined mel-spectrogram $\boldsymbol{y}^{(2)}$, which the Vocoder converts into a waveform.

BELLE-stream uses the same architecture and implements Eq. (2) by interleaving text and audio chunks as the input of the AR LM under a causal mask (Fig. 2). Each text chunk $\boldsymbol{x}^{(k)}$ contains S_{text} tokens and each audio chunk $\boldsymbol{y}^{(k)}$ contains S_{audio} frames. S_{audio} is kept small so that each audio chunk is generated after enough text has been received. As a result, the generated speech lags behind the text, and when the last text chunk arrives, more than S_{audio} frames typically remain to be generated. The final audio chunk therefore has no fixed length: generation continues until the Stop-Prediction Module halts it, so this chunk is longer than the preceding ones.

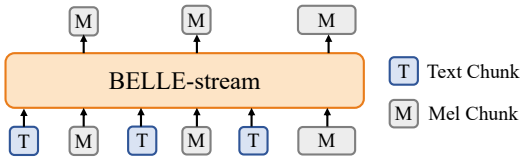

Figure 2: Interleaved text and mel-spectrogram chunks in BELLE-stream.

4 Bayesian Evidential Learning

4.1 Evidential Modeling of Mel Frames

EDL places a prior over the parameters of the likelihood and lets the network predict the parameters of this prior. In BELLE, each mel-spectrogram frame $\boldsymbol{y}_t \in \mathbb{R}^D$ is modeled as a Gaussian with unknown mean $\boldsymbol{\mu}_t$ and variance $\boldsymbol{\sigma}_t^2$, and an NIG prior is placed over them:

$$\boldsymbol{y}_t \sim \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\sigma}_t^2), \quad \boldsymbol{\mu}_t \sim \mathcal{N}\left(\boldsymbol{\gamma}_t, \frac{\boldsymbol{\sigma}_t^2}{\boldsymbol{\nu}_t}\right), \quad \boldsymbol{\sigma}_t^2 \sim \Gamma^{-1}(\boldsymbol{\alpha}_t, \boldsymbol{\beta}_t), \tag{3}$$

where the NIG parameters $\boldsymbol{\gamma}_t, \boldsymbol{\nu}_t, \boldsymbol{\alpha}_t, \boldsymbol{\beta}_t \in \mathbb{R}^D$ satisfy $\nu_{t,d} > 0$, $\alpha_{t,d} > 1$, and $\beta_{t,d} > 0$ for every dimension $d \in \{1, \ldots, D\}$. All distributions act element-wise, so the D mel bands are modeled independently (Appendix C discusses this choice).

Integrating out $\boldsymbol{\mu}_t$ and $\boldsymbol{\sigma}_t^2$ gives the predictive distribution of the frame, which is a Student-t distribution in each dimension:

$$p(\boldsymbol{y}_t \mid \boldsymbol{\gamma}_t, \boldsymbol{\nu}_t, \boldsymbol{\alpha}_t, \boldsymbol{\beta}_t) = \mathrm{St}\left(\boldsymbol{y}_t;\, \boldsymbol{\gamma}_t,\, \frac{\boldsymbol{\beta}_t \odot (1 + \boldsymbol{\nu}_t)}{\boldsymbol{\nu}_t \odot \boldsymbol{\alpha}_t},\, 2\boldsymbol{\alpha}_t\right), \tag{4}$$

where $\odot$ denotes element-wise multiplication.

4.2 From Gaussian to Evidential Sampling

In MELLE (Meng et al., 2024), the Sampling Module predicts a diagonal Gaussian from the hidden representation $\boldsymbol{e}_t$ with a linear projection,

$$\left[\boldsymbol{\mu}_t, \log \boldsymbol{\sigma}_t^2\right] = \boldsymbol{W}_{\text{G}} \boldsymbol{e}_t + \boldsymbol{b}_{\text{G}}, \tag{5}$$

and draws $\boldsymbol{z}_t = \boldsymbol{\mu}_t + \boldsymbol{\sigma}_t \odot \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, a reparameterization that keeps the sampling step differentiable. This Gaussian is trained with a KL divergence toward a target distribution centered at the ground-truth frame $\boldsymbol{y}_t$ with unit variance:

$$\begin{aligned} \mathcal{L}_{\text{KL}} &= \sum\nolimits_{t=1}^{T} D_{\text{KL}}\left(\mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\sigma}_t^2 \boldsymbol{I}) \,\|\, \mathcal{N}(\boldsymbol{y}_t, \boldsymbol{I})\right) \\ &= \frac{1}{2} \sum\nolimits_{t=1}^{T} \left(\|\boldsymbol{\mu}_t - \boldsymbol{y}_t\|_2^2 + \sum\nolimits_{d=1}^{D} \left(\sigma_{t,d}^2 - \log \sigma_{t,d}^2\right) - D \right). \end{aligned} \tag{6}$$

The first term is a regression loss on the mean. Each variance term $\sigma_{t,d}^2 - \log \sigma_{t,d}^2$ is minimized at $\sigma_{t,d}^2 = 1$ independently of $\boldsymbol{e}_t$, so the KL loss pulls the variance of every frame toward the same constant regardless of the input.

BELLE replaces this Gaussian with the NIG prior of Sec. 4.1, as shown in the right part of Fig. 1. A linear layer predicts the NIG parameters from $\boldsymbol{e}_t$:

$$[\boldsymbol{\gamma}_t, \boldsymbol{\nu}_t, \boldsymbol{\alpha}_t, \boldsymbol{\beta}_t] = \text{Activation}(\boldsymbol{W}_{\text{NIG}} \boldsymbol{e}_t + \boldsymbol{b}_{\text{NIG}}), \tag{7}$$

where Activation$(\cdot)$ applies softplus to $\boldsymbol{\nu}_t$ and $\boldsymbol{\beta}_t$ and a shifted softplus to $\boldsymbol{\alpha}_t$, so that the constraints in Sec. 4.1 hold. Each frame is then sampled hierarchically, following the generative process of Eq. (3):

$$\boldsymbol{\sigma}_t^2 \sim \Gamma^{-1}(\boldsymbol{\alpha}_t, \boldsymbol{\beta}_t), \quad \boldsymbol{\mu}_t \sim \mathcal{N}\left(\boldsymbol{\gamma}_t, \frac{\boldsymbol{\sigma}_t^2}{\boldsymbol{\nu}_t}\right), \quad \boldsymbol{z}_t \sim \mathcal{N}\left(\boldsymbol{\mu}_t, \boldsymbol{\sigma}_t^2\right). \tag{8}$$

The resulting $\boldsymbol{z}_t$ follows the Student-t distribution of Eq. (4), whose scale is determined by $\boldsymbol{\nu}_t$, $\boldsymbol{\alpha}_t$, and $\boldsymbol{\beta}_t$. Because these parameters are predicted from $\boldsymbol{e}_t$, the spread of $\boldsymbol{z}_t$ changes with the input. In place of the KL loss of Eq. (6), the NIG parameters are trained with the evidential loss, summed over frames and mel bands:

$$\begin{aligned}
\mathcal{L}_{\text{edl}} &= \sum\nolimits_{t=1}^{T} \sum\nolimits_{d=1}^{D} \left(\mathcal{L}_{\text{NLL}} + \lambda\, \mathcal{L}_{\text{R}}\right), \\
\mathcal{L}_{\text{NLL}} &= \frac{1}{2}\log\frac{\pi}{\nu} - \alpha\log\Omega + \left(\alpha + \frac{1}{2}\right)\log\left(\nu(y-\gamma)^2 + \Omega\right) + \log\frac{\Gamma(\alpha)}{\Gamma(\alpha+\frac{1}{2})}, \\
\mathcal{L}_{\text{R}} &= |y-\gamma|\,(2\nu + \alpha),
\end{aligned} \tag{9}$$

where the indices t, d are omitted inside the sum, $\Omega = 2\beta(1+\nu)$, and $\Gamma(\cdot)$ is the gamma function. $\mathcal{L}_{\text{NLL}}$ is the negative log-likelihood (NLL) of the ground-truth frame under the Student-t distribution of Eq. (4). The regularizer $\mathcal{L}_{\text{R}}$ penalizes the evidence $2\nu + \alpha$, which measures the confidence of the prior, in proportion to the error $|y - \gamma|$, so that the prior becomes less confident on frames that are predicted poorly. Unlike Eq. (6), this loss contains no term that pulls the variance toward a fixed value, so the spread of the predictive distribution is fit to the observed frames. Appendix D gives the derivations.

4.3 Training Objectives

BELLE is trained end-to-end with four losses:

$$\mathcal{L} = \mathcal{L}_{\text{reg}} + \lambda_{\text{samp}}\mathcal{L}_{\text{samp}} + \lambda_{\text{flux}}\mathcal{L}_{\text{flux}} + \mathcal{L}_{\text{stop}}, \tag{10}$$

where λ_{samp} and λ_{flux} weight the sampling and flux losses (values in Appendix E.2).

The regression loss $\mathcal{L}_{\text{reg}}$ applies L1 and L2 losses between the ground-truth frames $\boldsymbol{y}^{\text{gt}}$ and both the output of the denoising MLP, $\boldsymbol{y}^{(1)}$, and the output of the Postnet, $\boldsymbol{y}^{(2)}$:

$$\mathcal{L}_{\text{reg}} = \sum\nolimits_{j=1}^{2} \left(\|\boldsymbol{y}^{\text{gt}} - \boldsymbol{y}^{(j)}\|_1 + \|\boldsymbol{y}^{\text{gt}} - \boldsymbol{y}^{(j)}\|_2^2\right). \tag{11}$$

The sampling loss $\mathcal{L}_{\text{samp}}$ trains the Sampling Module. It is the evidential loss $\mathcal{L}_{\text{edl}}$ of Eq. (9) in BELLE and the KL loss of Eq. (6) in MELLE.

