Multi-Scale Spectrogram Modelling for Neural Text-to-Speech

Ammar Abbas, Bajibabu Bollepalli, Alexis Moinet, Arnaud Joly, Penny Karanasou, Peter Makarov, Simon Slangens, Sri Karlapati, Thomas Drugman

Alexa AI, Cambridge, United Kingdom

{syeabbs, bajibabb, drugman}@amazon.com

Abstract

We propose a novel Multi-Scale Spectrogram (MSS) modelling approach to synthesise speech with an improved coarse and fine-grained prosody. We present a generic multi-scale spectrogram prediction mechanism where the system first predicts coarser scale mel-spectrograms that capture the suprasegmental information in speech, and later uses these coarser scale mel-spectrograms to predict finer scale mel-spectrograms capturing fine-grained prosody. We present details for two specific versions of MSS called *Word-level MSS* and *Sentence-level MSS* where the scales in our system are motivated by the linguistic units. The Word-level MSS models word, phoneme, and frame-level spectrograms while Sentence-level MSS models sentence-level spectrogram in addition. Subjective evaluations show that Word-level MSS performs statistically significantly better compared to the baseline on two voices.

Index Terms: neural text-to-speech, multi-scale spectrogram, word-level, sentence-level

1. Introduction

Over the last few years, the progress in Text-To-Speech (TTS) technology has been astounding. Specifically, neural models such as Wavenet [1] and Tacotron [2] have revamped all the components of a modern TTS system [3]. Due to this, the Neural Text-To-Speech (NTTS) has become a standard paradigm where a neural sequence-to-sequence (seq2seq) model is employed to map the input text into acoustic features, and a neural vocoder model is employed to convert the acoustic features into a corresponding waveform. These NTTS systems are capable of generating high-quality speech that is often indistinguishable from human speech [4].

However, NTTS systems still struggle to produce speech with appropriate prosody compared to human speech [5]. The perceived prosody in the synthesized speech may sound inappropriate given the textual context, and includes problems such as wrong type of intonation, pausing, or emphasis. The lack of appropriate prosody stems from multiple reasons ranging from the model design to the way data is processed. Although the prosody is a suprasegmental phenomenon, the existing NTTS systems are designed in a way that they take fine textual representations such as phonemes as an input and predict finer-level acoustic representations such as mel-spectrograms as an output. Thus, the speech produced by the NTTS system can sound flat and require more cognitive effort to process it [6].

Numerous studies have been proposed in the literature to address the aforementioned issues [7, 8, 9, 10]. Most of these studies focus on modelling a latent representation space of prosody using a separate encoder called reference encoder [7]. The reference encoder guides the prosody of the output speech signal to generate expressive speech. The reference encoder can be designed in a variational [11] or non-variational [7] style. Moreover, the latent embedding vectors encoded by reference encoder can be represented either at a coarser level e.g. sentence [8] or at a finer level e.g. word or phonemes [9, 10].

Along with these latent representation based methods, another set of studies focus on modelling prosody in a hierarchical manner along with a reference encoder [12, 13, 14]. Here the input text is represented at various levels that are spanning from coarser (e.g., sentences) to finer (e.g., phonemes) levels. Kenter et al. [13] proposed such kind of hierarchical approach to model prosodic features such as F0, energy, and duration, and these prosodic features along with linguistic features are utilized by a neural vocoder to render the final speech waveform. However, one of the shortcomings of the prosody modelling studies based on latent representations is that they use reference mel-spectrograms to learn prosody embeddings during training, whereas during the inference time they either rely on textual based features to sample [8] or select [15] a prosody embedding from a set of pre-existing latent embeddings. This results in a mismatch between training and inference which could lead to an inappropriate prosody in the output speech signal.

Instead of modelling the prosodic latent space, few studies predict the conventional prosodic features (e.g., F0 and energy) in a multi-task manner and utilize those features to control the prosody of the synthesized speech [16, 17]. The performance of these methods depends on the accuracy and robustness of prosodic feature extraction and modelling. However, F0 extraction and modelling is generally prone to a number of errors [18].

