A Non-Textual Approach to Modelling Expressive Speech

Jianzhi Wang

December 20 2024

1 Introduction

Given a Mel spectrogram that corresponds to a monotonous audio $(M_t^A)_t$ and a label, say "happy", can it be converted into another Mel spectrogram that corresponds to an expressive audio $(M_t^H)_t$? For the rest of the report, $(M_t^A)_t$ will be referred to as a monotonous Mel spectrogram, while $(M_t^H)_t$ will be referred to as an expressive Mel spectrogram.

If the above process is possible, one can imagine a speech synthesis pipeline that takes a monotonous audio $(A_t^A)_t$ (for example, one obtained from a Text-to-Speech (TTS) system) and a context as input, converts it to a monotonous Mel spectrogram $(M_t^A)_t$, performs the transformation to an expressive Mel spectrogram $(M_t^H)_t$, and then obtains an expressive audio $(A_t^H)_t$ at the end (Figure 1).

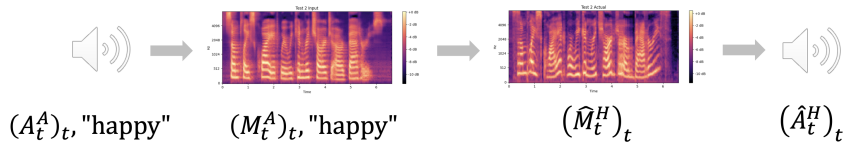

Figure 1: The pipeline from a (monotonous audio, label) pair $\{(A_t^A)_t$, "happy"$\}$ to Mel spectrogram space $((M_t^A)_t \to (\hat{M}_t^H)_t)$ and back to an expressive audio $(\hat{A}_t^H)_t$. The superscripts A and H represent "AI" and "Human" (the paragons of monotonous and expressive speeches) respectively.

This problem falls under the intersection of speech synthesis and voice conversion, with two unique features. Firstly, the project will focus on reducing the dependency on textual elements, as opposed to many existing speech synthesis models which take in texts as some form of input ([1], [2], [3]). The rationale is that a person can reasonably distinguish emotions directly from speech without a transcript. Baby sounds and spoken languages without a writing system are

further evidence against the necessity of textual elements. Hence, keeping this pipeline entirely within the Mel spectrogram space is a way of asserting that belief.

The second unique feature is that the starting point of the pipeline is a monotonous Mel spectrogram. Classical voice conversion focuses on two distinct participants (say A and B), while this project focuses on standardising one of the participant (where A is the source of the monotonous speech). This allows us to visualise how emotions change the energy levels with respect to a monotonous Mel spectrogram. Furthermore, classical voice conversion can still be achieved with this model if the inverse problem is also solved.

In this project, I first explore numerous variants of the transformer encoder architecture [4] to model expressive Mel spectrograms. The key finding is that the naive transformer with MSE loss produce low-quality (blurry) Mel spectrograms of unacceptable naturalness and intelligibility. Furthermore, this effect is not alleviated by increasing or decreasing the complexity of the model, GAN loss, soft-DTW loss, CNN layers nor using an additional duration.

Secondly, taking the best out of all the models tried, I propose Masque [1], a light-weight transformer encoder model that leverages SPARC features to inject emotions into monotonous Mel spectrograms and thereby generate intelligible and expressive audios. Masque also takes in a "context" which serves as a knob for more quantitative control in the intonation and prosody of the synthesised speech. It performs close to, but still below, human parity. The finding is that SPARC features are better for emotion modelling compared to Mel spectrogram energy levels. A subsequent experiment verifies that the SPARC speaker embedding is relatively uncorrelated with emotion labels, as expected from its definition. Masque also enables downstream applications including emotion switching and voice conversion with standardised speech as an intermediary.

2 Related Work

Recent speech synthesis models are highly impressive, such as VALL-E 2 ([1], [2]) which achieves human-level performance and Audiobox [3], capable of generating various audio modalities. However, there is still a lack of quantitative control in the intonations and prosody of the generated speech. Moreover, they take in texts as some form of input, be it for transcript [1] or for a description of target voice [3]. This prompts the philosophical question of whether humans require a written system to generate speeches of various emotions.

Another related body of work is CycleGAN-VC [5], whose focus is on voice-conversion. This project differs from CycleGAN-VC's methodology because it focuses on emotion injection. Furthermore, it is a constrained voice-conversion problem, where starting point is a monotonous Mel spectrogram instead of an arbitrary speaker.

