LP-CFM: PERCEPTUAL INVARIANCE-AWARE CONDITIONAL FLOW MATCHING FOR SPEECH MODELING

Doyeop Kwak, Youngjoon Jang, Joon Son Chung

Korea Advanced Institute of Science and Technology, South Korea

ABSTRACT

The goal of this paper is to provide a new perspective on speech modeling by incorporating perceptual invariances such as amplitude scaling and temporal shifts. Conventional generative formulations often treat each dataset sample as a fixed representative of the target distribution. From a generative standpoint, however, such samples are only one among many perceptually equivalent variants within the true speech distribution. To address this, we propose Linear Projection Conditional Flow Matching (LP-CFM), which models targets as projection-aligned elongated Gaussians along perceptually equivalent variants. We further introduce Vector Calibrated Sampling (VCS) to keep the sampling process aligned with the line-projection path. In neural vocoding experiments across model sizes, data scales, and sampling steps, the proposed approach consistently improves over the conventional optimal transport CFM, with particularly strong gains in low-resource and few-step scenarios. These results highlight the potential of LP-CFM and VCS to provide a more robust and perceptually grounded generative modeling of speech.

***Index Terms*—** speech modeling, flow matching, low-resource modeling, perceptual invariance, neural vocoding

1. INTRODUCTION

Recent advances in speech generation have identified conditional flow matching (CFM) [1] as a powerful alternative to diffusion-based models. CFM learns a time-dependent vector field that gradually transports samples from a simple source distribution to the complex target data distribution, achieving strong performance in various speech modeling tasks such as speech synthesis, enhancement, and separation [2–7].

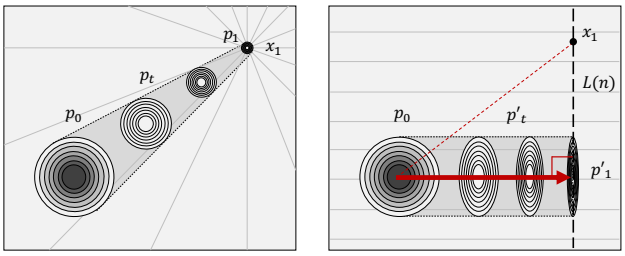
From the generative perspective, human auditory perception is generally robust to global amplitude scaling and small temporal shifts. In practice, two waveforms that differ only in loudness or slight temporal alignment are often perceived as perceptually identical [8–10]. This property has already been exploited in several speech-related tasks. For instance, the scale-invariant signal-to-distortion ratio (SI-SDR) [8] is widely adopted as an objective in speech separation for its robustness to amplitude variations [11–14]. Similarly, phase shift-invariant training (PSIT) [10] has been shown to enhance both the performance and training stability of speech enhancement models by relaxing strict temporal alignment. In contrast, conventional generative formulations, including CFM, are not inherently designed with such flexibility. They typically enforce learning a single instance from the dataset and penalize any deviation, even when the alternative outputs are perceptually equivalent. This rigid objective could lead to inefficient use of data and model capacity.

Motivated by these observations, we propose Linear Projection Conditional Flow Matching (LP-CFM), a new formulation of CFM that explicitly incorporates these perceptual invariances. Rather than matching to an isotropic Gaussian centered on a single data point, LP-CFM defines the target as an elongated Gaussian distribution along a line that represents a set of perceptually equivalent targets (e.g., variants differing only in global amplitude or temporal alignment), as illustrated in Fig. 1. This design encourages the model to learn a flow that directs samples toward the closest valid point within the equivalence set, instead of forcing convergence to one arbitrary instance. Furthermore, we introduce Vector Calibrated Sampling (VCS), a simple yet effective correction strategy that ensures sampling remains consistent with the projection-based geometry. Together, LP-CFM and VCS enable the model to capture speech distributions in a more efficient and perceptually meaningful way.

To validate the effectiveness of our proposed LP-CFM, we conduct various experiments within a controlled neural vocoding setting. Through a comparative analysis against the optimal transport CFM (OT-CFM), we confirm that our approach achieves consistently better outcomes under diverse conditions. This performance gain is especially pronounced in challenging scenarios, such as limited model capacity or a low number of sampling steps. These findings position LP-CFM as a robust and generalizable alternative for speech generation, and more broadly, as a step toward generative models that align more closely with human perceptual structures.

