eCat: An End-to-End Model for Multi-Speaker TTS & Many-to-Many Fine-Grained Prosody Transfer

Ammar Abbas*, Sri Karlapati*, Bastian Schnell, Penny Karanasou, Marcel Granero Moya, Amith Nagaraj, Ayman Boustati, Nicole Peinelt, Alexis Moinet, Thomas Drugman

Alexa AI, Amazon

syeabbs@amazon.co.uk, srikarla@amazon.co.uk

Abstract

We present eCat, a novel end-to-end multispeaker model capable of: a) generating long-context speech with expressive and contextually appropriate prosody, and b) performing fine-grained prosody transfer between any pair of seen speakers. eCat is trained using a two-stage training approach. In Stage I, the model learns speaker-independent word-level prosody representations in an end-to-end fashion from speech. In Stage II, we learn to predict the prosody representations using the contextual information available in text. We compare eCat to CopyCat2, a model capable of both fine-grained prosody transfer (FPT) and multi-speaker TTS. We show that eCat statistically significantly reduces the gap in naturalness between Copy-Cat2 and human recordings by an average of 46.7% across 2 languages, 3 locales, and 7 speakers, along with better target-speaker similarity in FPT. We also compare eCat to VITS, and show a statistically significant preference.

Index Terms: Multi-speaker TTS, prosody transfer, contextual prosody, end-to-end training

1. Introduction

Previously, Neural Text-to-Speech (NTTS) methods depended on training a separate component to generate mel-spectrograms from phonemized text [1–4] and another component (vocoder) to generate speech waveforms from mel-spectrograms [5–9]. The mel-spectrograms served as an intermediary between the two components. This approach suffered from compounding errors, as the vocoder is exposed to mel-spectrograms obtained from speech waveforms during training but predicted mel-spectrograms during inference. To assuage this, the vocoder was also sometimes fine-tuned on synthesised mel-spectrograms. Recently, NTTS methods have begun to use a single end-to-end model to generate speech waveforms directly from phonemized input text [10–12], resulting in an improvement in the naturalness of synthesised speech.

On a parallel note, there has also been work in multi-sentence TTS or long-context TTS [13–16]. The aim of such methods is to generate speech with consistent inter-sentence prosody by exploiting the dependency between consecutive sentences in long-context text. Makarov *et al.* [13] concatenate consecutive sentences into a single training sentence and train the TTS system to predict the mel-spectrogram for the concatenated sentences. Detai *et al.* [15] also use acoustic and textual context from surrounding sentences to synthesise speech with contextually appropriate prosody. Liumeng *et al.* [14] show that learning inter-sentence information in paragraphs with multi-head attention mechanism improves naturalness of TTS over

sentence-based models. Overall, these methods affirm that using additional context around the target sentence improves the contextual relevance of prosody across sentences.

In the realm of prosody representation learning, there has been significant work in learning speaker independent prosody representations for many-to-many fine-grained prosody transfer (FPT) [17–19]. FPT methods aim to extract prosody from a source speaker at a fine-grained level such as word, phoneme, or mel-spectrogram frame-level, and generate speech in a different target speaker's identity using prosody of the source speaker. Klimkov *et al.* [17] showed that by extracting prosody relevant speech features like pitch, energy, and durations at the phoneme-level, we can transfer prosody at a fine-grained level from any speaker to a target speaker's identity. Karlapati *et al.* [18] used a conditional variational autoencoder to learn speaker-independent prosody representations at the frame-level to capture source prosody. In CopyCat2 (CC2), Karlapati *et al.* [19] showed that learning fine-grained speaker-independent prosody representations at the word-level helps reduce speaker leakage in FPT and can be used in the downstream task of TTS with contextually appropriate prosody. They did this by predicting the prosody representations learnt from speech using the contextual information available in the text.

In this work, we present eCat, with 4 major contributions: i) to the best of our knowledge, eCat is the first end-to-end system capable of many-to-many FPT and multi-speaker TTS; ii) we show that training the acoustic model in an end-to-end fashion results in improved speaker similarity to the target speaker in FPT when compared to CC2; iii) we show that eCat provides naturalness improvements in multi-speaker TTS over CC2 on long-context text. These results are shown on English and Spanish internal datasets, consisting of data from en-US, en-GB, and es-US locales, and a total of 3 male and 4 female speakers; iv) we show that eCat is significantly preferred over VITS [10], a model built on a different architecture, in multi-speaker TTS.