Complementing to the aforementioned studies, in this paper, we propose a Multi-Scale Spectrogram (MSS) modelling technique to capture short and long-range dependencies observed in the speech signal. In MSS, the mel-spectrograms are predicted sequentially from a coarser scale capturing higher-level representation of speech to a finer scale capturing fine-grained prosodic details. Each subsequent finer scale is conditioned by the previous scale's predicted mel-spectrograms. This allows the MSS modelling to produce prosody appropriate at different linguistic units such as sentence, word and phonemes, thereby improving the overall naturalness of the NTTS systems.

Similar to our proposed approach, Vasquez et al. [19] predict the mel-spectrograms in a multi-scale manner. They initially predict lower resolution mel-spectrograms and progressively increase their resolution by 2 times at each scale irrespective of the semantic units in language or acoustic units in speech. Contrary to that, we provide a generic multi-scale mechanism to represent mel-spectrograms and later develop two specific MSS systems where the scales correspond to linguistic units which are sentences, words, and phonemes. Moreover, each scale in MSS has its own objective to learn and all scales are learned together by multi-task learning [20].

Another major difference between [19] and our approach is the use of explicit duration modelling instead of an attention mechanism to find alignment between text and speech.

Due to the stability issues of the attention mechanism of neural seq2seq models in the NTTS systems, the synthesized speech signals could have unpleasant artifacts such as mumbling, repetitions, or skipping [21]. To mitigate the stability issues, non-attentive neural seq2seq models have recently become popular, where the attention mechanism is replaced by an explicit duration model [8, 22, 17]. Hence, the non-attentive neural seq2seq model based NTTS system is employed in this paper.

Our main contributions are as follows: i) We propose a novel multi-scale mel-spectrogram modelling technique to improve the overall quality and naturalness of NTTS systems by appropriately capturing the coarse and fine-grained prosody; ii) We conduct and present an ablation study on two specific versions of MSS, called Sentence-level MSS and Word-level MSS. The Word-level MSS models word, phoneme, and frame-level spectrograms while Sentence-level MSS models sentence-level spectrogram additionally; iii) We evaluate Word-level MSS against a baseline system that is based on external duration model and show that it is significantly better than the baseline on two voices.

2. The Baseline

We use the same baseline system as in [8], which is a modified version of DurIAN [23]. Figure 1 illustrates the block diagram of our baseline system. It is composed of two major components: a seq2seq model and a duration model. First, the input text containing W words $\boldsymbol{w} = [w_0, w_1, \ldots, w_{W-1}]$ is passed through the front-end to extract P phonemes $\boldsymbol{p} = [p_0, p_1, \ldots, p_{P-1}]$ as an output. Next, the P phonemes are passed through an encoder, which captures the relations between phonemes, and produces P phoneme embeddings as an output. The encoder is composed of 1D convolutions followed by a bidirectional LSTM. Finally, the decoder takes these P phoneme embeddings and P phoneme durations in frames $\boldsymbol{d} = [d_0, d_1, \ldots, d_{P-1}]$ where $\sum_{d \in \boldsymbol{d}} = T$ as an input, and predicts T mel-spectrogram frames $\widehat{\mathbf{Y}} = [\boldsymbol{y}_0, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_{T-1}]$ auto-regressively as an output. There is no post-net after decoder as it resulted in instabilities when training with a reduction factor of 1.

During training, the phoneme durations are obtained from a forced-alignment algorithm, whereas during inference, the phoneme durations are predicted by a duration model trained separately. The duration model takes P phonemes as an input and predicts P durations. Both the acoustic and the duration model are optimized using an L2 loss function.