The spectrogram flux loss $\mathcal{L}_{\text{flux}}$ rewards a large L1 distance between the predicted location $\boldsymbol{\gamma}_t$ of frame t and the previous ground-truth frame $\boldsymbol{y}_{t-1}^{\text{gt}}$, which penalizes predictions that copy the previous frame:

$$\mathcal{L}_{\text{flux}} = -\sum\nolimits_{t=2}^{T} \|\boldsymbol{\gamma}_t - \boldsymbol{y}_{t-1}^{\text{gt}}\|_1. \tag{12}$$

MELLE applies the same loss to its Gaussian mean $\boldsymbol{\mu}_t$ in place of $\boldsymbol{\gamma}_t$.

The stop loss $\mathcal{L}_{\text{stop}}$ is a binary cross-entropy loss on the output of the Stop-Prediction Module (Sec. 3.2). Only the last frame of each utterance is a positive sample, so its term is weighted by 500.

When a text $\boldsymbol{x}$ has $N \geq 1$ realizations $\boldsymbol{y}_1^{\text{gt}}, \ldots, \boldsymbol{y}_N^{\text{gt}}$, where N may differ across texts, each realization is used as a separate ground truth. With $\mathcal{L}(\boldsymbol{y}_i^{\text{gt}})$ denoting the loss of Eq. (10) computed against $\boldsymbol{y}_i^{\text{gt}}$, the training loss for $\boldsymbol{x}$ is the weighted sum

$$\mathcal{L}_{\boldsymbol{x}} = \sum\nolimits_{i=1}^{N} w_i \mathcal{L}(\boldsymbol{y}_i^{\text{gt}}), \quad \text{with} \quad \sum\nolimits_{i=1}^{N} w_i = 1, \quad w_i \geq 0. \tag{13}$$

Using all N realizations of $\boldsymbol{x}$ as targets gives more direct evidence for the spread of Sec. 4.2 than a single recording provides. Appendix F shows that this objective is the likelihood of the N realizations under the AR factorization, although each realization is teacher-forced on its own previous frames.

5 PARALLEL SPEECH CORPORA

To help BELLE estimate the spread of the "one-to-many" distribution, two parallel-speech corpora with multiple realizations per text are constructed: ParallelVox from real human recordings and LS+6TTS from TTS synthesis.

5.1 PARALLELVOX: REAL PARALLEL SPEECH

ParallelVox merges five public English corpora in which the same text is recorded several times, either by different speakers or by one speaker on different occasions: LibriHeavy (Kang et al., 2024), the Parallel Audiobook Corpus, VCTK (Yamagishi et al., 2019), CMU ARCTIC (Kominek & Black, 2004) with L2-Arctic (Zhao et al., 2018), and the parallel subset of Mozilla Common Voice (Ardila et al., 2020). Each recording is transcribed with Whisper-large-v2 (Radford et al., 2023) and dropped if its word error rate (WER) exceeds 10%, and a text is kept only if at least two of its recordings remain. Texts whose recordings all come from one speaker form the single-speaker repeat test set (22 h, 250 speakers, 2.1 recordings per text on average), and the remaining texts form the multi-speaker training set (881 h, 3,406 speakers, 2.8 recordings per text on average). The test set records how one speaker varies across repetitions of the same text, so it serves as the human reference for the diversity evaluation (Sec. 6.2). On the training set, all recordings of a text are weighted equally in Eq. (13). Appendix G.1 gives the source statistics and the filtering details.

5.2 LS+6TTS: SYNTHETIC PARALLEL SPEECH

Real parallel corpora are costly to extend: ParallelVox reaches 881 h only after five corpora are merged. **LS+6TTS** instead adds synthetic realizations to a corpus with one recording per text. The LibriSpeech (Panayotov et al., 2015) training set is trimmed with voice activity detection and restricted to utterances of 0.5 to 14 s, which gives 706 h (denoted **LS**). Six pre-trained TTS teachers, CosyVoice2 (Du et al., 2024), IndexTTS (Deng et al., 2025), SparkTTS (Wang et al., 2025c), F5TTS (Chen et al., 2024b), MaskGCT (Wang et al., 2024), and XTTS-v2 (Casanova et al., 2024), each synthesize every text, so each text has $N = 7$ realizations and the corpus totals 4,817 h. Appendix G.2 gives the processing details.

In Eq. (13), the human recording is weighted 0.22 and each teacher 0.13, so the human recording receives about 1.7 times the weight of any single teacher. Appendix G.3 compares other weight allocations.

6 EXPERIMENTAL SETUP

6.1 TRAINING SETTINGS

BELLE and MELLE share the same backbone and are each trained on three corpora: LS, LS+6TTS, and the ParallelVox training set. Both use every realization of a text as a training target with the weights of Eq. (13), which reduces to standard training on LS, where $N = 1$. To compare real and synthetic parallel speech at equal volume, BELLE is also trained on **LS+6TTS-881h**, a subset of LS+6TTS that matches the ParallelVox training set in hours; the subset is sampled by text, so every retained text keeps all seven realizations, and it is trained with the ParallelVox budget. BELLE-stream is initialized from BELLE trained on LS+6TTS and further trained on the same corpus with $S_{\text{text}} = 20$ and $S_{\text{audio}} = 50$ (≈ 0.8 s per audio chunk). Model and training details are in Appendix E.1 and E.2.

6.2 EVALUATION SETTINGS

Intelligibility, speaker similarity, and quality are evaluated on the LibriSpeech test-clean subset, restricted to utterances of 4 to 10 s, with an open-source zero-shot TTS evaluation protocol (Lee, 2024). Two inference conditions are used. In **Continuation**, the first 3 s of an utterance and its transcription serve as the prompt, and the model generates the rest of the utterance. In **Cross-Sentence**, an utterance of a speaker and its transcription serve as the prompt, and the model generates a different sentence in the voice of that speaker.

Intelligibility is measured by WER-H, the WER of a HuBERT-based (Hsu et al., 2021) automatic speech recognition (ASR) model on the generated speech. Speaker similarity (SIM-o) is the cosine similarity between the WavLM-TDCNN (Chen et al., 2022) speaker embeddings of the generated speech and of the original speech prompt. For subjective evaluation, a mean opinion score (MOS) for overall quality and a similarity MOS (SMOS) for speaker similarity to the prompt are collected on a crowd-sourcing platform (Appendix H.2).

Diversity is evaluated on the ParallelVox single-speaker repeat test set (Sec. 5.1). Under the Cross-Sentence condition, each system synthesizes every prompt three times, and the three outputs form one group. Each sample is represented by the mean-pooled and L2-normalized hidden states of layer 13 of WavLM-Large (Chen et al., 2022), a layer chosen following the layer-wise probing analysis of Chiu et al. (2025). Diversity (Cos.) is the mean pairwise cosine distance within a group, averaged over all groups, and the Ground Truth value is computed in the same way over the repeated human readings of each text. On these readings, this distance correlates with how much duration, F0, and energy vary from one reading to another (Appendix H.4). Generated samples with a WER above 5% are dropped before this computation, since their distance reflects generation errors rather than acoustic variation. Appendix H gives further details.

7 Results

7.1 Main Results

BELLE outperforms MELLE and is competitive with larger TTS systems. BELLE and MELLE share the backbone, the training corpus, and the training schedule and differ in the Sampling Module and the loss that trains it (Sec. 4.2), so their gap measures the effect of the evidential formulation. BELLE outperforms MELLE on nearly every metric under all three training corpora (Table 1; Appendix I.1), including LS, which has one recording per text, so the evidential formulation helps even without parallel speech. Against these TTS systems, which have up to 40 times more parameters or 50 times more training data (Table 2), BELLE gives the best Continuation results, ranks third on both Cross-Sentence metrics, and has the lowest RTF while supporting streaming generation.

Parallel realizations improve most objective metrics. To isolate the effect of parallel realizations, BELLE is trained on LS and on LS+6TTS-881h, two corpora of similar volume built on LS that differ in whether each text has one realization or seven. With seven realizations per text, four of the five objective metrics improve, namely Cross-Sentence WER-H, both SIM-o scores, and diversity, whereas Continuation WER-H becomes worse. Several realizations of a text thus supply the spread that the evidential loss estimates.

Real and synthetic parallel speech are comparable, and synthetic speech scales further. To compare the two sources at equal volume, BELLE trained on ParallelVox is compared with BELLE trained on LS+6TTS-881h under the same training budget. Each model leads on the test set closer to its own training domain: the synthetic corpus, built on LS, gives lower WER-H and higher SIM-o on LibriSpeech test-clean, whereas ParallelVox gives higher diversity on its own single-speaker repeat test set. Neither source dominates, so the two support BELLE at a comparable level. They differ in cost: ParallelVox reaches 881 h only after five public corpora are merged, whereas synthetic realizations can be generated for any corpus with one recording per text at the cost of TTS inference. Scaling LS+6TTS-881h to the full LS+6TTS improves every objective metric and yields the strongest BELLE in Table 1.

7.2 Where the Gain over MELLE Comes From

BELLE learns an input-dependent variance. The end-task metrics show that BELLE outperforms MELLE but do not show where the gain comes from. To test whether it comes from the per-frame variance that BELLE predicts from the input (Sec. 4.2) and which the KL loss of MELLE pulls toward one for every frame, the per-frame standard deviation σ_t is exported from BELLE and MELLE trained on LS+6TTS on 196 Cross-Sentence utterances. Because the two models place σ_t on different scales, only relative quantities are compared. MELLE's σ_t is almost constant, as Eq. (6) predicts, with a coefficient of variation (CV) of 6% within an utterance and 1.2% across utterances. BELLE's σ_t varies much more, with a CV of 19% within an utterance and 7.0% across utterances, so its predicted variance changes with the input both from frame to frame and from sentence to sentence. Both models

Table 1: Controlled comparison under a shared training and evaluation pipeline. WER-H and SIM-o are measured on LibriSpeech test-clean, and Cos. ($\times 10^{-4}$) on the ParallelVox single-speaker repeat test set. MOS and SMOS were collected only for Ground Truth and the models trained on LS+6TTS.