2. Proposed Method: eCat

eCat consists of 3 components: 1) end-to-end acoustic model, 2) duration model, and 3) long-context flow-based prosody predictor called FlowCat. These 3 components are trained using a two-stage approach [19–21]. In Stage I, we learn word-level speaker-independent prosody representations from multi-speaker data, while in Stage II, we learn to predict these representations using the contextual information available in text.

2.1. Stage I: Learning Prosody Representations

In eCat, as shown in Figure 1, the acoustic model consists of a phoneme encoder, a non-autoregressive (NAR) decoder, a conditional variational reference encoder, and the BigVGAN-base vocoder [9]. All components in the acoustic model are trained
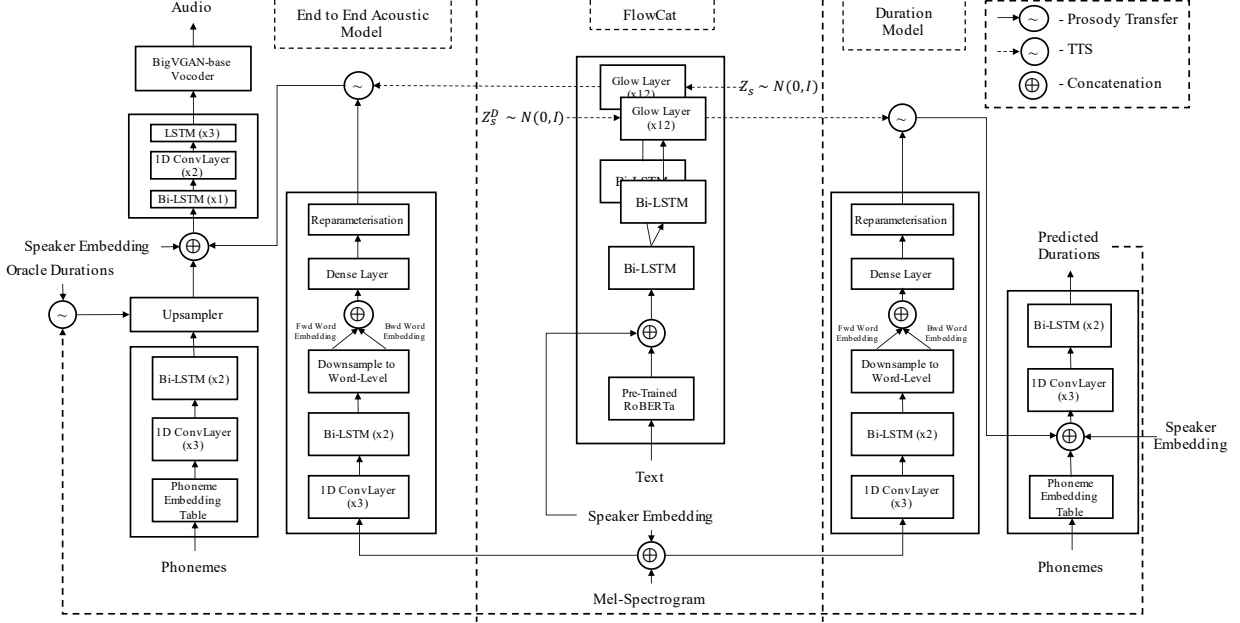
10.21437/Interspeech.2023-965

Figure 1: *eCat architecture. Left: end-to-end acoustic model; middle: flow-based prosody predictor (FlowCat); right: duration model.*