3. Multi-Scale Spectrogram (MSS) modelling

This section introduces the proposed MSS modelling technique. As shown in Figure 1, the decoder of the baseline system predicts the mel-spectrogram frames directly from phoneme embeddings using phoneme durations. Contrary to this, the decoder based on MSS modelling technique predicts the mel-spectrogram frames after conditioning them on all the higher-level mel-spectrograms as illustrated in Figure 2. Specifically, the MSS modelling technique first predicts the mel-spectrogram vectors representing speech on a coarser scale which are later used for the prediction of mel-spectrogram vectors representing speech at a finer scale. The coarser scale representation captures most of the suprasegmental aspects of the prosody resulting in a more appropriate prosody for the given text. In principle, these scales can be defined in both time and frequency axes of the

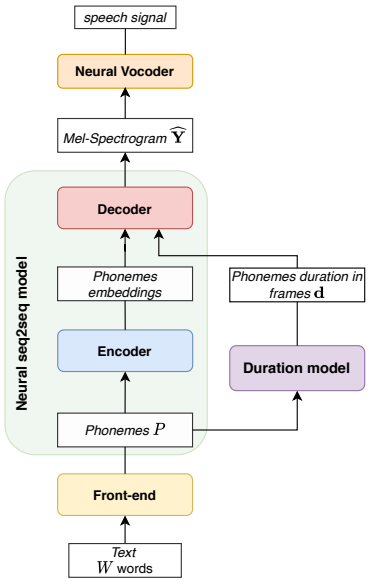


Figure 1: *Block diagram showing the architecture of our baseline. The decoder module colored in red is substituted by a multi-scale decoder in the proposed MSS modelling technique.*

mel-spectrogram. However, in this paper, the scales are defined only along the time axis while keeping the number of mel-bins constant ($= 80$) along the frequency axis. Extending the scales to the frequency axis is left as future work.

3.1. Generic multi-scale representations

Before moving to modelling, we first discuss how to construct targets for learning a generic multi-scale model. Let us assume that there are in total $L+1$ scales in the MSS modelling. At each scale l where $0 \leq l \leq L$, we compute a mel-spectrogram $\mathbf{S}^l = [\boldsymbol{s}_0^l, \boldsymbol{s}_1^l, \ldots, \boldsymbol{s}_{N_l-1}^l]$ of dimension $N_l \times 80$ from the ground-truth mel-spectrogram $\mathbf{Y} = [\boldsymbol{y}_0, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_{T-1}]$ of dimension $T \times 80$. Here, the L^{th} scale (the highest level representation of speech) has the least number of mel-spectrogram vectors, and their number progressively increases on each subsequent scale

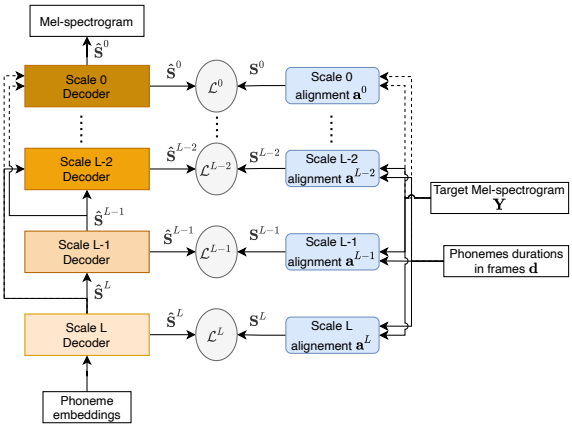


Figure 2: *Block diagram of the generic MSS decoder with $L+1$ scales. The scale 0 decoder depends upon the outputs of all the previous scale decoders.*

$l < L$ until the last scale 0 such that $N_L < N_{L-1} < \cdots < N_0 = T$. For example in Sentence-level MSS, $N_L = 1$, which is the number of sentences in an utterance.

We compute the mel-spectrogram $\mathbf{S}^l$ at scales $l > 0$ using the following equation:

$$\mathbf{S}^l = [\boldsymbol{s}_0^l, \boldsymbol{s}_1^l, \ldots, \boldsymbol{s}_{N_l-1}^l],$$

where each

$$\boldsymbol{s}_i^l = \begin{cases} \dfrac{1}{a_i^l} \displaystyle\sum_{j=c_{i-1}^l}^{c_i^l} \boldsymbol{y}_j, & 1 \leq i < N_l, \\ \dfrac{1}{a_i^l} \displaystyle\sum_{j=0}^{c_i^l} \boldsymbol{y}_j, & i = 0, \end{cases} \tag{1}$$