[1] https://github.com/jianzhi-1/masque

3 Methods

3.1 Data

The supervised learning setup of this project requires pairs of monotonous audio and expressive audio. Expressive audio are obtained from the EXPRESSO [6] dataset. It comprises 4 distinct speakers, 40 hours of speaking time, and 11 styles ("confused", "default", "emphasis", "enunciated", "essentials", "happy", "laughing", "longform", "sad", "singing", "whisper"). To maintain target speaker invariance, only data from speaker 4 is used. The train-test split is: $n_{\text{train}} = 2000$, $n_{\text{valid}} = 450$, and $n_{\text{test}} = 453$. The distribution of emotion labels provides approximately 20 minutes of training data for each label (Figure 2).

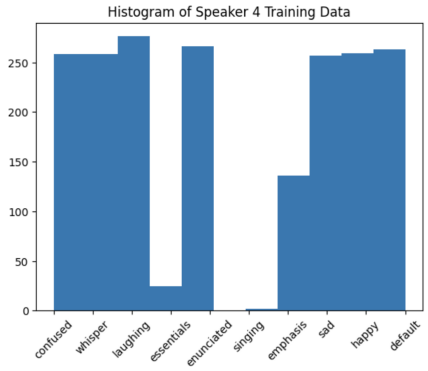

Figure 2: A histogram of emotion labels in the training set.

The input to the data processing pipeline (Figure 3) comprises the human audio waveform $(A_t^H)_t$ and its corresponding transcript. Keeping the research goal in mind, the only use of the transcript is to generate the corresponding standardised audio $(A_t^A)_t$ using a text-to-speech (TTS) system. After which, the transcript is discarded, so as to keep the pipeline solely in Mel spectrogram space. For this experiment, Tacotron2 [7] is used as the TTS system. Speaker 4 is of the same gender and sounds similar to the audio produced by Tacotron2.

The Mel spectrogram of both audio signals $((M_t^H)_{t=1}^{T'}, (M_t^A)_{t=1}^{T})$ are then obtained. HiFi-GAN [7] is chosen as the Mel spectrogram-to-waveform converter, so the selected number of Mel features is `n_mels=80`, which corresponds to the number of features expected by HiFi-GAN.

The vanilla transformer encoder architecture expects the input and output sequences to be of the same length. To solve this alignment issue, the dynamic time warping (DTW) algorithm is applied to the expressive Mel spectrogram $(M_t^H)_{t=1}^{T'}$ with respect to the monotonous Mel spectrogram $(M_t^A)_{t=1}^{T}$. The rationale is to preserve the quality of the model input over the target. Following the algorithm, the output of the data processing pipeline are two Mel spectrograms of the same shape $\mathbb{R}^{80\times T}$ where T is the sequence length of the monotonous Mel

spectrogram.

DTW inevitably results in some perceptual artifacts induced by the artificial collapse and expansion of the Mel spectrograms frames. However, the outputs are still reasonable enough to be used as data (Figure 4). Efforts were made to tackle the alignment problem in other ways, such as by using the soft-DTW loss [8] and by using an additional duration model. By comparison, the naive DTW is still the most computationally efficient and produces similar results.

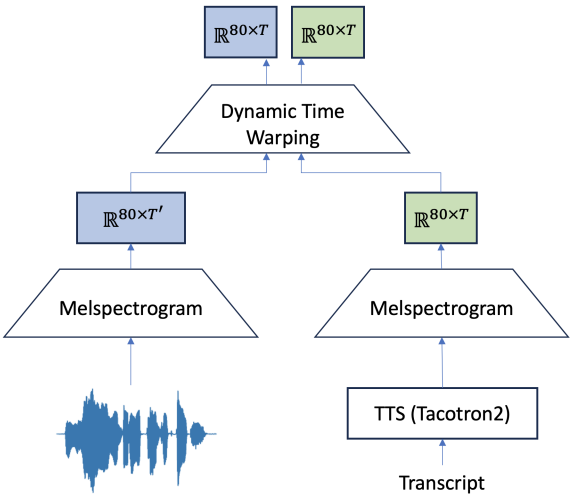

Figure 3: Data processing pipeline. Inputs: expressive audio $(A_t^H)_{t=1}^{T'}$ and its transcript. Output: a pair of Mel spectrograms of the same shape $(M_t^H)_{t=1}^{T}$, $(M_t^A)_{t=1}^{T}$

3.2 Baseline

The baseline model is a transformer encoder architecture with $52,395$ parameters, made up of $L = 6$ transformer encoder layers (`nheads=8`, `d=512`) with MSE loss (Figure 5). The input consists of the label and the monotonous Mel spectrogram $(M_t^A)_{t=1}^{T}$, embedded and projected into $\mathbb{R}^{512}$ respectively. The label is then prepended to the sequence. Afterwards, a positional embedding layer with learnable weights is applied. The resultant sequence then passes through $L = 6$ transformer encoder layers. The vector corresponding to the label position is discarded before the last projection layer. The rest of the sequence is projected back into $\mathbb{R}^{80}$ to obtain the predicted expressive Mel spectrogram $(\hat{M}_t^H)_{t=1}^{T}$ and the loss is computed with the ground truth expressive Mel spectrogram $(M_t^H)_{t=1}^{T}$ as the target.