end-to-end. Similar to CC2, the phoneme encoder takes a vector of P phonemes, $\boldsymbol{y}$, as input and provides phoneme encodings as output. These encodings are upsampled through replication according to the per-phoneme durations, $\boldsymbol{d}$, provided as input to the Upsampler. These upsampled encodings are then passed to an NAR decoder along with speaker embeddings [18], $\boldsymbol{c} \in \mathbb{R}^E$, where E is the size of a speaker embedding. As the model is trained end-to-end, the decoder generates an intermediate representation, $B \in \mathbb{R}^{T_{mel} \times 80}$, instead of mel-spectrogram $X \in \mathbb{R}^{T_{mel} \times 80}$, where T_{mel} is the number of mel-spectrogram frames. B is then passed to the BigVGAN-base vocoder as input to generate a waveform, $\boldsymbol{x}$, of length T samples. We aim to learn from speech a speaker-independent word-level representation of prosody, $Z = [\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_W]$, where $\boldsymbol{z}_i \in \mathbb{R}^U$ is a word-level prosody vector, W is number of words in a sentence, and U is the dimension of each prosody vector. As computing the true distribution of prosody is intractable, we use variational methods and the re-parameterisation trick to approximate the word-level distribution of prosody using a Conditional VAE, $q_\phi(Z \mid \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{d}, \boldsymbol{c})$ [22]. We assume that the vocoder parameters are also included in θ and train the end-to-end acoustic model to maximise the ELBO: $log\, p_\theta(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{d}, \boldsymbol{c}, Z) - KL(q_\phi(Z \mid \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{d}, \boldsymbol{c}) \mid\mid p(Z))$.

In CC2, the authors had a similar ELBO where they used the $L2$ loss between generated mel-spectrograms and those obtained from speech waveforms as a proxy for negative log-likelihood. In eCat, we maximise the log-likelihood by using a modified GAN training scheme from BigVGAN [9]. We use the end-to-end acoustic model as the generator and represent the generated waveform as $\boldsymbol{x'}$. We use 5 multi-periodicity (D_{MPD}) and 3 multi-resolution (D_{MRD}) discriminators used in BigV-GAN as the discriminators, $D \in \{D_{MPD}\} \cup \{D_{MRD}\}$. We use $\mathcal{A}$ to denote our dataset consisting of tuples of speech waveforms and corresponding text $(\boldsymbol{x}, t_x)$. For each discriminator $D_k \in D$, the generator and discriminator losses are given as:

$$L_g^k = \mathop{\mathbb{E}}_{Z \sim q_\phi} \left[(D_k(\boldsymbol{x'}) - 1)^2 \right], \tag{1}$$

$$L_d^k = \mathop{\mathbb{E}}_{(\boldsymbol{x}, \boldsymbol{y}) \sim \mathcal{A}, \ Z \sim q_\phi} \left[(D_k(\boldsymbol{x}) - 1)^2 + (D_k(\boldsymbol{x'}))^2 \right]. \tag{2}$$

We also use modified versions of the feature matching loss (L_f) and mel-spectrogram loss (L_p) from BigVGAN. We take the feature matching loss as the $L1$ loss between the outputs of each intermediate layer in a discriminator, $D_k^i \in D_k$, and use $\mid D_k \mid$ to refer to the number of layers in D_k. For L_p, we compute the $L1$ loss between the mel-spectrograms X and X' of the target and generated waveforms $\boldsymbol{x}$ and $\boldsymbol{x'}$:

$$L_f^{k,i} = \mid\mid D_k^i(\boldsymbol{x}) - D_k^i(\boldsymbol{x'}) \mid\mid_1$$

$$L_f^k = \mathop{\mathbb{E}}_{(\boldsymbol{x}, \boldsymbol{y}) \sim \mathcal{A}, \ Z \sim q_\phi} \left[\frac{1}{\mid D_k \mid} \sum_i L_f^{k,i} \right], \tag{3}$$

$$L_p = \mathop{\mathbb{E}}_{(\boldsymbol{x}, \boldsymbol{y}) \sim \mathcal{A}, \ Z \sim q_\phi} \left[\mid\mid X - X' \mid\mid_1 \right].$$

We use η to denote the parameters in all the discriminators. We know that θ and ϕ are the set of all acoustic model parameters. Thus, the acoustic model and discriminator losses are:

$$\operatorname*{argmin}_{\theta, \ \phi} \mathbb{L}_{AM} = \sum_k \left(L_g^k + \lambda_f L_f^k \right) + \lambda_p L_p$$

$$+ \alpha \sum_{i=1}^W KL(q_\phi(\boldsymbol{z}_i \mid \boldsymbol{x}, \boldsymbol{c}) \mid\mid p(\boldsymbol{z}_i)), \tag{4}$$