$$c_k^l = \sum_{i=0}^{k} a_i^l \tag{2}$$

In Eq. 1, $\boldsymbol{a}^l = [a_0^l, a_1^l, \ldots, a_{N_l-1}^l]$ is an alignment vector that denotes the alignments at scale $l > 0$ where each a_i^l represents the number of mel-spectrogram frames of $\mathbf{Y}$ corresponding to the spectrogram $\boldsymbol{s}_i^l$ at scale l. Thus the total sum of $\sum \boldsymbol{a}^l = T$. The alignment vectors a_i^l can be computed based on the definition of each scale (cf. Section 3.2). The vector $\boldsymbol{c}^l = [c_0^l, c_1^l, \ldots, c_{N_l-1}^l]$ is the cumulative sum of alignment vector $\boldsymbol{a}^l$, and each element c_k^l corresponds to the starting frame for token $k-1$ and ending frame for token k. So each target vector $\boldsymbol{s}_i^l$ at $l > 0$ is computed by taking the mean of mel-spectrogram frames $\mathbf{Y}$ from index c_{i-1}^l to c_i^l. The mel-spectrogram $\mathbf{S}^0$ at 0th scale of MSS modelling technique is equal to the target mel-spectrogram $\mathbf{Y}$ i.e. $\mathbf{S}^0 = \mathbf{Y}$, thus $N_0 = T$. This paper considers two specific cases of MSS modelling technique for validating our proposed approach: 1) *Sentence-level MSS* and 2) *Word-level MSS*, named after the coarser used linguistic unit.

3.2. Word-level MSS and Sentence-level MSS

The Word-level MSS has a total number of three scales ($L = 2$) where the 1st and 2nd scales correspond to the linguistic unit of phonemes and words respectively. The 0th scale corresponds to frame-level mel-spectrogram as discussed above. The number of mel-spectrogram vectors at each scale are given as such: N_2 = number of words in a given utterance, N_1 = number of phonemes present in an utterance, and $N_0 = T$. In Sentence-level MSS, $L = 3$ and there is an additional 3rd scale that corresponds to sentences. The N_3 = number of sentences in a given utterance, which is equal to 1 in our case. In both these specific systems, the alignment vectors $\boldsymbol{a}^l$ are obtained from the phoneme durations and relations between phoneme, word, and sentences, which are obtained by the front-end. More specifically, in Sentence-level MSS, the alignment vector $\boldsymbol{a}^1$ is equal to phoneme durations $\boldsymbol{d}$ in frames. The alignment vector $\boldsymbol{a}^2$ is equal to the word durations in frames and $\boldsymbol{a}^3$ is equal to the sentence duration in frames.

The multi-scale representations that need to be modelled in Sentence-level MSS are $\mathbf{S} = [\mathbf{S}^0, \mathbf{S}^1, \mathbf{S}^2, \mathbf{S}^3]$. To model 3rd scale (sentence-level) mel-spectrogram $\hat{\mathbf{S}}^3$, we first project P phoneme embeddings into a sentence-level vector by taking the last hidden state of the LSTM encoder. Later, the sentence-level vector is passed through a sequence of 1D convolutions to obtain sentence-level mel-spectrogram $\hat{\mathbf{S}}^3$. The loss function at the 3rd scale (sentence-level) is defined as:

$$\mathcal{L}^3 = \left\| \hat{\mathbf{S}}^3 - \mathbf{S}^3 \right\|_2 \tag{3}$$

The $\hat{\mathbf{S}}^3$ mel-spectrogram is assumed to capture the sentence-level acoustic properties such as speaker-identity, recording environment, or speaking style of the sentence.