System	Training data	Continuation		Cross-Sentence		Diversity	Subjective	
		WER-H↓	SIM-o↑	WER-H↓	SIM-o↑	Cos.↑	MOS↑	SMOS↑
Ground Truth	-	2.15	0.668	2.15	0.672	46.7	4.20	3.81
VALL-E	ParallelVox	7.46	0.340	16.3	0.452	17.4	-	-
MELLE	ParallelVox	3.06	0.406	5.60	0.538	19.9	-	-
BELLE	ParallelVox	3.00	0.449	5.73	0.570	21.5	-	-
BELLE	LS+6TTS-881h	2.57	0.510	4.05	0.634	20.3	-	-
MELLE	LS	2.78	0.452	8.76	0.561	10.4	-	-
BELLE	LS	2.39	0.457	5.01	0.576	17.3	-	-
MELLE	LS+6TTS	2.59	0.488	3.83	0.606	21.3	4.02	3.80
BELLE	LS+6TTS	**2.13**	**0.519**	**2.99**	**0.641**	**23.9**	**4.21**	**4.13**

Table 2: Comparison with larger TTS systems on LibriSpeech test-clean. Data is the training volume. RTF is measured with batch size 1. Streaming indicates whether a system supports streaming generation. Llasa-8B uses direct inference. †Estimated from the reported architecture.

System	Params	Data	Continuation		Cross-Sentence		RTF↓	Streaming
			WER-H↓	SIM-o↑	WER-H↓	SIM-o↑		
Ground Truth	-	-	2.15	0.668	2.15	0.672	-	-
VALL-E (Wang et al., 2023)	0.37B†	60k h	3.8	-	5.9	-	-	✗
VALL-E 2 (Chen et al., 2024a)	-	50k h	2.32	0.504	**2.44**	0.643	-	✗
RALL-E (Xin et al., 2024)	0.37B†	44k h	-	-	2.8	0.49	-	✗
CLAM-TTS (Kim et al., 2024)	1.2B†	55k h	2.36	0.477	5.11	0.495	0.42	✗
Llasa-8B (Ye et al., 2025)	8B	250k h	2.29	0.483	-	-	0.79	✗
CosyVoice 2 (Du et al., 2024)	0.65B	167k h	-	-	3.37	**0.713**	0.33	✓
BELLE	0.2B	4.8k h	**2.13**	**0.519**	2.99	0.641	**0.28**	✓

assign a larger σ_t to frames with fast spectral change, such as onsets, offsets, and formant transitions, than to steady frames of similar energy (Appendix I.2), but the increase is 17% for BELLE, in every analyzed utterance, and only 5% for MELLE.

The learned variance follows real speech. To test whether the variation in BELLE's σ_t reflects real speech, BELLE trained on ParallelVox is teacher-forced on one reading of each sentence in the single-speaker repeat test set, and its predicted σ_t^2 is compared frame by frame with the variance across the repeated human readings of that sentence. Over 2,185 sentences, the two are positively correlated (median Spearman $\rho = 0.18$). Frames with fast spectral change, low energy, or an irregular alignment could raise both quantities at once, so the correlation is recomputed after these three factors are removed from both, and it stays at a median of 0.17, positive in 91% of the sentences. When the predicted variance of one sentence is instead compared with the readings of a different sentence, the correlation drops to about zero (Appendix I.2), so it reflects the content of each sentence rather than the analysis procedure. Spectral flux alone is not correlated with the variance across readings (median $\rho = -0.06$), so a predictor based on spectral change would not reproduce this result. The predicted variance of BELLE also responds more strongly than that of MELLE: the slope of $\log \sigma_t^2$ on the logarithm of the variance across readings is 0.069 for BELLE and 0.024 for MELLE. Together with the previous analysis, this ties the gain over MELLE to the input-dependent variance of Sec. 4.2.

7.3 Why the NIG Prior Fits Speech

BELLE places an NIG prior on every mel frame (Eq. (3)) instead of a Gaussian with a fixed or a free variance. To test whether speech calls for this choice, the repeated readings of each text in the ParallelVox single-speaker repeat test set are aligned by dynamic time warping (DTW), which gives

several observations of every frame without any model (Appendix I.2). These observations show two properties, each of which matches one part of the NIG prior and rules out a simpler distribution.

(i) Some frames vary several times more than others. Within an utterance, the standard deviation across repeated readings is about 3 times larger at the most variable frames than at the most stable ones (99th vs. 1st percentile), whereas sampling noise alone would give 1.3 times. A fixed variance, as in MELLE, cannot represent this difference, whereas BELLE predicts the spread of each frame from the input.

(ii) The variance of a frame is uncertain. A text has only a few readings, so the variance of a frame computed from them carries large sampling noise: most of its variation from frame to frame is noise rather than a real difference in spread (Appendix I.2). A free per-frame variance, as learned by a Gaussian NLL loss, is fit directly to these noisy values and therefore fits the noise as well. The Inverse-Gamma prior over σ_t^2 in BELLE instead treats the variance as uncertain and pulls its estimate toward the prior mean, so the noise in a few readings moves the estimate less. Consistent with this, a Gaussian NLL loss with a free per-frame variance, trained in place of MELLE's KL loss, trails both MELLE and BELLE (Appendix I.4). Integrating over the uncertain variance gives the Student-t predictive distribution of Eq. (4), whose tails are heavier than those of a Gaussian, and the real readings are themselves heavier-tailed than a Gaussian once the small number of readings and the alignment are accounted for (Appendix I.2).

Modeling these properties raises the held-out likelihood. To test whether the two properties hold beyond the texts they are measured on, the densities are fit on half of the texts and evaluated on the other half. Each held-out reading is predicted from the other readings of its text. Relative to a Gaussian with one global variance, a Gaussian whose variance at each frame and band is estimated from the other readings and pulled toward the global variance, as in (i) and (ii), raises the held-out log-likelihood by 0.033 nats per observation, and replacing it with the Student-t predictive distribution of an uncertain variance raises it by another 0.066 nats (Appendix I.2).

7.4 Streaming Generation

As a continuous-valued AR model that samples each frame in a single pass, BELLE can generate speech chunk by chunk while the text is still arriving (Appendix J). Among the streaming systems in Table 3, BELLE-stream has the lowest WER-H, and its SIM-o lies between those of SMLLE and IST-LM. Because BELLE-stream conditions each audio chunk only on the text received so far (Eq. (2)), its WER-H is higher than that of non-streaming BELLE. On one NVIDIA H200 GPU with batch size 1, BELLE-stream runs at a real-time factor (RTF) of 0.308 with a median first-packet latency of 232 ms, so it synthesizes speech faster than real time.

Table 3: Cross-Sentence results of streaming systems: SMLLE (Sun et al., 2025), IST-LM (Yang et al., 2024), and BELLE-stream trained on LS+6TTS.

System	WER-H↓	SIM-o↑
SMLLE	6.37	0.489
IST-LM	4.53	**0.653**
BELLE-stream	**4.01**	0.557

8 Conclusion

We propose BELLE, a TTS model that explicitly models the "one-to-many" distribution with evidential deep learning, predicting an NIG prior over the mean and variance of each mel-spectrogram frame from the input. We construct two parallel-speech corpora, ParallelVox from real human recordings and LS+6TTS from six pre-trained TTS teachers, which provide the multiple realizations per text that this variance estimation requires. Our experiments lead to three findings. First, BELLE outperforms MELLE under the same training conditions, is competitive with larger TTS systems on the objective metrics of LibriSpeech test-clean, and supports streaming generation. Second, real speech has the properties that the NIG prior assumes, and the variance predicted by BELLE changes with the input and follows the variation across repeated human readings of a text. Third, parallel speech improves BELLE, and at the same volume, synthetic parallel speech performs comparably to real parallel speech and is far easier to scale.

AI USE STATEMENT

We used large language models (LLMs) to polish the writing of our paper and to help edit code. The research ideas, method design, experiments, and analyses are our own. We reviewed all LLM-assisted text and code and take full responsibility for the content of the paper.

ETHICS STATEMENT

BELLE is a zero-shot TTS model: given a few seconds of a speaker's voice, it generates new speech in that voice. The same capability that serves assistive communication and spoken dialogue systems can be misused for impersonation, fraud, and the spread of fabricated speech. Detectors of synthetic speech, audio watermarking, and consent from the speakers whose voices are used as prompts reduce these risks, and we encourage their use together with models of this kind.

All human recordings used in this paper come from public corpora: LibriSpeech and the five sources of ParallelVox, whose licenses are listed in Table 4. LS+6TTS adds speech synthesized by six publicly released TTS models in the voices of LibriSpeech speakers, and this synthetic speech is used only to train and evaluate the models in this paper. The subjective evaluation was run with crowd-sourced raters who were paid for their participation (Appendix H.2).

REPRODUCIBILITY STATEMENT

Code and audio samples are available at the anonymous links given in the abstract. The model architecture is described in Sec. 3.2 and Appendix E.1, the Sampling Module and its losses in Sec. 4, and the derivation of the evidential loss in Appendix D. ParallelVox is built only from public corpora, and Appendix G.1 gives its sources, preprocessing, filtering thresholds, and train/test split. LS+6TTS is built from LibriSpeech with six public TTS models, and Appendix G.2 gives the preprocessing, the teachers, and how they are prompted. Training hyperparameters, including the optimizer, batch size, number of updates, learning-rate schedule, and loss weights of every model, are listed in Appendix E. The evaluation follows an open-source protocol, and Appendix H lists the ASR and speaker-verification models used, the subjective evaluation procedure, and the computation of the diversity metric. Appendix I.1 describes the significance tests, and Appendix I.2 describes the variance analyses of Sec. 7.2 and Sec. 7.3.

REFERENCES

Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. In *Proc. NeurIPS*, 2020.

Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, Zhihua Wei, Yao Qian, Jinyu Li, and Furu" Wei. SpeechT5: Unified-modal encoder-decoder pre-training for spoken language processing. In *Proc. ACL*, 2022.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. In *Proc. LREC*, 2020.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *Proc. ICML*, 2015.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, et al. Audiolm: a language modeling approach to audio generation. *IEEE/ACM transactions on audio, speech, and language processing*, 31:2523–2533, 2023.

Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Göknar, Iulian Gulea, Logan Hart, Aya Aljafari, Joshua Meyer, Reuben Morais, Samuel Olayemi, et al. Xtts: a massively multilingual zero-shot text-to-speech model. *arXiv preprint arXiv:2406.04904*, 2024.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. WavLM: Large-scale self-supervised pre-training

for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6): 1505–1518, 2022.

Sanyuan Chen, Shujie Liu, Long Zhou, Yanqing Liu, Xu Tan, Jinyu Li, Sheng Zhao, Yao Qian, and Furu Wei. VALL-E 2: Neural codec language models are human parity zero-shot text to speech synthesizers. *arXiv preprint arXiv:2406.05370*, 2024a.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. *arXiv preprint arXiv:2410.06885*, 2024b.

Aemon Yat Fei Chiu, Kei Ching Fung, Roger Tsz Yeung Li, Jingyu Li, and Tan Lee. A large-scale probing analysis of speaker-specific attributes in self-supervised speech representations. *arXiv preprint arXiv:2501.05310*, 2025.

Wei Deng, Siyi Zhou, Jingchen Shu, Jinchao Wang, and Lu Wang. Indextts: An industrial-level controllable and efficient zero-shot text-to-speech system. *arXiv preprint arXiv:2502.05512*, 2025.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, et al. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*, 2024.

Sefik Emre Eskimez, Xiaofei Wang, Manthan Thakker, Canrun Li, Chung-Hsien Tsai, Zhen Xiao, Hemin Yang, Zirun Zhu, Min Tang, Xu Tan, et al. E2 tts: Embarrassingly easy fully non-autoregressive zero-shot tts. In *Proc. SLT*, 2024.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proc. ICML*, 2016.

Alex Graves. Practical variational inference for neural networks. In *Proc. NeurIPS*, volume 24, 2011.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. Conformer: Convolution-augmented transformer for speech recognition. In *Proc. Interspeech*, 2020.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 29:3451–3460, 2021.

Dongya Jia, Zhuo Chen, Jiawei Chen, Chenpeng Du, Jian Wu, Jian Cong, Xiaobin Zhuang, Chumin Li, Zhen Wei, Yuping Wang, et al. Ditar: Diffusion transformer autoregressive modeling for speech generation. *arXiv preprint arXiv:2502.03930*, 2025.

Wei Kang, Xiaoyu Yang, Zengwei Yao, Fangjun Kuang, Yifan Yang, Liyong Guo, Long Lin, and Daniel Povey. Libriheavy: A 50,000 hours ASR corpus with punctuation casing and context. In *Proc. ICASSP*, pp. 10991–10995, 2024.

Jaehyeon Kim, Keon Lee, Seungjun Chung, and Jaewoong Cho. Clam-tts: Improving neural codec language model for zero-shot text-to-speech. In *Proc. ICLR*, 2024.

John Kominek and Alan W Black. The CMU Arctic speech databases. In *Proc. SSW*, 2004.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. In *Proc. NeurIPS*, 2020.

Keon Lee. evaluate-zero-shot-tts, August 2024. URL `https://github.com/keonlee9420/evaluate-zero-shot-tts`.

Keon Lee, Dong Won Kim, Jaehyeon Kim, Seungjun Chung, and Jaewoong Cho. Ditto-tts: Diffusion transformers for scalable text-to-speech without domain-specific factors. In *Proc. ICLR*, 2025.

Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, and Ming Liu. Neural speech synthesis with transformer network. In *Proc. AAAI*, 2019.

Weiwei Lin and Chenghan He. Continuous autoregressive modeling with stochastic monotonic alignment for speech synthesis. *arXiv preprint arXiv:2502.01084*, 2025.

Zhijun Liu, Shuai Wang, Sho Inoue, Qibing Bai, and Haizhou Li. Autoregressive diffusion transformer for text-to-speech synthesis. *arXiv preprint arXiv:2406.05551*, 2024.

David JC MacKay. A practical bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472, 1992.

Lingwei Meng, Long Zhou, Shujie Liu, Sanyuan Chen, Bing Han, Shujie Hu, Yanqing Liu, Jinyu Li, Sheng Zhao, Xixin Wu, et al. Autoregressive speech synthesis without vector quantization. *arXiv preprint arXiv:2407.08551*, 2024.

Radford M Neal. *Bayesian leaning for neural networks*. University of Toronto, 1996.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An ASR corpus based on public domain audio books. In *Proc. ICASSP*, 2015.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proc. ICML*, 2023.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. Utmos: Utokyo-sarulab system for voicemos challenge 2022. *arXiv preprint arXiv:2204.02152*, 2022.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In *Proc. NeurIPS*, volume 31, 2018.

Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ-Skerrv Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions. In *Proc. ICASSP*, 2018.

Zhengyan Sheng, Zhihao Du, Shiliang Zhang, Zhijie Yan, Yexin Yang, and Zhenhua Ling. Syncspeech: Low-latency and efficient dual-stream text-to-speech based on temporal masked transformer. *arXiv preprint arXiv:2502.11094*, 2025.

SileroTeam. Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier. `https://github.com/snakers4/silero-vad`, 2024.

Haiyang Sun, Shujie Hu, Shujie Liu, Lingwei Meng, Hui Wang, Bing Han, Yifan Yang, Yanqing Liu, Sheng Zhao, Yan Lu, et al. Zero-shot streaming text to speech synthesis with transducer and auto-regressive modeling. *arXiv preprint arXiv:2505.19669*, 2025.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*, 2023.

Hui Wang, Shujie Liu, Lingwei Meng, Jinyu Li, Yifan Yang, Shiwan Zhao, Haiyang Sun, Yanqing Liu, Haoqin Sun, Jiaming Zhou, et al. Felle: Autoregressive speech synthesis with token-wise coarse-to-fine flow matching. *arXiv preprint arXiv:2502.11128*, 2025a.

Hui Wang, Yifan Yang, Shujie Liu, Jinyu Li, Lingwei Meng, Yanqing Liu, Jiaming Zhou, Haoqin Sun, Yan Lu, and Yong Qin. Streammel: Real-time zero-shot text-to-speech via interleaved continuous autoregressive modeling. *arXiv preprint arXiv:2506.12570*, 2025b.

Xinsheng Wang, Mingqi Jiang, Ziyang Ma, Ziyu Zhang, Songxiang Liu, Linqin Li, Zheng Liang, Qixi Zheng, Rui Wang, Xiaoqin Feng, et al. Spark-tts: An efficient llm-based text-to-speech model with single-stream decoupled speech tokens. *arXiv preprint arXiv:2503.01710*, 2025c.

Yuancheng Wang, Haoyue Zhan, Liwei Liu, Ruihong Zeng, Haotian Guo, Jiachen Zheng, Qiang Zhang, Xueyao Zhang, Shunsi Zhang, and Zhizheng Wu. Maskgct: Zero-shot text-to-speech with masked generative codec transformer. *arXiv preprint arXiv:2409.00750*, 2024.

Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*, 2017.

Detai Xin, Xu Tan, Kai Shen, Zeqian Ju, Dongchao Yang, Yuancheng Wang, Shinnosuke Takamichi, Hiroshi Saruwatari, Shujie Liu, Jinyu Li, et al. RALL-E: Robust codec language modeling with chain-of-thought prompting for text-to-speech synthesis. *arXiv preprint arXiv:2404.03204*, 2024.

Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92). 2019.

Yifan Yang, Ziyang Ma, Shujie Liu, Jinyu Li, Hui Wang, Lingwei Meng, Haiyang Sun, Yuzhe Liang, Ruiyang Xu, Yuxuan Hu, et al. Interleaved speech-text language models are simple streaming text to speech synthesizers. *arXiv preprint arXiv:2412.16102*, 2024.

Zhen Ye, Xinfa Zhu, Chi-Min Chan, Xinsheng Wang, Xu Tan, Jiahe Lei, Yi Peng, Haohe Liu, Yizhu Jin, Zheqi Dai, Hongzhan Lin, Jianyi Chen, Xingjian Du, Liumeng Xue, Yunlin Chen, Zhifei Li, Lei Xie, Qiuqiang Kong, Yike Guo, and Wei Xue. Llasa: Scaling train-time and inference-time compute for llama-based speech synthesis. *arXiv preprint arXiv:2502.04128*, 2025.

Guanlong Zhao, Sinem Sonsaat, Alif Silpachai, Ivana Lucic, Evgeny Chukharev-Hudilainen, John Levis, and Ricardo Gutierrez-Osuna. L2-ARCTIC: A Non-native English Speech Corpus. In *Proc. Interspeech*, 2018.

Xinfa Zhu, Wenjie Tian, and Lei Xie. Autoregressive speech synthesis with next-distribution prediction. *arXiv preprint arXiv:2412.16846*, 2024.

A LIMITATIONS

BELLE is trained and evaluated only on English speech, so its behavior on other languages and in cross-lingual synthesis is untested.

The official implementation of MELLE is not public, so every MELLE result in this paper comes from our re-implementation, trained with the same backbone, data, and training schedule as BELLE. Its absolute scores can therefore differ from those reported by Meng et al. (2024).

The comparison of realization weights (Appendix G.3) and the ablations of the Sampling Module and the flux loss (Appendices I.3 and I.4) are run on corpora built from a 72 h subset of LS (Appendix E.3). Whether the same rankings hold at the scale of LS+6TTS is not tested.

The largest training corpus in this paper is LS+6TTS with 4.8k h, whereas the larger TTS systems in Table 2 are trained on 44k to 250k h. How BELLE scales beyond this volume, for example when parallel speech is synthesized for a larger corpus with one recording per text, is not tested.

B BAYESIAN INTERPRETATION OF BELLE

Bayesian neural networks (BNNs) place priors over the network weights. BELLE follows evidential deep learning (EDL; Amini et al., 2020) instead: the network weights are point estimates, and the prior is placed over the parameters of the likelihood of each frame.