$$\operatorname*{argmin}_{\eta} \mathbb{L}_D = \sum_k L_d^k.$$

As shown to the right in Figure 1, the duration model in eCat is unchanged from CC2. We learn word-level speaker-independent duration prosody representations, $Z^D \in \mathbb{R}^{W \times U^D}$, where U^D is the size of each representation. To learn Z^D, like in the acoustic model we use a conditional variational reference encoder. We train the duration model, r_θ to maximise the evidence lower bound (ELBO): $log\, r_\theta(\boldsymbol{d} \mid \boldsymbol{y}, \boldsymbol{c}, Z^D) - KL(q_\psi(Z^D \mid \boldsymbol{x}, \boldsymbol{c}) \mid\mid r(Z^D))$, where $r(Z^D) = N(\boldsymbol{0}, I)$.

2.2. Stage II: FlowCat - Predicting Prosody Representations from Text

The acoustic and duration prosody distributions, $Z \sim q_\phi$ and $Z^D \sim q_\psi$ are learnt at the end of Stage I. Both distributions are conditional upon $\boldsymbol{x}$, which is unavailable during inference. We

learn to predict these latents using the contextual information available in text. As shown in the middle in Figure 1, we define a prosody predictor called FlowCat, $s_\nu(Z, Z^D \mid t_x, \boldsymbol{c})$, to predict Z and Z^D given text t_x and speaker identity $\boldsymbol{c}$. We use normalising flows and condition them on the contextual information available in text. We use RoBERTa [23] to get contextualised word embeddings from text. We combine these word embeddings with speaker embeddings to condition the flow layers. Normalizing flows are a class of generative models which provide exact log-likelihood estimates by using invertible functions and the change of variables trick [24]. We use Glow-based flow layers as our flow functions [25] which assumes conditional independence between word-level prosody representations.

We note that the individual word-level prosody vectors $\boldsymbol{z}_i$ and $\boldsymbol{z}_i^D$ learnt in Stage I are regularised to be from a simple distribution due to the prior being $N(0, I)$ in the KL divergence. However, we hypothesise that the sequence of word-level prosody latents Z and Z^D are complex distributions. Therefore, we conjecture that flows allow us to sample varied points from the complex prosody distributions, leading to expressive prosody in synthesised speech. When training, we use normalizing flows to transform a sequence of latent vectors from the target prosody distributions, $Z \sim q_\phi$ and $Z^D \sim q_\psi$ to a sequence of latent vectors from simple distributions $Z_s \sim \mathcal{N}(\boldsymbol{0}, I)$ and $Z_s^D \sim \mathcal{N}(\boldsymbol{0}, I)$. We do so through a composition of K invertible flow functions $\boldsymbol{f} = f_1 \circ f_2 \circ \cdots \circ f_K$. We define $H_i = f_i(H_{i-1}; t_x, \boldsymbol{c})$ as the output from each flow function. Each $H_i = [\boldsymbol{h}_{1,i}, \boldsymbol{h}_{2,i}, \ldots, \boldsymbol{h}_{W,i}]$ is a sequence of W vectors. We define $H_0 = Z$, $H_0^D = Z^D$, $H_K^D = Z_s^D$, $H_K = Z_s$, $s(Z_s) = \mathcal{N}(\boldsymbol{0}, I)$, and $s(Z_s^D) = \mathcal{N}(\boldsymbol{0}, I)$. We maximise the following log-likelihood of sampling points from the target distribution:

$$
\begin{aligned}
log\, s_\nu(Z, Z^D \mid t_x, \boldsymbol{c}) = & \underset{\substack{Z \sim q_\phi \\ Z^D \sim q_\psi}}{\mathbb{E}} \left[s(Z_s) + s(Z_s^D) + \right. \\
& \left. \sum_{i=1,w=1}^{K,W} log\, det\left(\frac{\partial \boldsymbol{h}_{w,i}}{\partial \boldsymbol{h}_{w,i-1}}\right) + log\, det\left(\frac{\partial \boldsymbol{h}_{w,i}^D}{\partial \boldsymbol{h}_{w,i-1}^D}\right) \right].
\end{aligned}
\tag{5}
$$

3. Experiments

3.1. Data

We conducted experiments on 2 internal datasets ($\mathcal{A}_1$ & $\mathcal{A}_2$). Both datasets consist of speakers reading excerpts from Wikipedia articles, news articles, etc. $\mathcal{A}_1$ is an English dataset and contains 80 hours of speech from 4 speakers: 1 en-US female, 2 en-US male, and 1 en-GB female. $\mathcal{A}_2$ is an es-US dataset and contains a total of 40 hours of speech from 1 male and 2 female speakers. All recordings were sampled at 24 kHz. We split each speaker's data in each dataset into train, validation, and test sets, in the ratio of 7 : 1 : 2 without replacement.