Similarly, to model the 2nd scale (word-level) mel-spectrogram $\hat{\mathbf{S}}^2$, we first project phonemes of each word into a word-level vector. Later, the word-level vectors are concatenated with the upsampled sentence-level mel-spectrogram $\hat{\mathbf{S}}_\uparrow^3$. $\hat{\mathbf{S}}_\uparrow^3$ is computed by upsampling the predicted sentence-level mel-spectrogram $\mathbf{S}^3$ to have the same dimension as $\mathbf{S}^2$ using an alignment defined between 2nd and 3rd scale, i.e. between words and the sentence they belong to. The loss function on 2nd scale is defined as:

$$\mathcal{L}^2 = \left\| \hat{\mathbf{S}}^2 - \mathbf{S}^2 \right\|_2 \tag{4}$$

$\hat{\mathbf{S}}^2$ is assumed to capture the word level acoustic properties such as word prominence, rise and fall of pitch and energy at word level.

We follow the same procedure to predict the phoneme-level mel-spectrogram $\hat{\mathbf{S}}^1$ and final frame-level mel-spectrogram $\hat{\mathbf{S}}^0$ or $\hat{\mathbf{Y}}$. In each of the scales l, the predicted mel-spectrograms are also conditioned on all the coarser scale predicted mel-spectrograms. The loss function at each scale l is defined as $\mathcal{L}^l = \left\| \hat{\mathbf{S}}^l - \mathbf{S}^l \right\|_2$. The sentence-level MSS is trained by minimizing the loss at all scales. So the training loss for the whole system is defined as:

$$\mathcal{L} = \mathcal{L}^0 + \mathcal{L}^1 + \mathcal{L}^2 + \mathcal{L}^3, \tag{5}$$

which can be interpreted as maximizing the likelihood of mel-spectrograms at all scales:

$$\begin{aligned} p(\hat{\mathbf{S}}^3, \hat{\mathbf{S}}^2, \hat{\mathbf{S}}^1, \hat{\mathbf{S}}^0) = & p(\hat{\mathbf{S}}^3).p(\hat{\mathbf{S}}^2|\hat{\mathbf{S}}^3).p(\hat{\mathbf{S}}^1|\hat{\mathbf{S}}^3, \hat{\mathbf{S}}^2) \\ & .p(\hat{\mathbf{S}}^0|\hat{\mathbf{S}}^3, \hat{\mathbf{S}}^2, \hat{\mathbf{S}}^1) \end{aligned} \tag{6}$$

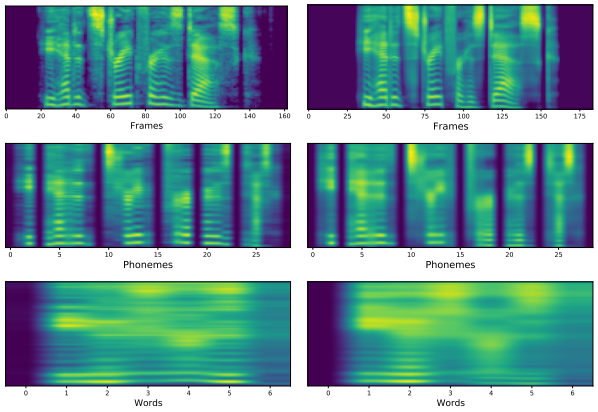
4. Experiments

4.1. Data

The experiments were carried out on an internal voice dataset that was recorded by two native US-English female voice talents. We refer to them as speaker-A and speaker-B. The training and test sets for speaker-A are 33 and 5.5 hours respectively, while they are 32 and 3 hours for speaker-B respectively. The test set is reserved for testing in this and future research studies. The sampling rate of the recorded audio is 24kHz. We extracted 80 band mel-spectrograms with a frame shift of 12.5ms.

4.2. Training and inference

As mentioned in Section 3.2, we have developed two systems based on MSS modelling technique: Sentence-level MSS and Word-level MSS. In both systems: 1) we train the seq2seq model according to the loss defined in Equation 5 where oracle alignments $\boldsymbol{a}^l$ are provided to the model at each scale l ; 2) we train the duration model using L2 loss as shown in Section 2. We optimize the loss in both the models using Adam optimizer [24] with different learning rates. We use a learning rate of 10^{-3} and 10^{-4} for the acoustic and the duration model respectively.