2. BACKGROUND

Flow matching [1] formulates generative modeling as learning a continuous-time flow that maps a simple prior distribution p_0 into the data distribution p_1. While the ideal marginal vector field governing the transport of the entire distribution



(a) OT-CFM (b) LP-CFM

Fig. 1: Conceptual illustration of OT-CFM and LP-CFM: (a) OT-CFM models a spherical (isotropic) Gaussian distribution around target sample x_1, whereas (b) LP-CFM places mass along the shortest projection path toward a line formed by equivalent variants of the target sample, resulting in an elongated Gaussian distribution.

is intractable, conditional flow matching (CFM) provides a tractable, simulation-free objective. CFM works by defining simpler conditional probability paths $p_t(x|x_1)$ for each data point x_1 and training vector v_θ to match the corresponding conditional vector fields $u_t(x|x_1)$. A powerful and widely-used implementation is optimal transport CFM (OT-CFM) [1], which constructs a straight-line probability path between a prior sample $x_0 \sim \mathcal{N}(0, I)$ and a data sample x_1. The intermediate path x_t is defined as an interpolation that transforms the prior into a narrow Gaussian $\mathcal{N}(x_1, \sigma_{\min}^2 I)$ centered at the data point:

$$x_t = (1 - (1 - \sigma_{\min})t)x_0 + tx_1.$$

For a point on this path, the conditional vector field $u_t(x_t|x_1)$ is a time-invariant vector equivalent to the path velocity $\dot{x}_t$:

$$u_t(x_t|x_1) = \dot{x}_t = x_1 - (1 - \sigma_{\min})x_0.$$

The training objective is therefore formulated as a regression problem with the following loss function:

$$\mathcal{L}_{\text{CFM}}(\theta) = \mathbb{E}_{t,x_0,x_1}\left[\left\|v_\theta(x_t, t) - u_t(x|x_1)\right\|^2\right]. \quad (1)$$

3. METHODOLOGY

3.1. Linear Projection CFM (LP-CFM)

In practical generation tasks, a given target x_1 may have multiple equivalent variants that are perceived as indistinguishable in quality. We assume such variants lie on a line

$$L(n; x_1) = a(x_1)n + b(x_1), \quad n \in \mathbb{R},$$

where $a(x_1) \in \mathbb{R}^d$ is a direction vector and $b(x_1) \in \mathbb{R}^d$ is an offset. Any point on $L(n; x_1)$ is then considered a valid variant of x_1.

3.1.1. Target distribution construction

Under this assumption, instead of modeling the target as an isotropic Gaussian centered at x_1, we define an elongated Gaussian distribution that concentrates around the line $L(n; x_1)$. Let $p_0 = \mathcal{N}(\mu_0, \Sigma_0)$ be the source distribution. We construct target distribution as

$$p'_1(x \mid x_1) = \mathcal{N}\big(b + P(\mu_0 - b),\ M\Sigma_0 M^\top\big),$$

$$\text{where} \quad M = \lambda I + (1 - \lambda)P, \quad P = \frac{aa^\top}{a^\top a}, \quad \lambda \in (0, 1].$$

Here, P is the projection matrix onto the line direction a, and M shrinks orthogonal components by a factor λ. Intuitively, this operation translates p_0 toward the closest point on $L(n; x_1)$ and compresses variance in the orthogonal subspace, yielding a thin, elongated Gaussian aligned with the line. We set p_0 as $\mathcal{N}(0, I)$ which in turn simplifies to

$$p'_1(x \mid x_1) = \mathcal{N}(b - Pb,\ MM^\top).$$

3.1.2. Conditional path and velocity

The conditional probability path is defined as the Wasserstein-2 displacement interpolation between p_0 and p'_1. Since Gaussians are closed under W_2 displacement interpolation [15], the intermediate distribution p'_t is Gaussian as well. At the sample level, the interpolation can be written as

$$x_t = (1 - t)x_0 + t\,(b - Pb + Mx_0), \quad x_0 \sim p_0,$$

with target velocity

$$u_t(x|x_1) = \dot{x}_t = (b - Pb + Mx_0) - x_0.$$

The training objective follows the CFM form as Equation (1).