3.2. Training & Hyperparameters

The hyper parameters used to train our eCat models are shown in Table 1. In Stage I, we trained the end-to-end acoustic model and the duration model without multi-sentence input. Training the end-to-end acoustic model to synthesise the whole sentence while generating the waveform at 24 kHz, is very slow. We make 2 deviations to speed-up model training. First, we randomly sample chunks of length M from the waveform to be synthesised during training. Second, we provide mel-spectrograms which are a compressed representation of speech to the reference encoder. We don't pass durations and phonemes to the reference encoder as this information is already captured in the mel-spectrograms. It creates word-level representations

Table 1: *Hyperparameters in eCat.*

Name	Symbol	Value
Speaker Embeddings	E	192 dims
Acoustic Word Prosody Latents	U	4 dims
Duration Word Prosody Latents	U^D	2 dims
Number of Glow Layers	K	12 layers
Feature Matching Loss Weight	λ_f	4
Mel-spectrogram Loss Weight	λ_p	45
KL Divergence Weight	α	10^{-3}
Waveform Chunk Size	M	19200 samples

using frame-to-word alignments which are obtained via forced alignment. We trained the end-to-end acoustic model on 7 V100 GPUs for 440 epochs with a batch size of 84.

In Stage II, we train FlowCat with multi-sentence input. We concatenate consecutive sentences to create sentences within a range of $72 - 95$ words, and provide this as input to FlowCat. We use the prosody embeddings obtained from Stage I to act as the target for FlowCat. This provides our language model with extended context to obtain more contextually relevant word-embeddings. We used pre-trained RoBERTa-base models from Hugging Face [26] for $\mathcal{A}_1$ [23] and $\mathcal{A}_2$ [27], and fine-tune them during training. We trained FlowCat on 4 V100 GPUs for 106 epochs with a batch size of 128.

3.3. Inference

eCat has 2 inference modes: FPT and TTS. In FPT mode, we provide the source waveform and source speaker embeddings from which the prosody is extracted using the duration and acoustic reference encoders. Both the acoustic and duration models are then conditioned with input phonemes from the text and target speaker embedding to generate speech with the target speaker's identity and source prosody.

In TTS mode, we first run FlowCat on a window of text, including the target sentence to be synthesised and its surrounding context. FlowCat predicts the acoustic and duration prosody latents for all the words in the window. We only select latents for the target sentence, and use them in place of the outputs from the duration and acoustic reference encoders. The duration model then predicts per-phoneme target sentence durations. They are used by the end-to-end acoustic model along with the text-predicted acoustic word-level prosody latents, phonemes, and target speaker embeddings to synthesise speech.

3.4. Results

3.4.1. Ablation Studies

We conducted 2 MUSHRA evaluations [28] on $\mathcal{A}_1$, to understand the separate contributions of end-to-end training with BigVGAN-base vocoder and FlowCat. Each MUSHRA was conducted with 24 listeners who were asked to rate, on a scale of 0 to 100, the naturalness of speech samples. Each sample had a concatenated duration of $20 \sim 30$secs. We used pair-wise two-sided Wilcoxon signed rank test with Bonferroni correction to measure the statistical significance of the results.