B.1 EMPIRICAL BAYES AND TYPE-II MAXIMUM LIKELIHOOD

For one mel band of one frame, the observation y follows a Gaussian with mean μ and variance σ^2. BELLE treats $\boldsymbol{\theta} = (\mu, \sigma^2)$ as a random variable and predicts the parameters $\boldsymbol{m} = (\gamma, \nu, \alpha, \beta)$ of an NIG prior over $\boldsymbol{\theta}$ from the input $\boldsymbol{x}$. Training maximizes the marginal likelihood of the observation,

$$\log p(y \mid \boldsymbol{m}) = \log \int p(y \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \boldsymbol{m})\, d\boldsymbol{\theta}, \tag{14}$$

in which $\boldsymbol{\theta}$ is integrated out in closed form (Appendix D). Fitting the parameters of a prior by maximizing the marginal likelihood is known as empirical Bayes, or type-II maximum likelihood. In contrast, a regression loss fits a point estimate of y, and the KL loss of MELLE fits one Gaussian per frame whose variance is pulled toward one (Eq. (6)).

B.2 SAMPLING AT GENERATION TIME

At generation time, each frame is sampled from the predicted NIG prior in three steps (Eq. (8)): a variance $\boldsymbol{\sigma}_t^2$ is drawn from the Inverse-Gamma distribution $\Gamma^{-1}(\boldsymbol{\alpha}_t, \boldsymbol{\beta}_t)$, a mean $\boldsymbol{\mu}_t$ from $\mathcal{N}(\boldsymbol{\gamma}_t, \boldsymbol{\sigma}_t^2/\boldsymbol{\nu}_t)$, and the frame $\boldsymbol{z}_t$ from $\mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\sigma}_t^2)$. The sample $\boldsymbol{z}_t$ therefore follows the Student-t distribution of Eq. (4), and its spread changes with the input through $(\boldsymbol{\nu}_t, \boldsymbol{\alpha}_t, \boldsymbol{\beta}_t)$. Replacing $\boldsymbol{z}_t$ with the location $\boldsymbol{\gamma}_t$ at inference raises WER-H (Appendix I.4), so the sampled spread affects the generated speech.

A model trained only with an L1 or L2 loss predicts the conditional median or mean of a frame, which averages over the valid realizations of a text. Sampling from the NIG prior instead draws one realization, with a larger spread at frames with fast spectral change (Sec. 7.2). Consistent with this, BELLE reaches a higher diversity than MELLE under all three training corpora (Table 1).

B.3 COMPARISON WITH GAUSSIAN SAMPLING

The Sampling Modules of BELLE and MELLE differ in the family of the predictive distribution and in how its variance is trained.

Predictive distribution. Integrating out the NIG prior gives a Student-t predictive distribution with 2α degrees of freedom (Amini et al., 2020). As $\alpha \to \infty$, the Student-t approaches a Gaussian, so the predictive distribution of BELLE contains the Gaussian of MELLE as a limiting case and can also represent heavier tails. Real repeated readings have tails heavier than a Gaussian once the small number of readings and the alignment are accounted for (Appendix I.2).

Variance. In the KL loss of MELLE (Eq. (6)), each variance term $\sigma^2_{t,d} - \log \sigma^2_{t,d}$ is minimized at $\sigma^2_{t,d} = 1$ for every frame, which pulls the model toward homoscedasticity, a variance that is the same for every input. The evidential loss has no term that pulls the variance toward a fixed value, so the variance is fit to the observed frames and can change with the input, which is known as heteroscedasticity. The predicted variances of the two models follow this difference: the coefficient of variation of σ_t within an utterance is 19% for BELLE and 6% for MELLE (Sec. 7.2).

C Modeling Temporal and Spectral Dependencies

The NIG prior of Eq. (3) is placed on each frame and acts on each mel band separately, whereas neighboring mel frames are strongly correlated, and so are neighboring bands within a frame. Correlations across frames are modeled exactly by the AR factorization, and the only independence assumption is that the bands of one frame are independent given its hidden representation.

Across frames. BELLE does not treat frames as independent. The factorization of Eq. (1) is the chain rule of probability and holds for any joint distribution of the frames, so the NIG prior of frame t models the conditional distribution $p(\boldsymbol{y}_t \mid \boldsymbol{y}_{<t}, \boldsymbol{x})$ rather than a marginal distribution of the frame. The AR LM computes $\boldsymbol{e}_t$ from the text and all previous frames (Sec. 3.2), and the NIG parameters of frame t are a function of $\boldsymbol{e}_t$. A change in any previous frame therefore changes both the location $\boldsymbol{\gamma}_t$ and the spread of frame t. At inference, each generated frame is fed back as input to the next step, so the dependence of every frame on the frames before it is kept during generation.

Across bands. Within frame t, the D bands are modeled as independent given $\boldsymbol{e}_t$, and correlations across bands are still captured in two ways. First, all bands share $\boldsymbol{e}_t$, and their NIG parameters are predicted together by one linear layer, so the locations $\boldsymbol{\gamma}_t$ of all bands, which carry the spectral envelope of the frame, are predicted jointly. The independence applies only to the random deviation of each band around its location. Second, this deviation passes through two networks that act on all bands together before the output is formed: the denoising MLP maps all D bands of $\boldsymbol{z}_t$ to the frame $\boldsymbol{y}_t^{(1)}$, and the Postnet convolves over neighboring frames with all bands as input channels. Both networks are trained with the regression loss of Eq. (11) against the ground-truth frames, so they can remove deviations that are inconsistent across bands.

Why the bands are not modeled jointly. A joint prior over the D bands would require a full covariance matrix per frame, with $D(D+1)/2 = 3{,}240$ variance parameters for $D = 80$ instead of 80, and sampling from its conjugate prior, the Normal-Inverse-Wishart distribution, requires drawing an 80×80 covariance matrix and factorizing it at every frame. The factorization over bands keeps the Sampling Module to one linear layer and the sampling of Eq. (8) to element-wise operations. MELLE makes the same assumption with its diagonal Gaussian (Eq. (5)), so the comparison between BELLE and MELLE differs in the prior over each band, not in how the bands are coupled.

D Derivation of the Evidential Loss

We drop the indices t and d and consider one mel band of one frame. The observation y and its NIG prior are

$$y \sim \mathcal{N}(\mu, \sigma^2), \quad \mu \sim \mathcal{N}\left(\gamma, \frac{\sigma^2}{\nu}\right), \quad \sigma^2 \sim \Gamma^{-1}(\alpha, \beta), \tag{15}$$

where $\Gamma^{-1}(\alpha, \beta)$ is the Inverse-Gamma distribution, and the parameters $\boldsymbol{m} = (\gamma, \nu, \alpha, \beta)$ satisfy $\gamma \in \mathbb{R}$, $\nu > 0$, $\alpha > 1$, and $\beta > 0$. The joint density of (μ, σ^2) is the NIG density

$$p(\mu, \sigma^2 \mid \boldsymbol{m}) = \frac{\beta^\alpha \sqrt{\nu}}{\Gamma(\alpha)\sqrt{2\pi\sigma^2}} \left(\frac{1}{\sigma^2}\right)^{\alpha+1} \exp\left\{-\frac{2\beta + \nu(\gamma - \mu)^2}{2\sigma^2}\right\}, \tag{16}$$

where $\Gamma(\cdot)$ is the gamma function. The marginal likelihood of y integrates out both parameters:

$$p(y \mid \boldsymbol{m}) = \int_0^\infty \int_{-\infty}^\infty p(y \mid \mu, \sigma^2)\, p(\mu, \sigma^2 \mid \boldsymbol{m})\, \mathrm{d}\mu\, \mathrm{d}\sigma^2. \tag{17}$$

Integrating out μ first gives $y \mid \sigma^2 \sim \mathcal{N}(\gamma, \sigma^2(1+\nu)/\nu)$. Integrating out σ^2 under the Inverse-Gamma prior then gives a Student-t distribution,

$$p(y \mid \boldsymbol{m}) = \mathrm{St}\left(y;\, \gamma,\, \frac{\beta(1+\nu)}{\nu\alpha},\, 2\alpha\right), \tag{18}$$

with location γ, squared scale $\beta(1+\nu)/(\nu\alpha)$, and 2α degrees of freedom. The negative logarithm of this density is $\mathcal{L}_{\text{NLL}}$ of Eq. (9), with $\Omega = 2\beta(1+\nu)$.

The regularizer $\mathcal{L}_{\text{R}} = |y-\gamma|\,(2\nu+\alpha)$ follows Amini et al. (2020). The quantity $2\nu+\alpha$ is the evidence, the number of virtual observations that the prior represents, and multiplying it by the error lowers the confidence of the prior on frames that are predicted poorly. The two terms are combined with the weight λ in Eq. (9).

E Model and Training Details

E.1 Model Configuration

The Prenet is a 3-layer network with a dropout rate of 0.5, which is kept active at inference as in Tacotron (Wang et al., 2017). The AR LM is a decoder-only Transformer with 12 blocks, 16 attention heads, a hidden size of 1024, a feed-forward dimension of 4096, and a dropout rate of 0.1[2]. The Sampling Module is one linear layer that outputs the NIG parameters, and the denoising MLP has three residual layers. The Postnet consists of five convolutional blocks with a kernel size of 5 and 256 channels. Waveforms are synthesized with a pretrained HiFi-GAN vocoder (Kong et al., 2020)[3].

E.2 Training of the Main Models

BELLE and MELLE are trained on LS, LS+6TTS, and the ParallelVox training set with the weighted loss of Eq. (13). On LS+6TTS, the human recording of a text is weighted 0.22 and each teacher realization 0.13 in Eq. (13) (Appendix G.3); on ParallelVox, all recordings of a text are weighted equally; on LS, each text has one recording. Training uses AdamW with a batch of about 160K frames on 16 NVIDIA A800 GPUs for 450K updates. The learning rate rises linearly to 5×10^{-4} over the first 10% of the updates and then decays linearly to zero. The models trained on ParallelVox and on LS+6TTS-881h share a budget of about 440K updates (Sec. 6.1). For BELLE, $\lambda = 0.5$ in Eq. (9), and $\lambda_{\text{samp}} = 0.2$ and $\lambda_{\text{flux}} = 0.5$ in Eq. (10). For MELLE, $\lambda_{\text{samp}} = 0.1$ and $\lambda_{\text{flux}} = 0.5$, as in the original paper (Meng et al., 2024).