This formulation naturally includes OT-CFM as a special case: when $\lambda = \sigma_{\min}$, $b = x_1$, and the line is undefined ($a = 0 \Rightarrow P = 0$), the formulation becomes identical to the isotropic Gaussian target of OT-CFM. Our method can therefore be viewed as a more general formulation that adapts the covariance structure to reflect equivalence classes of data along $L(n; x_1)$.

3.2. Application on Speech Modeling

Theoretically, LP-CFM can be applied to any kind of generation task if its variants can be expressed as a line equation $L(n) = an + b$. To provide a concrete instantiation of our general theory, this paper focuses on the task of speech modeling. We formulate the proposed equations for this specific domain by leveraging the core properties of the short-time Fourier transform (STFT).

3.2.1. Scaling property

When a signal is scaled in amplitude by a factor s, its magnitude spectrogram becomes $X_{\text{mag},y} = |s|X_{\text{mag}}$. Taking the logarithm of this equation yields an additive relationship:

$$\log X_{\text{mag},y} = \log X_{\text{mag}} + \log|s|.$$

By setting the variant parameter $n = \log|s| \in \mathbb{R}$, this forms the line equation $L(n) = \log X_{\text{mag}} + n$, which corresponds to the line with a slope of $a = 1$ and an offset of $b = \log X_{\text{mag}}$.

3.2.2. *Shifting property*

When a signal is shifted in time of τ, the phase spectrogram X_{pha} is modified as follows:

$$X_{\text{pha},y}[k] = X_{\text{pha}}[k] - \frac{2\pi k}{N}\tau,$$

where k is the frequency-bin index and N is the FFT size. This expression is inherently a line equation. By defining $n = \tau$ and a constant vector $\kappa[k] = 2\pi k/N$, the equation takes the form $L(n) = X_{\text{pha}} - n\kappa$. This corresponds to a line with a slope of $a = -\kappa$ and an offset of $b = X_{\text{pha}}$. Through these constructions, both log-magnitude and phase spectrograms admit linear variant sets. This derivation can also be applied to broader domains, such as log-mel spectrograms, which share the same scaling property.

3.3. Vector Calibrated Sampling (VCS)

In LP-CFM, the target velocity u_t is by definition orthogonal to its corresponding target line. However, the predicted vector v may contain small, erroneous components parallel to this line due to prediction errors. To address this, we propose Vector Calibrated Sampling (VCS), a simple correction applied during inference to enforce this geometric constraint. VCS removes the erroneous component of the predicted vector v that is parallel to the target line, while preserving the vector's original magnitude:

$$v' = \frac{\|v\|}{\|(I-P)v\|}(I-P)v.$$

This operation is feasible in our speech application because the line slopes are known constants ($a = 1$ for log-magnitude and $a = -\kappa$ for phase). The purpose of VCS is not to significantly boost performance, but to act as a safeguard that ensures the sampling process remains consistent with the geometric properties of the LP-CFM framework.

4. EXPERIMENTS

4.1. Experimental Setup

We evaluate our method on a neural vocoding setup, converting mel-spectrograms into waveforms. This task serves as a controlled testbed for modeling both magnitude and phase, allowing for a relative comparison against OT-CFM across varying conditions to isolate the contributions of LP-CFM.

Model architecture. To control for architectural factors, we simplify the model design. The mel encoder consists of 1D convolution with kernel size of 7 followed by a single ConvNeXt V2 [16] block, which maps mel bins to the STFT frequency dimension. The encoded mel features are channel-wise concatenated with the input of a flow matching decoder to predict vectors for both magnitude and phase spectrograms. The decoder is a minimally modified open-source 2D UNet backbone[1], featuring one ResNet block per scale, three scales in total with no attention modules. We build three model sizes with channel configurations of [16,32,64], [32,64,128], and [64,128,256], using group normalization with 2, 4, and 8 groups, respectively. The decoded magnitude and phase spectrograms are converted to a waveform via an inverse STFT.

Table 1: Results under different model sizes.