First, we built a version of eCat, called E2E-CC2, where we used the Prosody Predictor from CC2 instead of FlowCat. We compared E2E-CC2 to CC2 and human recordings, making end-to-end training the only difference between the systems. As shown in Table 2, we found that E2E-CC2 is statistically significantly better than CC2. We hypothesise that this is due to end-to-end training solving the compounding errors problem, resulting in better segmental quality. Second, we compared E2E-CC2, eCat, and human recordings, to determine the impact of FlowCat. We also found that eCat is statistically sig-

Table 2: *Mean MUSHRA scores for ablation study of end-to-end training & FlowCat with* 95% *CI on* $\mathcal{A}_1$. *(**Bold** indicates statistical significance)*

| Studied component | Mean MUSHRA scores | | | |
	CC2	E2E-CC2	eCat	Human
End-to-End	66.9 ± 0.9	$\mathbf{67.4 \pm 0.9}$	N/A	71.4 ± 0.9
FlowCat	N/A	71.3 ± 1.5	$\mathbf{73.9 \pm 1.4}$	76.9 ± 1.3

Table 3: *Mean MUSHRA scores in FPT for prosody similarity to source prosody and for speaker similarity to target speaker identity with* 95% *CI on* $\mathcal{A}_1$.

| | Mean MUSHRA scores | |
	CC2	eCat
Prosody similarity	71.58 ± 1.02	71.70 ± 1.03
Speaker similarity	73.06 ± 0.96	$\mathbf{73.55 \pm 0.96}$

nificantly better when we use FlowCat instead of the prosody predictor from CC2. We conjecture that this is a result of more contextually appropriate and expressive speech due to the use of flows and better transitions between consecutive sentences due to multi-sentence context. Thus, both end-to-end training and FlowCat contribute to improving eCat.

3.4.2. Fine-grained Prosody Transfer

To evaluate FPT, we measured both the prosody similarity to the source prosody and the speaker similarity to the target speaker. We conducted 2 MUSHRAs consisting of CC2 and eCat with a reference speech sample. When evaluating prosody similarity, the reference is a recording from the source speaker whose prosody is to be transferred. When evaluating target speaker similarity, we provide a recording in the target speaker's voice as reference. Each evaluation consisted of 100 listeners, each of whom was presented with 100 samples, 25 each from the speakers in $\mathcal{A}_1$. In evaluating prosody similarity, the listeners had to rate each sample on a scale of 0 to 100 based on how closely it followed the reference's prosody. As shown in Table 3, we found that eCat was on par with CC2 (p-val > 0.05). This shows that the prosody latent spaces learnt in both CC2 and eCat contain similar information for prosody transfer, implying that the improvements in prosody in TTS are owed to the improvement in prosody prediction, confirming FlowCat improvements from Section 3.4.1. To evaluate speaker similarity, the listeners were asked to rate each sample on a scale of 0 to 100 based on how closely it resembles the voice identity of the reference speech. As shown in Table 3, eCat is statistically significantly better (p-val < 0.05) than CC2 in terms of speaker similarity.

3.4.3. TTS Naturalness

We trained separate models for $\mathcal{A}_1$ and $\mathcal{A}_2$ datasets. For each dataset, we conducted a MUSHRA evaluation between CC2, eCat, and human recordings. Both MUSHRAs were run on concatenated samples having a total duration of 20 $\sim$ 30secs. For $\mathcal{A}_1$, we randomly chose 25 such samples from the test set of each of the 4 speakers leading to 100 samples. We found that eCat statistically significantly reduces the gap between CC2 and human recordings (p-val < 0.05) for each speaker, and by 46.9% overall. For $\mathcal{A}_2$, we randomly chose 20 concatenated samples from each speaker's test set for a total of 60 samples. We found that eCat reduces the gap between CC2 and human recordings statistically significantly (p-val < 0.05) for each speaker, and by 46.5% overall. This is very close to the overall gap reduction observed on $\mathcal{A}_1$ as shown in Table 4. We hypothesise that this is because of improvements in both prosody

Table 4: *Mean MUSHRA scores for each speaker evaluated on TTS naturalness from* $\mathcal{A}_1$ & $\mathcal{A}_2$ *dataset with* 95% *CI.*

| | Mean MUSHRA scores | | |
	CC2	eCat	Human
English ($\mathcal{A}_1$)			
en-US male 1	68.5 ± 1.4	$\mathbf{73.1 \pm 1.4}$	77.5 ± 1.3
en-US male 2	69.1 ± 1.4	$\mathbf{74.4 \pm 1.3}$	78.9 ± 1.2
en-US female 1	70.3 ± 1.4	$\mathbf{73.1 \pm 1.4}$	77.9 ± 1.2
en-GB female 1	74.6 ± 1.3	$\mathbf{76.7 \pm 1.2}$	79.2 ± 1.2
Spanish ($\mathcal{A}_2$)			
es-US male 1	75.1 ± 1.8	$\mathbf{78.2 \pm 1.7}$	81.8 ± 1.6
es-US female 1	71.0 ± 1.8	$\mathbf{74.8 \pm 1.6}$	78.7 ± 1.5
es-US female 2	73.7 ± 1.7	$\mathbf{76.5 \pm 1.6}$	80.1 ± 1.5