BELLE-stream is initialized from BELLE *LS+6TTS* and trained on LS+6TTS with a batch of 80K frames on 8 NVIDIA A800 GPUs for 150K updates, with $\lambda_{\text{flux}} = 0.1$ and all other hyperparameters as in BELLE. A text chunk contains $S_{\text{text}} = 20$ phonemes after G2P conversion, and an audio chunk contains $S_{\text{audio}} = 50$ mel frames, about 0.8 s. Utterances whose ratio of mel frames to phonemes is below 2.5 are removed from training, so that each text chunk has enough audio frames to be generated from.

VALL-E (Wang et al., 2023) is trained from scratch on the ParallelVox training set as a discrete-token baseline. The recordings of a text are treated as independent text–recording pairs, so ParallelVox is used as a standard TTS corpus.

E.3 Training of the Analysis Models

The weight comparison of Appendix G.3 and the ablations of Appendices I.3 and I.4 use smaller corpora built from the LibriSpeech `train-clean-100` subset. Processing this subset as LS (Appendix G.2) gives 72 h of speech, denoted *LS-100*. *LS-100+2TTS* adds the realizations of XTTS-v2 and MaskGCT, for 218 h in total, and *LS-100+6TTS* adds those of all six teachers, for 455 h in total. Both are trained with the weighted loss of Eq. (13), a batch of about 80K frames, and 47K updates.

[2] Our code is modified from `https://github.com/lifeiteng/vall-e`.

[3] `https://huggingface.co/mechanicalsea/speecht5-tts`

F Parallel Realizations under the AR Factorization

BELLE models the distribution of each frame given the text and the previous frames (Eq. (1)), and during training each realization is teacher-forced on its own previous frames. The N realizations of a text $\boldsymbol{x}$ therefore give one observation of each of N different conditional distributions $p(\boldsymbol{y}_t \mid \boldsymbol{y}^{(i)}_{<t}, \boldsymbol{x})$, not N observations of one conditional distribution. They are nonetheless N observations of the same distribution $p(\boldsymbol{y} \mid \boldsymbol{x})$, and training with Eq. (13) maximizes a weighted likelihood of these observations.

Inputs at training and inference. During training, the input of BELLE is the text of an utterance and the target is its mel-spectrogram; no prompt is given. The prompts used by the teachers to build LS+6TTS (Appendix G.2) only determine which voice each teacher synthesizes and are never seen by BELLE, so every realization of a text, human or synthetic, is a sample of $p(\boldsymbol{y} \mid \boldsymbol{x})$, which includes variation in speaker and prosody. At inference, the prompt text is concatenated with the input text, the prompt speech is given as the first frames, and the model continues generation from it. The prompt speech therefore acts as the prefix $\boldsymbol{y}_{<t}$ and sets the voice of the generated speech.

The loss decomposes over realizations. Under the model, the realizations of a text are independent given $\boldsymbol{x}$, because every frame of every realization draws its own mean and variance from the NIG prior (Eq. (8)). The log-likelihood of the N realizations is therefore

$$\log p_\theta\big(\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(N)} \mid \boldsymbol{x}\big) = \sum\nolimits_{i=1}^{N} \log p_\theta\big(\boldsymbol{y}^{(i)} \mid \boldsymbol{x}\big) = \sum\nolimits_{i=1}^{N} \sum\nolimits_{t=1}^{T_i} \log p_\theta\big(\boldsymbol{y}^{(i)}_t \mid \boldsymbol{y}^{(i)}_{<t}, \boldsymbol{x}\big), \tag{19}$$

where T_i is the number of frames of realization i and each term is the Student-t likelihood of Eq. (4). Computing $\mathcal{L}_{\text{NLL}}$ on all realizations of a text jointly and summing the losses computed on each realization separately thus give the same objective, and the regularizer $\mathcal{L}_{\text{R}}$ is also a sum over realizations. Eq. (13) is this sum with the weights w_i, and with equal weights it is the maximum-likelihood objective for N observations of $p(\boldsymbol{y} \mid \boldsymbol{x})$. If all realizations of a text instead shared one draw of the variance, the joint likelihood would not factorize in this way.

What the realizations estimate. Let $p_{\text{data}}(\boldsymbol{y} \mid \boldsymbol{x})$ be the distribution from which the realizations of $\boldsymbol{x}$ are drawn. As the number of realizations grows, the average of this objective approaches the expected log-likelihood under p_{data}, and maximizing it minimizes

$$D_{\text{KL}}\big(p_{\text{data}}(\boldsymbol{y} \mid \boldsymbol{x}) \,\|\, p_\theta(\boldsymbol{y} \mid \boldsymbol{x})\big) = \sum\nolimits_t \mathbb{E}_{\boldsymbol{y}_{<t} \sim p_{\text{data}}}\Big[D_{\text{KL}}\big(p_{\text{data}}(\boldsymbol{y}_t \mid \boldsymbol{y}_{<t}, \boldsymbol{x}) \,\|\, p_\theta(\boldsymbol{y}_t \mid \boldsymbol{y}_{<t}, \boldsymbol{x})\big)\Big] \tag{20}$$

by the chain rule of the KL divergence. The conditional distribution of each frame, including its spread, is therefore fit on average over the prefixes that realizations of $\boldsymbol{x}$ produce. One recording provides one such prefix per text, whereas N realizations provide N prefixes that share the linguistic content and differ in voice and prosody. The AR LM computes $\boldsymbol{e}_t$ with the same parameters for every prefix, so realizations with similar prefixes inform the same region of the conditional distribution, in the same way that an AR language model learns next-token distributions although most of its prefixes occur only once in the training data. At the first frame, all realizations share the context $\boldsymbol{x}$ and are N observations of the same conditional distribution $p(\boldsymbol{y}_1 \mid \boldsymbol{x})$.

Empirical check. The second analysis of Sec. 7.2 teacher-forces BELLE *ParallelVox* on one reading of each sentence and compares its predicted variance with the variance across the repeated readings at the same frames. The two are positively correlated (median Spearman $\rho = 0.18$), also after spectral flux, energy, and alignment irregularity are controlled for (median partial $\rho = 0.17$), and the correlation vanishes when the predicted variance of one sentence is paired with the readings of another (Appendix I.2). The conditional spread that BELLE learns from realizations with different prefixes thus follows the spread across real readings of a sentence.

G Parallel Speech Corpora

G.1 ParallelVox

ParallelVox merges five public English corpora in which the same text is recorded more than once, by different speakers or by one speaker on different occasions (Table 4). All audio is downmixed

to mono, resampled to 16 kHz, and trimmed with Silero VAD (SileroTeam, 2024), and each text is converted to phonemes with the espeak-ng G2P backend, which uses 78 phoneme symbols. Each text is stored once with all of its recordings, and in training all recordings of a text enter Eq. (13) with equal weights. A text has 2 to 91 recordings in the training set and 2 to 18 in the test set.

Table 4: Source corpora of ParallelVox after VAD trimming and before WER filtering.

Source	Recordings	Hours	Speakers	License
LibriHeavy (Kang et al., 2024)	312,959	1,017.7	3,392	CC BY 4.0
Parallel Audiobook Corpus	84,791	134.4	53	Public Domain
VCTK 0.92 (Yamagishi et al., 2019)	34,981	20.7	109	CC BY 4.0
CMU ARCTIC (Kominek & Black, 2004) + L2-Arctic (Zhao et al., 2018)	18,625	15.5	42	Mixed
Common Voice 17.0 (Ardila et al., 2020), parallel subset	284	0.2	255	CC-0
Total	**451,640**	**1,188.5**	**3,851**	

Quality filtering. Each recording is transcribed with Whisper-large-v2 (Radford et al., 2023) using greedy decoding and scored with UTMOS22-Strong (Saeki et al., 2022). A recording is dropped when its WER or MOS fails the threshold, and a text is kept only if at least two of its recordings remain, since parallel speech needs at least two realizations of a text. ParallelVox uses only the filter WER $\leq 10\%$, which keeps 123,640 texts, 344,303 recordings, and 903 h. Table 5 shows how stricter WER or MOS thresholds reduce the corpus.

Table 5: Size of ParallelVox under different WER and MOS thresholds. The first row is the setting used in this paper.

WER $\leq$	MOS $\geq$	Texts	Recordings	Hours
10%	—	123,640	344,303	903.0
10%	3.8	60,012	175,356	400.1
10%	4.0	38,618	114,741	244.4
5%	—	92,734	262,337	638.5
3%	—	73,683	214,831	474.9

Train and test split. Texts whose recordings all come from one speaker form the single-speaker repeat test set: 3,306 texts, 6,831 recordings, 22 h, and 250 speakers. The remaining 120,334 texts, each recorded by at least two speakers, form the training set: 337,472 recordings, 881 h, and 3,406 speakers. The test set records how one speaker varies across repeated readings of the same text, which makes it the human reference for the diversity metric (Appendix H.3).

G.2 LS+6TTS

LS is built from the LibriSpeech training set (Panayotov et al., 2015). Long silences are removed with Silero VAD (SileroTeam, 2024), and utterances of 0.5 to 14 s are kept, which gives 706 h. Audio is resampled to 16 kHz and converted to 80-band mel-spectrograms, and transcripts are converted to phonemes with espeak-ng[4].

Six pretrained TTS teachers synthesize every text of LS: CosyVoice2 (Du et al., 2024), an LLM with a flow-matching decoder; IndexTTS (Deng et al., 2025), an AR model over discrete acoustic tokens; SparkTTS (Wang et al., 2025c), a single-stage LLM-based model; F5TTS (Chen et al., 2024b), a flow-matching model; MaskGCT (Wang et al., 2024), a masked generative non-autoregressive model; and XTTS-v2 (Casanova et al., 2024), a multilingual model. Each teacher is prompted with a different utterance of the same speaker, drawn at random from the LS recordings of that speaker with a duration of 3 to 8 s, and its output is resampled to 16 kHz. With the human recording, each text has seven realizations, and the corpus totals 4,817 h.