(a) *Oracle spectrogram* (b) *Predicted spectrogram*

Figure 3: *Visualisation of mel-spectrograms at different scales in Word-level MSS given the text: "He headed straight for his desk.". Left panel: oracle spectrograms, right panel: predicted spectrograms. The mel-spectrograms in bottom, middle, and top row correspond to 2nd (word), 1st (phoneme), and 0th (frame) scale respectively.*

During inference, we follow these 2 steps in the following order: I) we predict durations $\hat{\boldsymbol{d}}$ from the duration model trained in step 2; II) we generate mel-spectrograms $\hat{\mathbf{S}}^0$ from the seq2seq model trained in step 1 using the predicted durations $\hat{\boldsymbol{d}}$ from step I. We use Wavenet vocoder [1] to synthesize speech from $\hat{\mathbf{S}}^0$.

4.3. Evaluations

4.3.1. Qualitative evaluation of predicted and target spectrograms on different scales

Figure 3 shows the target (left column) and predicted (right column) mel-spectrograms at different scales l in the Word-level MSS for speaker-B as an example. The bottom row shows the 2nd scale (word-level) mel-spectrogram $\mathbf{S}^2$. There are 7 mel-spectrogram vectors $[\boldsymbol{s}_0^2, \boldsymbol{s}_1^2, \ldots, \boldsymbol{s}_6^2]$ which represent coarse-grained prosodic features such as prominence at word-level. We can observe the harmonics and energy, and how they vary after each word. The middle row shows the phoneme-level mel-spectrogram $\mathbf{S}^1$. On this scale, there are 29 mel-spectrogram vectors $[\boldsymbol{s}_0^1, \boldsymbol{s}_1^1, \ldots, \boldsymbol{s}_{28}^1]$ which represent fine-grained prosodic features at phoneme-level. At this scale, we can see how the prosody varies around phonemes and can identify the stress on phonemes based on their acoustic representations. The top row shows frame-level mel-spectrogram $\mathbf{S}^0$, which is equal to the target spectrogram $\mathbf{Y}$. When comparing the target (left column) to the predicted (right column) mel-spectrograms, the Word-level MSS is able to capture the high-level prosodic features at 1st and 2nd scale, albeit with a smoother representation due to the nature of the L2 loss used in eq. 5.

4.3.2. Ablation study

A MUSHRA [25] evaluation was conducted on speaker-A to evaluate how the number/definition of scales affects the performance of MSS modelling. For the MUSHRA evaluation, we selected the following three systems: the baseline from Section 2, Word-level MSS, and Sentence-level MSS. A total of 50 utterances were selected randomly from the test set, and the duration of each utterance was approximately 15 seconds. Each

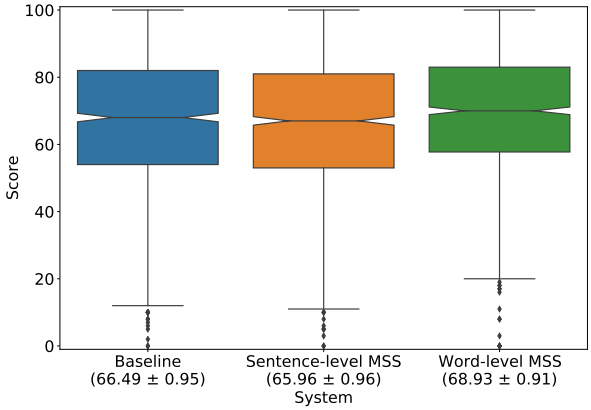


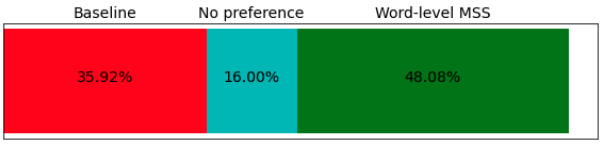
Figure 4: *Results of the MUSHRA evaluation of ablation study on speaker-A voice. Mean rating and 95% confidence intervals are reported below system names.*