Model	Method	M-STFT↓	PESQ↑	MCD↓	Period↓	V/UV F1↑	UTMOS↑
UNet-16	OT	1.0399	3.743	2.223	0.1108	0.9596	2.8715
	LP	**1.0253**	**3.858**	**2.174**	**0.1050**	**0.9614**	**3.0153**
UNet-32	OT	0.9917	4.011	2.048	0.0908	0.9655	3.2254
	LP	**0.9848**	**4.097**	**2.018**	**0.0881**	**0.9665**	**3.2647**
UNet-64	OT	0.9670	4.180	1.975	0.0801	0.9704	3.3900
	LP	**0.9631**	**4.191**	**1.942**	**0.0772**	**0.9709**	**3.4231**

Training details. All experiments use the single-speaker LJ Speech [17] dataset. Following prior work [18], mel-spectrograms and target STFTs are extracted using a 1024-point FFT, a 256-sample hop size, and 80 mel bins (0–8 kHz). We use a train-validation split of 12,950 and 150 samples [18]. To ensure a fair comparison, both LP-CFM and OT-CFM are trained with identical settings and a fixed random seed. We set λ as 1×10^{-4}, matching the σ_{min} value used in OT-CFM. Models are trained for 500 epochs on a single RTX 4090 GPU with a batch size of 16. We use the AdamW optimizer with betas of (0.9, 0.99), learning rate of 5×10^{-4}, decayed exponentially by 0.99 per epoch.

Evaluation metrics. We report performance using standard vocoder metrics: multi-resolution STFT (M-STFT) [19], PESQ [20], mel-cepstral distance (MCD) [21], periodicity error and V/UV F1 [22], along with UTMOS [23] as an automated proxy for subjective quality. For sampling, we utilize first-order Euler ODE-solver with a sampling step of 6 as a default. Since LP-CFM can produce outputs with various scales, all target and predicted waveforms for both methods are peak-normalized to 0.95 before evaluation.

4.2. Results and Analysis

Impact of model size. The analysis begins with model capacity, examining how it influences the relative behavior between LP-CFM and OT-CFM. As shown in Table 1, LP-CFM provides consistent gains across multiple architectures. The improvements are particularly notable when the model size is small (e.g., UNet-16), and the gap narrows for larger models. This performance trend can be attributed to line-projection geometry of LP-CFM: by targeting the closest point on a line of valid variants rather than a path converging to a single fixed point, the transport path length and variability are reduced. This property may ease the optimization, especially for models with limited capacity.

Data efficiency. To compare data efficiency between the two methods, we train models on randomly sampled subsets of

[1] https://huggingface.co/docs/diffusers/api/models/unet2d

Table 2: Results under different dataset sizes on UNet-32.

Trainset	Method	M-STFT↓	PESQ↑	MCD↓	Period↓	V/UV F1↑	UTMOS↑
LJ - 33%	OT	1.0176	3.929	2.124	0.0992	0.9618	3.1118
	LP	**1.0153**	**3.975**	**2.101**	**0.0976**	**0.9634**	**3.1501**
LJ - 66%	OT	1.0047	3.994	2.051	0.0941	0.9646	3.1718
	LP	**0.9968**	**4.071**	**2.037**	**0.0902**	**0.9669**	**3.2416**
LJ - 100%	OT	0.9917	4.011	2.048	0.0908	0.9655	3.2254
	LP	**0.9848**	**4.097**	**2.018**	**0.0881**	**0.9665**	**3.2647**

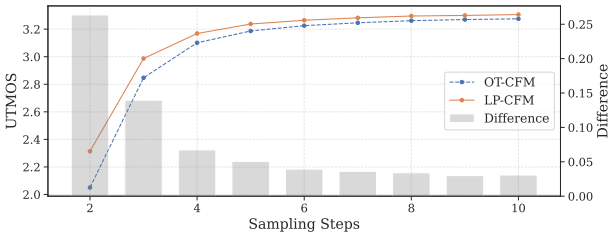


Fig. 2: Step-wise UTMOS comparison between OT-CFM and LP-CFM on UNet-32. UTMOS scores are shown as line (left axis), and the bar represents their score difference (right axis).