Table 5: *Percentage preferences in the preference test.*

	VITS	No Preference	eCat
en-US male 1	31.93%	32.53%	35.53%
en-US male 2	30.00%	31.80%	**38.20%**
en-US female 1	29.67%	33.07%	**37.27%**
en-GB female 1	30.40%	32.53%	**37.07%**

and segmental quality. As shown in the ablation study in Section 3.4.1, the improvement in prosody comes from FlowCat, leading to more expressive, coherent and contextually appropriate prosody in the synthesised speech, while the improvement in segmental quality is owed to the end-to-end training.

We also conducted a preference test between eCat and VITS [10]. VITS is an end-to-end model with a different modelling approach that has demonstrated state-of-the-art results in naturalness of synthesised speech. We trained VITS using its public implementation [1] with BigVGAN-base vocoder instead of HiFi-GAN V1 [8] for a fairer comparison. The preference test was conducted on $\mathcal{A}_1$ dataset with 15 concatenated samples for each of the 4 speakers totalling to 60 samples. As shown in Table 5, we found that eCat is statistically significantly preferred to VITS on 3 out of 4 speakers with a consistent improvement. On further analysis, we find that eCat has better expressiveness and consistent prosody across sentences when compared to VITS. This shows the importance of prosody prediction in improving the naturalness of synthesised speech. We also find that both systems have similar segmental quality as they both use the same vocoder and are trained end-to-end.

4. Conclusion

In this paper we introduced eCat, which to the best of our knowledge is the first end-to-end model capable of both fine-grained prosody transfer and multi-speaker TTS. We demonstrated that both end-to-end training and FlowCat bring statistically significant improvements through ablation studies. We showed that eCat reduces the gap in naturalness between CC2 and human speech in 3 locales and 7 voices by an average of 46.7%. We compared eCat to VITS, a state-of-the-art TTS model, and showed that eCat is statistically significantly preferred in 3 out of the 4 voices. We also showed that eCat improves target speaker similarity in prosody transfer settings in comparison to CC2.

5. Acknowledgements

---

[1] https://github.com/jaywalnut310/vits

6. References

[1] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural TTS synthesis by conditioning wavenet on mel-spectrogram predictions," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[2] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," in *Proc. of the AAAI Conference on Artificial Intelligence*, 2019.

[3] R. Skerry-Ryan, E. Battenberg, Y. Xiao *et al.*, "Towards End-to-End Prosody Transfer for Expressive Speech Synthesis with Tacotron," in *Proc. of the International Conference on Machine Learning*, 2018.

[4] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "FastSpeech 2: Fast and High-Quality End-to-End Text to Speech," in *International Conference on Learning Representations*, 2021.

[5] Y. Jiao, A. Gabryś, G. Tinchev, B. Putrycz, D. Korzekwa, and V. Klimkov, "Universal Neural Vocoding with Parallel Wavenet," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[6] J. Lorenzo-Trueba, T. Drugman, J. Latorre, T. Merritt, B. Putrycz, R. Barra-Chicote *et al.*, "Towards Achieving Robust Universal Neural Vocoding," in *Proc. of Interspeech*, 2019.

[7] Kumar, Kundan and Kumar, Rithesh and de Boissiere, Thibault and Gestin, Lucas and Teoh, Wei Zhen and Sotelo, Jose and de Brébisson, Alexandre and Bengio, Yoshua and Courville, Aaron C, "MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis," in *Advances in Neural Information Processing Systems*, 2019.

[8] Kong, Jungil and Kim, Jaehyeon and Bae, Jaekyoung, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," in *Advances in Neural Information Processing Systems*, 2020.

[9] S. gil Lee, W. Ping, B. Ginsburg *et al.*, "BigVGAN: A universal neural vocoder with large-scale training," in *International Conference on Learning Representations*, 2023.