[4] https://github.com/espeak-ng/espeak-ng

G.3 Weights of Human and Synthetic Realizations

In Eq. (13), the human recording of an LS+6TTS text is weighted 0.22 and each of the six teacher realizations 0.13, so the weights sum to one and the human recording counts about 1.7 times as much as any single teacher. Table 6 compares this choice with three other allocations that also sum to one, on *LS-100+6TTS* (Appendix E.3). The weights 0.22/0.13 give the lowest WER-H in both conditions, whereas 0.16/0.14 gives the highest SIM-o in both. Raising the human weight above 0.22 worsens all four metrics. The SIM-o differences between 0.16/0.14 and 0.22/0.13 are at most 0.006, whereas the WER-H differences are 0.07 and 0.44, so 0.22/0.13 is used.

Table 6: Weights of the human recording and of each teacher realization in Eq. (13), on *LS-100+6TTS*. The second row is the setting used in this paper.

Human	**Each teacher**	**Continuation**		**Cross-Sentence**	
		WER-H ↓	SIM-o ↑	WER-H ↓	SIM-o ↑
0.16	0.14	2.64	**0.415**	5.17	**0.453**
0.22	0.13	**2.57**	0.409	**4.73**	0.449
0.28	0.12	2.78	0.406	5.47	0.444
0.34	0.11	2.76	0.405	5.47	0.446

H Evaluation Details

H.1 Objective Evaluation

All models are evaluated on the same set of LibriSpeech test-clean utterances of 4 to 10 s, following the protocol of VALL-E (Wang et al., 2023) as implemented by Lee (2024). Intelligibility is measured with two ASR models, a Conformer-Transducer[5] (Gulati et al., 2020) and a fine-tuned HuBERT-Large[6] (Hsu et al., 2021), whose WERs are denoted WER-C and WER-H. Speaker similarity is the cosine similarity between WavLM-TDCNN[7] (Chen et al., 2022) speaker embeddings of the generated speech and of the prompt: SIM-o uses the original prompt, and SIM-r uses the prompt reconstructed by the vocoder. The main text reports WER-H and SIM-o; WER-C and SIM-r appear in Table 9. BELLE-stream is evaluated only under the Cross-Sentence condition.

H.2 Subjective Evaluation

MOS and SMOS are collected for Ground Truth, MELLE *LS+6TTS*, and BELLE *LS+6TTS*. One Cross-Sentence sample is selected per speaker, which gives 40 samples per system, and all samples are resampled to 16 kHz. For MOS, raters score quality and naturalness on a 1-to-5 scale and are told to ignore differences in style, such as timbre, emotion, and prosody. For SMOS, each sample is paired with the reference utterance, and raters score how similar the speaker identity (timbre) is on a 1-to-5 scale while ignoring differences in content, grammar, and audio quality. The order of the samples is shuffled for each rater. Fifteen native speakers rated the samples on Amazon Mechanical Turk, and every rater scored both BELLE and MELLE. Figure 3 shows the instructions. The participants received about $100 in total. Part of the samples can be heard at `https://anonymous-pages229.github.io/BELLE-demo/`.

H.3 Diversity Metric

Diversity is measured on the single-speaker repeat test set of ParallelVox (Appendix G.1). Of its 3,306 texts, 3,270 have a Cross-Sentence prompt, a different utterance of the same speaker of 3 to 8 s, and are used. Each system synthesizes each text three times from the same prompt, so the three outputs of a group differ only through sampling. Each waveform is passed through WavLM-Large, and the

[5] `https://huggingface.co/nvidia/stt_en_conformer_transducer_xlarge`

[6] `https://huggingface.co/facebook/hubert-large-ls960-ft`

[7] `https://github.com/microsoft/UniSpeech/tree/main/downstreams/speaker_verification`



(a) Speaker similarity (SMOS) (b) Speech quality (MOS)

Figure 3: Instructions shown to the raters in the subjective evaluation.

hidden states of layer 13 (of 24) are mean-pooled over time, excluding padding, and L2-normalized; Appendix H.4 explains this choice. The distance between two outputs a and b is $1 - |\cos(a, b)|$, where the absolute value keeps two embeddings that point in opposite directions from counting as maximally different. The diversity of a group is the mean distance over all pairs in it, and Cos. is the mean over groups, reported $\times 10^{-4}$ in Table 1. The Ground Truth value applies the same computation to the human readings of each text.

Generated samples with a WER above 5% are dropped before the distances are computed, since the distance of an unintelligible sample reflects generation errors rather than acoustic variation, and a group with fewer than two remaining samples has no pair to compare and is left out of the mean.

H.4 Validity of the Diversity Metric

The choice of layer follows the probing analysis of Chiu et al. (2025), which finds that early layers of self-supervised speech models encode basic acoustics, that middle layers combine them into more abstract traits, and that intermediate representations track prosody better than dedicated speaker embeddings. Layer 13 of WavLM-Large is therefore used instead of a speaker-embedding model.

To test whether the metric reflects how differently a sentence is spoken, five prosodic quantities are measured on every sample: duration, median F0, F0 standard deviation, mean energy, and energy standard deviation. For each group, the variation of each quantity across the samples is correlated with the WavLM distance of the group. Table 7 reports Spearman's ρ over the groups of each group source, and a permutation test gives $p \leq 0.004$ in every cell.

Table 7: Spearman's ρ between the within-group variation of prosodic quantities and the within-group WavLM distance.

Group source	Duration	Median F0	F0 s.d.	Energy	Energy s.d.
Ground Truth (human readings)	**0.54**	**0.70**	**0.61**	0.21	**0.65**
BELLE *ParallelVox*	0.07	0.37	0.25	**0.34**	0.33
BELLE *LS+6TTS*	0.19	0.38	0.25	0.29	0.29

All 15 correlations are positive, and on four of the five quantities the correlation is strongest on the human readings, where the speaker and the text are fixed and only the reading varies. The metric therefore tracks how prosody varies across repeated readings of one sentence, which is the variation that diversity is meant to measure.

I Additional Results

I.1 Statistical Significance

BELLE and MELLE are scored on the same test items, so each objective difference is tested with a paired bootstrap over items, namely sentences for WER-H and SIM-o and prompt groups for Cos., and a paired Wilcoxon test on the same items gives the same conclusions. MOS and SMOS are tested with a paired Wilcoxon test over the rated samples, since every rater scored both systems. On LS, all five objective differences favor BELLE with $p \leq 0.004$, and on LS+6TTS, all seven differences favor BELLE with $p < 0.05$. On ParallelVox, the two SIM-o differences and the diversity difference favor BELLE with $p \leq 0.007$, whereas the two WER-H differences are not significant (Table 8).

Table 8: BELLE − MELLE differences and p-values, from a paired bootstrap for the objective metrics and a paired Wilcoxon test for MOS and SMOS. MOS and SMOS were collected only on LS+6TTS.

Metric	ParallelVox	LS	LS+6TTS
Continuation WER-H	-0.07 ($p = 0.52$)	-0.39 ($p < 0.001$)	-0.46 ($p < 0.001$)
Continuation SIM-o	$+0.044$ ($p < 0.001$)	$+0.005$ ($p = 0.004$)	$+0.031$ ($p < 0.001$)
Cross-Sentence WER-H	$+0.13$ ($p = 0.56$)	-3.75 ($p < 0.001$)	-0.84 ($p < 0.001$)
Cross-Sentence SIM-o	$+0.032$ ($p < 0.001$)	$+0.015$ ($p < 0.001$)	$+0.036$ ($p < 0.001$)
Diversity (Cos.)	$+1.67$ ($p = 0.007$)	$+6.90$ ($p < 0.001$)	$+2.60$ ($p < 0.001$)
MOS	–	–	$+0.19$ ($p = 0.021$)
SMOS	–	–	$+0.33$ ($p < 0.001$)

I.2 Details of the Variance Analyses

Predicted variance. For the first analysis of Sec. 7.2, the per-frame σ_t of BELLE *LS+6TTS* and MELLE *LS+6TTS* is exported during inference on 196 Cross-Sentence utterances of LibriSpeech test-clean. For BELLE, the variance is the analytic variance of the Student-t predictive distribution, $\beta(1+\nu)/((\alpha-1)\nu)$, computed from the predicted NIG parameters rather than from samples; for MELLE, it is $\exp(\log \sigma^2)$ from the predicted log-variance. The per-frame σ_t is the square root of this variance averaged over the 80 mel bands. The CV within an utterance is the standard deviation of σ_t over the frames of the utterance divided by their mean, averaged over utterances, and the CV across utterances is computed over the per-utterance means of σ_t.

The spectral change of a frame is its spectral flux, the absolute difference between consecutive log-mel frames averaged over the 80 bands. Within each utterance, the 20% of frames with the highest flux form the fastest group, the 20% with the lowest flux form the slowest group, and the ratio of their mean σ_t is computed. Frames with fast spectral change, such as onsets and offsets, are often quieter than steady frames, and quiet frames have noisier log-mel values, so a larger σ_t at these frames could reflect low energy rather than spectral change. The comparison is therefore made among frames of similar energy: the energy of a frame is the mean of its 80 log-mel values, the frames of each utterance are split into five equal-frequency energy bins, the ratio is computed within each bin and averaged over bins, and a bin is skipped when either group has fewer than three frames. The per-utterance ratios are averaged over utterances, with bootstrap confidence intervals. With the ground-truth recordings as the AR input (teacher forcing), the increase is 12% for BELLE and 4% for MELLE.

Alignment of readings. The readings of a text are aligned on their mel-spectrograms by DTW. The per-band mean of each reading is first subtracted, which removes offsets in timbre and level that would otherwise distort the alignment, since the DTW cost is the mean absolute mel difference. The medoid of the group, the reading with the smallest summed DTW distance to the others, provides the reference time axis, and every other reading is aligned to it under a Sakoe–Chiba band of 25% of the length and at least 10 frames. The variance across readings is then computed at every frame of the medoid.