utterance was rated by 24 native US-English professional listeners. Figure 4 presents the results of the MUSHRA evaluation. We have used a pairwise two-sided Wilcoxon signed-rank test corrected for multiple comparisons to measure statistical significance between the systems. The Word-level MSS system performed statistically significantly better than both the other systems (p-value $< 10^{-6}$) while there was no statistically significant difference between the baseline and sentence-level MSS systems (p-value $= 0.27$). We believe that during training, the Sentence-level MSS system is overfitting on the sentence-level spectrogram $\hat{\mathbf{S}}^3$ prediction. Due to this, it fails to capture coarse-grained prosodic features observed in the target $\mathbf{S}^3$, thus adversely affecting the prediction of spectrograms in lower scales $l < L$ which are conditioned on $\hat{\mathbf{S}}^3$. Moreover, an utterance in our data corresponds to a sentence which makes the prediction of $\hat{\mathbf{S}}^{\mathbf{3}}$ even more difficult because it does not have the surrounding sentences as an input to the system unlike scales $l < L$. These reasons could suggest why there is no improvement in the Sentence-level MSS system compared to the baseline.

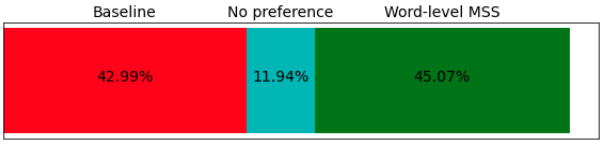
4.3.3. Preference tests

As the Word-level MSS system showed a significant improvement in the ablation study, we have selected it for further comparisons with the baseline system. A preference test was conducted on speaker-A and speaker-B voices. For speaker-A, 50 utterances were selected randomly from the test set and each utterance had a duration of approximately 15 seconds. Similar to the MUSHRA evaluation, we used a third-party vendor to complete the test, and a total 24 listeners participated. The results are shown in Figure 5a. We use a binomial significance test to measure the statistical significance. The Word-level MSS is found to be statistically significantly better than the baseline (p-value $< 10^{-4}$)

For speaker-B, 100 utterances were selected randomly from the test set and the duration of each utterance was approximately 15 seconds. A total of 48 subjects participated in the preference test. However, this evaluation was conducted via Clickworker platform - a crowdsourced evaluation, unlike earlier evaluations. The results of the preference test are shown in Figure 5b. There was not a statistically significant preference for the Word-level MSS system (p-value=0.068). However, upon removing



(a) *speaker-A*



(b) *speaker-B*

Figure 5: *The results of preference tests between the Word-level MSS and the baseline system on two voices.*

listeners that had low reliability because they did not listen completely to both samples, we found that the difference becomes significant, i.e. Word-level MSS is preferred statistically significantly (p-value=0.007). The results of both MUSHRA and preference tests suggest that the Word-level MSS system is able to produce more natural speech than the baseline system. We observe that the Word-level MSS has more contextually appropriate emphasis on the words and generally better intonation without impacting the segmental quality. Although the Word-level MSS system is preferred over the baseline system, we do note that on certain utterances it is unable to produce the right kind of intonation, specifically when a question does not start with an interrogative word.

5. Conclusion

In this paper, we presented a novel method for multi-scale modelling of mel-spectrograms to improve the quality of NTTS systems. We presented a generic MSS modelling approach and later provided details for its two specific versions called Sentence-level MSS and Word-level MSS where the scales correspond to the linguistic units. The ablation study showed that the Word-level MSS system performed statistically significantly better than Sentence-level MSS. In the preference evaluations on 2 voices, the Word-level MSS system showed statistically significantly better results than the baseline system. In the future, we want to introduce scales in MSS along the frequency axis as well which could result in an even improved segmental quality. We also want to extend the sentence-level MSS to broader linguistic units for a better modelling of the coarse-grained prosody of speech. Furthermore, we want to introduce another scale that corresponds to syllables as they are strongly linked to prosodic events like stress and intonation [26].

6. References

[1] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[2] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017.

[3] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[4] 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 *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[5] Y. Ning, S. He, Z. Wu, C. Xing, and L.-J. Zhang, "A review of deep learning based speech synthesis," *Applied Sciences*, vol. 9, no. 19, p. 4050, 2019.