LJSpeech and compare their performance. Table 2 demonstrates that LP-CFM consistently surpasses OT-CFM even under limited data scenarios. For example, training with only 66% of the data still yields higher performance than OT-CFM trained on the full dataset across most metrics. Since LP-CFM constructs its target distribution by capturing a set of multiple variants with a single elongated Gaussian, the model can leverage a richer and more diverse set of data instances than the dataset alone provides. This approach resembles data augmentation, but with a crucial distinction: instead of exposing on arbitrary variants, LP-CFM dynamically steers the learning process toward the closest variant under the current flow.

Sampling efficiency. Building on the above findings, we next examine how the line-projection geometry affects inference behavior. We evaluate UTMOS performance under varying step budgets. As shown in Fig. 2, LP-CFM consistently achieves higher scores across different numbers of steps, with its advantage most evident in few-step regimes where approximation errors tend to accumulate. These results suggest that the proposed line-projection geometry—which yields shorter and more consistent transport paths—not only facilitates optimization and improves data efficiency, but also proves effective in sampling, particularly in low-step settings where error accumulation is a concern.

Subjective evaluation. To examine how the observed objective gains translate into perceptual quality, we conducted a comparative mean opinion score (CMOS) evaluation on 15 randomly chosen validation samples, each rated by 25 listeners across four representative scenarios. We verify the results with one-sample t-tests against zero, confirming that all results are statistically significant (p-value < 0.05). As shown in Table 3, listeners express a clear preference for LP-CFM in small-model and few-step sampling conditions. In the other scenarios, LP-CFM also receives consistently positive ratings, with relatively higher preference under the low-data setting. Taken together, these results indicate that the perceptual advantages of LP-CFM are consistent with its objective improvements.

Table 3: Results of CMOS test between OT-CFM and LP-CFM on representative scenarios with 95% confidence interval. Higher scores indicate stronger preference for LP-CFM.

Scenario	CMOS ↑
UNet-32, 6 steps	0.12±0.09
UNet-16, 6 steps	**0.46±0.10**
UNet-32, 33% data, 6 steps	0.17±0.10
UNet-32, 3 steps	0.35±0.12

Table 4: Results of ablation study on UNet-32.

	Method		VCS	M-STFT↓	PESQ↑	MCD↓	Period↓	V/UV F1↑	UTMOS↑
	Mag.	Pha.							
(1)	OT	OT	x	0.9917	4.011	2.048	0.0908	0.9655	3.2254
(2)	OT	OT	o	5.4160	1.102	11.138	0.6437	0.0058	1.6226
(3)	OT	LP	x	0.9935	4.016	2.030	0.0909	0.9658	3.2263
(4)	LP	OT	x	0.9856	4.088	2.022	0.0880	**0.9665**	3.2550
(5)	LP	LP	x	0.9859	4.094	2.019	**0.0879**	**0.9665**	3.2627
(6)	LP	LP	o	**0.9848**	**4.097**	**2.018**	0.0881	**0.9665**	**3.2647**

Ablation study. To disentangle the contributions of each component, we evaluate LP-CFM when applied separately to magnitude (row 4) and phase (row 3), in comparison with OT-CFM applied to both (row 1). As shown in Table 4, applying it to the magnitude yielded the dominant improvements, while phase-only application resulted in smaller gains. This outcome can be attributed to the dominant role of the magnitude on speech quality, as well as the inherent complexity of phase modeling. Ultimately, applying LP-CFM to both components produced the most balanced performance (row 5).

We also examine the effect of VCS. When combined with LP-CFM, VCS behaved as an intended safeguard: it neither boosted performance nor harmed it, yielding comparable or slightly higher scores (row 6). In contrast, applying VCS to OT-CFM significantly degraded performance (row 2), which is expected since it does not assume projection-aligned trajectories. This contrast provides indirect evidence that LP-CFM has indeed learned the intended projection-aligned paths. Since VCS explicitly removes the parallel component, a model not following such trajectories would be expected to suffer the same degradation observed in OT-CFM.

5. CONCLUSION

In this work, we introduced LP-CFM, a perceptual invariance–aware refinement of conditional flow matching that aligns training with perceptual equivalence in speech. As a proof of concept, we evaluated the proposed method in a controlled neural vocoding setup, where it delivered consistent gains over OT-CFM across diverse conditions. Its advantages were most pronounced in resource-constrained scenarios, including limited model capacity, data scarcity, and few-step sampling—conditions often encountered in practical applications. We expect that LP-CFM will serve as a foundation for more perceptually informed generative speech models and inspire further exploration of invariance-aware modeling.