[10] J. Kim, J. Kong, and J. Son, "Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech," in *Proc. of the International Conference on Machine Learning*, 2021.

[11] X. Tan, J. Chen, H. Liu, J. Cong, C. Zhang, Y. Liu, X. Wang, Y. Leng, Y. Yi, L. He *et al.*, "Naturalspeech: End-to-end text to speech synthesis with human-level quality," *arXiv preprint arXiv:2205.04421*, 2022.

[12] J. Donahue, S. Dieleman, M. Binkowski, E. Elsen, and K. Simonyan, "End-to-end Adversarial Text-to-Speech," in *International Conference on Learning Representations*, 2021.

[13] P. Makarov, S. Ammar Abbas, M. Łajszczak, A. Joly, S. Karlapati, A. Moinet, T. Drugman, and P. Karanasou, "Simple and Effective Multi-sentence TTS with Expressive and Coherent Prosody," in *Proc. Interspeech*, 2022.

[14] Xue, Liumeng and Soong, Frank K. and Zhang, Shaofei and Xie, Lei, "ParaTTS: Learning Linguistic and Prosodic Cross-Sentence Information in Paragraph-Based TTS," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2854–2864, 2022.

[15] D. Xin, S. Adavanne, F. Ang, A. Kulkarni, S. Takamichi, and H. Saruwatari, "Improving Speech Prosody of Audiobook Text-to-Speech Synthesis with Acoustic and Textual Contexts," *arXiv preprint arXiv:2211.02336*, 2022.

[16] Pan, Junjie and Wu, Lin and Yin, Xiang and Wu, Pengfei and Xu, Chenchang and Ma, Zejun, "A Chapter-Wise Understanding System for Text-To-Speech in Chinese Novels," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[17] V. Klimkov, S. Ronanki, J. Rohnke, and T. Drugman, "Fine-Grained Robust Prosody Transfer for Single-Speaker Neural Text-To-Speech," in *Proc. Interspeech*, 2019.

[18] S. Karlapati, A. Moinet, A. Joly, V. Klimkov, D. Sáez-Trigueros, and T. Drugman, "CopyCat: Many-to-Many Fine-Grained Prosody Transfer for Neural Text-to-Speech," in *Proc. Interspeech*, 2020.

[19] S. Karlapati, P. Karanasou, M. Łajszczak, S. Ammar Abbas, A. Moinet, P. Makarov, R. Li, A. van Korlaar, S. Slangen, and T. Drugman, "CopyCat2: A Single Model for Multi-Speaker TTS and Many-to-Many Fine-Grained Prosody Transfer," in *Proc. Interspeech*, 2022.

[20] S. Karlapati, A. Abbas, Z. Hodari, A. Moinet, A. Joly, P. Karanasou, and T. Drugman, "Prosodic Representation Learning and Contextual Sampling for Neural Text-to-Speech," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[21] Z. Hodari, A. Moinet, S. Karlapati, J. Lorenzo-Trueba, T. Merritt, A. Joly, A. Abbas, P. Karanasou, and T. Drugman, "Camp: A Two-Stage Approach to Modelling Prosody in Context," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[22] K. Sohn, H. Lee, and X. Yan, "Learning Structured Output Representation using Deep Conditional Generative Models," in *Proc. of Advances in Neural Information Processing Systems*, 2015.

[23] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[24] G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan, "Normalizing Flows for Probabilistic Modeling and Inference," *Journal of Machine Learning Research*, 2022.

[25] D. P. Kingma and P. Dhariwal, "Glow: Generative Flow with Invertible 1x1 Convolutions," in *Advances in Neural Information Processing Systems*, 2018.

[26] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020.

[27] J. D. la Rosa y Eduardo G. Ponferrada y Manu Romero y Paulo Villegas y Pablo González de Prado Salas y María Grandury, "BERTIN: Efficient Pre-Training of a Spanish Language Model using Perplexity Sampling," *Procesamiento del Lenguaje Natural*, vol. 68, no. 0, pp. 13–23, 2022.

[28] ITU-R, Recommendation BS, "1534-1,"Method for the Subjective Assessment of Intermediate Quality Levels of Coding Systems (MUSHRA)"," *International Telecommunication Union*, 2003.