[6] A. Curtin and H. Ayaz, "Cognitive considerations in auditory user interfaces: Neuroergonomic evaluation of synthetic speech comprehension," in *International Conference on Engineering Psychology and Cognitive Ergonomics*. Springer, 2017, pp. 106–116.

[7] R. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. Weiss, R. Clark, and R. A. Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," in *international conference on machine learning*. PMLR, 2018, pp. 4693–4702.

[8] 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," *arXiv preprint arXiv:2011.02252*, 2020.

[9] 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," *arXiv preprint arXiv:2011.01175*, 2020.

[10] G. Sun, Y. Zhang, R. J. Weiss, Y. Cao, H. Zen, A. Rosenberg, B. Ramabhadran, and Y. Wu, "Generating diverse and natural text-to-speech samples using a quantized fine-grained VAE and autoregressive prosody prior," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6699–6703.

[11] K. Akuzawa, Y. Iwasawa, and Y. Matsuo, "Expressive Speech Synthesis via Modeling Expressions with Variational Autoencoder," *Proc. Interspeech 2018*, pp. 3067–3071, 2018.

[12] W.-N. Hsu, Y. Zhang, R. J. Weiss, H. Zen, Y. Wu, Y. Wang, Y. Cao, Y. Jia, Z. Chen, J. Shen *et al.*, "Hierarchical generative modeling for controllable speech synthesis," *arXiv preprint arXiv:1810.07217*, 2018.

[13] T. Kenter, V. Wan, C.-A. Chan, R. Clark, and J. Vit, "CHiVE: Varying prosody in speech synthesis with a linguistically driven dynamic hierarchical conditional variational network," in *International Conference on Machine Learning*. PMLR, 2019, pp. 3331–3340.

[14] C.-M. Chien and H.-y. Lee, "Hierarchical Prosody Modeling for Non-Autoregressive Speech Synthesis," *arXiv preprint arXiv:2011.06465*, 2020.

[15] S. Tyagi, M. Nicolis, J. Rohnke, T. Drugman, and J. Lorenzo-Trueba, "Dynamic prosody generation for speech synthesis using linguistics-driven acoustic embedding selection," *arXiv preprint arXiv:1912.00955*, 2019.

[16] T. Raitio, R. Rasipuram, and D. Castellani, "Controllable neural text-to-speech synthesis using intuitive prosodic features," *arXiv preprint arXiv:2009.06775*, 2020.

[17] 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. [Online]. Available: https://openreview.net/forum?id=piLPYqxtWuA

[18] T. Drugman, G. Huybrechts, V. Klimkov, and A. Moinet, "Traditional machine learning for pitch detection," *IEEE Signal Processing Letters*, vol. 25, no. 11, pp. 1745–1749, 2018.

[19] S. Vasquez and M. Lewis, "Melnet: A generative model for audio in the frequency domain," *arXiv preprint arXiv:1906.01083*, 2019.

[20] S. Ruder, "An overview of multi-task learning in deep neural networks," *arXiv preprint arXiv:1706.05098*, 2017.

[21] M. He, Y. Deng, and L. He, "Robust Sequence-to-Sequence Acoustic Modeling with Stepwise Monotonic Attention for Neural TTS," in *Proc. Interspeech 2019*, 2019, pp. 1293–1297. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-1972

[22] Z. Zeng, J. Wang, N. Cheng, T. Xia, and J. Xiao, "AlignTTS: Efficient feed-forward text-to-speech system without explicit alignment," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6714–6718.

[23] C. Yu, H. Lu, N. Hu, M. Yu, C. Weng, K. Xu, P. Liu, D. Tuo, S. Kang, G. Lei *et al.*, "Durian: Duration informed attention network for multimodal synthesis," *arXiv preprint arXiv:1909.01700*, 2019.

[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[25] B. Series, "Method for the subjective assessment of intermediate quality level of audio systems," *International Telecommunication Union Radiocommunication Assembly*, 2014.

[26] J. Itô, *Syllable theory in prosodic phonology*. Routledge, 2018, vol. 10.