Correlation with the variance across readings. BELLE *ParallelVox* and MELLE *ParallelVox* are teacher-forced on the medoid of each group of the single-speaker repeat test set, and their predicted σ_t^2 is compared with the variance across readings at the same frames. The analysis uses the 2,185 groups whose readings are all at most 14 s long, the longest utterance length used in training. For each group, the Spearman correlation is computed over frames, and its significance is assessed against the correlations obtained under all circular shifts of the predicted variance, which keep the autocorrelation of both sequences, with a false discovery rate correction over groups. As a control, the predicted variance of one sentence is paired with the readings of another sentence. The partial correlation removes from both variables the ranks of spectral flux, energy, and alignment irregularity, defined as the fraction of alignment paths that enter a medoid frame through a non-diagonal step, averaged over readings and smoothed over 5 frames. The local DTW cost is not used as a control, because it is computed from the same frame differences as the variance across readings. The amplitude of the response is the slope of $\log \sigma_t^2$ on the logarithm of the variance across readings within each group, which does not depend on the overall scale of σ_t^2.

For BELLE, the median correlation is 0.18 (95% CI [0.17, 0.18]); it is positive in 88% of the groups and significant in 57%. The median partial correlation is 0.17 (95% CI [0.16, 0.17]); it is positive in 91% of the groups and significant in 59%. Within bins of spectral flux and energy, the median correlation is 0.16. On the 92 groups with at least three readings, the median correlation is 0.19 and the median partial correlation 0.20. In the mismatched control, the median correlation is 0.01 and no group is significant, and spectral flux itself has a median correlation of -0.06 with the variance across readings. The median slope is 0.069 for BELLE and 0.024 for MELLE, with a paired difference of 0.044 (95% CI [0.042, 0.046]) that is positive in 89% of the groups.

Properties of real speech tested in Sec. 7.3. These statistics use the aligned human readings alone, without any model. For each text, the standard deviation of a frame is the square root of the variance across readings averaged over the 80 bands, as for σ_t in Sec. 7.2, and the ratio of its 99th to its 1st percentile over the frames of the text is computed. The median ratio is 3.2 (interquartile range [2.7, 3.6]) over the 114 texts with at least three readings and 2.5 over the 10 texts with at least six readings. If the variance were the same at every frame, sampling noise alone would give median ratios of 1.30 and 1.15.

The reliability of the per-frame variance is measured with a shrinkage estimator, which writes the variance of a frame as a weighted average of the global variance and the empirical variance of the frame across readings, with the weight chosen to minimize the mean squared error. With three readings of a text, the best weight puts about 79% on the global variance and 21% on the empirical variance of the frame. Because this weight equals the share of sampling noise in the variation of the empirical variance across frames, about 79% of this variation is sampling noise, and this share decreases as the number of readings grows. The Inverse-Gamma prior gives an estimate of the same form: for n observations with a known mean and the prior $\sigma^2 \sim \Gamma^{-1}(\alpha, \beta)$, the posterior mean of σ^2 is $w\,\beta/(\alpha-1) + (1-w)\,s^2$, where s^2 is the empirical variance and $w = (\alpha-1)/(\alpha-1+n/2)$, so α sets how far the estimate moves toward the prior mean.

For the tails, each observation is compared with the other readings of its frame. For a text with n readings, its deviation from the mean of the other readings, divided by their standard deviation and by $\sqrt{1+1/(n-1)}$, follows a Student-t distribution with $n-2$ degrees of freedom when the readings are Gaussian. Passing this value through the cumulative distribution function of that Student-t and then through the inverse standard normal gives a value that is standard normal under a Gaussian, whatever n is. On the 114 texts with at least three readings, the probabilities that this value exceeds 2.58 and 3.29 in absolute value are 1.43 (95% CI [1.26, 1.59]) and 1.50 (95% CI [1.33, 1.67]) times their nominal values of 1% and 0.1%. To account for the alignment, Gaussian readings are generated around the medoid, warped along the real alignment paths, and realigned by DTW; whether their per-frame variance is the empirical variance of the data, a smoothed version of it, or a shrunk version of it, the two ratios stay between 1.01 and 1.04. The excess on real readings remains on the half of the frames with the lowest spectral flux, with ratios of 1.55 and 1.65, and on the texts with at least six readings, with ratios of 1.22 and 1.61.

For the held-out log-likelihood, the 114 texts of the single-speaker repeat test set with at least three readings are split by text into two halves of 57. Each observation, one mel band of one frame of one reading, is predicted from the other readings of its text, and all three densities use the mean of these readings as their location. The first Gaussian uses one global variance for all observations. The second uses the variance of the other readings at the same frame and band, pulled toward the global variance with a weight of 0.30 on the local variance. The Student-t uses a scale derived from a shrunk local variance and a fitted degree of freedom. The global variance, the weights, the scale factor, and the degree of freedom are fit on one half, and the log-likelihood is computed on the other half. The second Gaussian gains 0.033 nats per observation over the first, and the Student-t gains another 0.066 nats, which sums to 7.9 nats per frame over its 80 bands.

I.3 Ablation of the Sampling Module and the Flux Loss

Table 9 removes either the Sampling Module or the flux loss from BELLE *LS-100+2TTS* (Appendix E.3). Removing the Sampling Module raises WER-H from 2.66 to 5.00 in Continuation and from 4.84 to 13.07 in Cross-Sentence, whereas removing the flux loss raises it to 3.25 and 6.37. All eight metrics follow the same order, with the full model best and the model without the Sampling

Module worst, so both components help and the Sampling Module accounts for the larger share of the gain.

Table 9: Ablation of the Sampling Module and the flux loss on BELLE *LS-100+2TTS*. ✓ indicates enabled, and ✗ indicates disabled.

Sampling	Flux loss	Continuation				Cross-Sentence			
		WER-C	WER-H	SIM-r	SIM-o	WER-C	WER-H	SIM-r	SIM-o
✗	✓	4.30	5.00	0.382	0.344	12.04	13.07	0.392	0.351
✓	✗	2.50	3.25	0.388	0.347	5.61	6.37	0.417	0.372
✓	✓	**2.04**	**2.66**	**0.398**	**0.362**	**4.12**	**4.84**	**0.433**	**0.395**

I.4 Free-Variance Gaussian Baseline and Inference-Time Sampling

The Sampling Module of MELLE already predicts a per-frame variance (Eq. (5)), and only the KL loss of Eq. (6) pulls this variance toward one. To test whether a free variance alone accounts for the gain of BELLE, the KL loss is replaced by the Gaussian NLL

$$\mathcal{L}_{\text{GNLL}} = \sum_{t=1}^{T} \sum_{d=1}^{D} \left(\frac{(y_{t,d} - \mu_{t,d})^2}{2\sigma_{t,d}^2} + \frac{1}{2} \log \sigma_{t,d}^2 \right), \tag{21}$$

which has no term toward a fixed value and no prior over the variance. The four systems in Table 10 are trained on *LS-100+2TTS* and differ only in the Sampling Module and its loss; the first and last rows are taken from Table 9.

Table 10: Sampling Modules and their losses on *LS-100+2TTS*.

Sampling Module and loss	Continuation WER-H↓	Cross-Sentence WER-H↓
None	5.00	13.07
Gaussian, KL toward unit variance (MELLE)	3.24	6.83
Gaussian, NLL with free variance	3.88	7.31
NIG, evidential loss (BELLE)	**2.66**	**4.84**

The free-variance model trails MELLE in both conditions and trails BELLE by 2.5 points of Cross-Sentence WER-H, so a free variance alone does not account for the gain. Two properties of $\mathcal{L}_{\text{GNLL}}$ are consistent with this result. Its gradient with respect to $\mu_{t,d}$ is scaled by $1/\sigma_{t,d}^2$, so frames that receive a large variance contribute little to learning the mean. It also fits the variance of each frame to the few realizations of a text: *LS-100+2TTS* has three realizations per text, and on real speech with three readings per text, most of the frame-to-frame variation of such a variance is sampling noise (Appendix I.2). The Inverse-Gamma prior of BELLE instead pulls this estimate toward the prior mean (Sec. 7.3).

The sampled spread is also used at generation time. For BELLE *LS-100+2TTS*, replacing the sample $\boldsymbol{z}_t$ with the location $\boldsymbol{\gamma}_t$ at inference raises Cross-Sentence WER-H from 4.84 to 6.16.

J Streaming Generation

Why streaming. In a spoken dialogue system, the text to be spoken is often produced incrementally, for example by an LLM that generates one token at a time. A non-streaming TTS model starts synthesis only after the whole sentence is available, so the delay before the first audio grows with the length of the sentence. A streaming model starts from the first text chunk, so the first audio is available after a fixed delay.

Why AR models suit streaming. An AR model generates speech in time order, and each frame is conditioned only on the previous frames and the text (Eq. (1)). A generated frame therefore never depends on later frames, and streaming only requires conditioning each audio chunk on the text

received so far (Eq. (2)). Models that generate all frames of an utterance jointly, such as diffusion and flow-matching TTS, must be modified to operate on chunks before they can stream; CosyVoice 2, for example, uses a chunk-aware flow-matching decoder (Du et al., 2024).

Why BELLE suits streaming. BELLE generates mel-spectrogram frames directly, so each frame can be passed to the vocoder without an intermediate stage that decodes discrete tokens into acoustic features. Its Sampling Module predicts the NIG parameters of a frame with one linear layer and samples the frame in closed form (Eq. (8)), so each frame costs one forward pass of the AR LM, the Sampling Module, and the denoising MLP. In contrast, continuous-valued AR models with diffusion or flow-matching heads (Liu et al., 2024; Wang et al., 2025a; Jia et al., 2025) run several denoising steps for every frame. BELLE-stream therefore uses the same Sampling Module and the same evidential loss as BELLE, and only the input sequence changes.

Implementation. BELLE-stream interleaves text chunks and audio chunks as the input of the AR LM under a causal mask (Sec. 3.2), following the interleaving scheme of IST-LM (Yang et al., 2024). Its chunk sizes and training settings are given in Appendix E.2, and its results are in Sec. 7.4.