6. ACKNOWLEDGEMENT

This work was supported by IITP grant funded by the Korea government (MSIT, RS-2024-00457882, National AI Research Lab Project).

7. REFERENCES

[1] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le, "Flow matching for generative modeling," in *Proc. ICLR*, 2023.

[2] Sang-Hoon Lee, Ha-Yeong Choi, and Seong-Whan Lee, "Periodwave: Multi-period flow matching for high-fidelity waveform generation," in *Proc. ICLR*, 2025.

[3] Tianze Luo, Xingchen Miao, and Wenbo Duan, "Wavefm: A high-fidelity and efficient vocoder based on flow matching," in *Proc. NAACL*, 2025.

[4] Shivam Mehta, Ruibo Tu, Jonas Beskow, Éva Székely, and Gustav Eje Henter, "Matcha-tts: A fast tts architecture with conditional flow matching," in *Proc. ICASSP*, 2024.

[5] Youngjoon Jang, Ji-Hoon Kim, Junseok Ahn, Doyeop Kwak, Hong-Sun Yang, Yoon-Cheol Ju, Il-Hwan Kim, Byeong-Yeol Kim, and Joon Son Chung, "Faces that speak: Jointly synthesising talking face and speech from text," in *Proc. CVPR*, 2024.

[6] Seonggyu Lee, Sein Cheong, Sangwook Han, and Jong Won Shin, "Flowse: Flow matching-based speech enhancement," in *Proc. ICASSP*, 2025.

[7] Alexander H Liu, Matthew Le, Apoorv Vyas, Bowen Shi, Andros Tjandra, and Wei-Ning Hsu, "Generative pre-training for speech with flow matching," in *Proc. ICLR*, 2024.

[8] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey, "Sdr–half-baked or well done?," in *Proc. ICASSP*, 2019.

[9] Shiqi Zhang, Zheng Qiu, Daiki Takeuchi, Noboru Harada, and Shoji Makino, "Unrestricted global phase bias-aware single-channel speech enhancement with conformer-based metric gan," in *Proc. ICASSP*, 2024.

[10] Doyeop Kwak, Youngjoon Jang, Seongyu Kim, and Joon Son Chung, "Ednet: A distortion-agnostic speech enhancement framework with gating mamba mechanism and phase shift-invariant training," *arXiv preprint arXiv:2506.16231*, 2025.

[11] Yi Luo and Nima Mesgarani, "Tasnet: time-domain audio separation network for real-time, single-channel speech separation," in *Proc. ICASSP*, 2018.

[12] Yi Luo and Nima Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[13] Cem Subakan, Mirco Ravanelli, Samuele Cornell, Mirko Bronzi, and Jianyuan Zhong, "Attention is all you need in speech separation," in *Proc. ICASSP*, 2021.

[14] Shengkui Zhao and Bin Ma, "Mossformer: Pushing the performance limit of monaural speech separation using gated single-head transformer with convolution-augmented joint self-attentions," in *Proc. ICASSP*, 2023.

[15] Asuka Takatsu, "Wasserstein geometry of gaussian measures," 2011.

[16] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie, "Convnext v2: Co-designing and scaling convnets with masked autoencoders," in *Proc. CVPR*, 2023.

[17] Keith Ito and Linda Johnson, "The lj speech dataset," `https://keithito.com/LJ-Speech-Dataset/`, 2017.

[18] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Proc. NeurIPS*, 2020.

[19] Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *Proc. ICASSP*, 2020.

[20] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP*, 2001.

[21] Robert Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proceedings of IEEE pacific rim conference on communications computers and signal processing*, 1993.

[22] Max Morrison, Rithesh Kumar, Kundan Kumar, Prem Seetharaman, Aaron Courville, and Yoshua Bengio, "Chunked autoregressive gan for conditional waveform synthesis," in *Proc. ICLR*, 2022.

[23] Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari, "Utmos: Utokyo-sarulab system for voicemos challenge 2022," in *Proc. Interspeech*, 2022.