For evaluation, the predicted Mel spectrogram $(\hat{M}_t^H)_{t=1}^{T}$ is passed through HiFi-GAN [7] to obtain the predicted expressive waveform.

3.3 Duration Model

The DTW step in the data generation process inevitably causes some unnatural audio artefacts. In fact, around 36% of the training data has at least a frame that corresponds to ≥ 20 frames in the expressive Mel spectrogram $(M_t^H)_{t=1}^{T'}$ post DTW. If those frames do not correspond to silence, they can be seen as signs of poor alignment. Hence, in this experiment, a duration model is implemented as part of the main transformer model.

The proposed duration model retains the same transformer encoder architecture. However, the targets are now $(d_t)_{t=1}^{T}$, with d_t being the number of frames in the expressive Mel spectrogram $(M_t^H)_{t=1}^{T'}$ that the frame M_t^A corresponds to in the DTW algorithm. Hence, it must hold that $\sum_{t=1}^{T} d_t = T'$, since the frames of the monotonous Mel spectrogram $(M_t^A)_{t=1}^{T}$ maps surjectively to the frames in the expressive Mel spectrogram $(M_t^H)_{t=1}^{T'}$.

For training effectiveness, the range of $\hat{d}_t$ is limited to an integer in $[0, 19]$. Any data containing a frame with $d_t \geq 20$ is discarded. For each input frame M_t^A, the duration model outputs a $\mathbb{R}^{20}$ vector with index i (0-indexed) corresponding to the logit that $\hat{d}_t = i$. The loss used is CrossEntropyLoss.

To use the duration model in prediction, the total time T' is provided in advance, as motivated by the Total-Duration-Aware technique [9]. The duration model supplies a score $s(t, d_t)$ corresponding to the logit of projecting the frame at t into d_t frames. Then, a dynamic program was executed to maximise the sum of logits. Thus, for a monotonous Mel spectrogram $(M_t^A)_{t=1}^{T}$, the dynamic program returns $(\hat{d}_t)_{t=1}^{T}$ with $\sum_{t=1}^{T} \hat{d}_t = T'$.

$$dp(t, x) = \begin{cases} -\infty & x > T' \\ 0 & t = T \text{ and } x = T' \\ \max_{d_t \in \{1,\dots,T_{\max}\}} \{dp(t+1, x+d_t) + s(t, d_t)\} & \text{else} \end{cases}$$

Finally, the same transformer encoder architecture as present in section 3.2 is used with the input monotonous Mel spectrogram modified, such that the frame at index t is padded with $\hat{d}_t - 1$ dummy vectors (Figure 6). The benefit now is that the target expressive Mel spectrogram $(M_t^H)_{t=1}^{T'}$ is temporally preserved.

3.4 GAN Loss

The GAN loss was an attempt to reduce the "blurriness" of the baseline model predictions. Prior GAN architectures were based on waveform generation [10] and hence included multi-scale and multi-period blocks [2] in the discriminator. To keep the problem in image domain, a simple discriminator architecture is proposed (Figure 7), featuring an adaptive pooling layer to handle the variable-length input.

[2] https://github.com/jik876/hifi-gan/blob/master/models.py#L128

The generator loss is shown below. The first term represents the feature loss and the second term is the standard generator loss [10]. α is a hyperparameter that influences the relative weights of the losses.

$$\mathcal{L}(G;D) = \mathbb{E}_{(M^A,M^H)\sim\mathcal{D}}\left[\left\|M^H - G\left(M^A\right)\right\|_F^2\right] + \alpha\mathbb{E}_{M^A\sim\mathcal{D}}\left[(D(G(M^A))-1)^2\right]$$

The discriminator loss is also standard [10].

$$\mathcal{L}(D;G) = \mathbb{E}_{(M^A,M^H)\sim\mathcal{D}}\left[\left(D(M^H)-1\right)^2 + \left(D(G(M^A))\right)^2\right]$$

3.5 Soft-DTW Loss [11]

As before, it was observed that the DTW algorithm is not perfect in aligning the pairs of Mel spectrograms $(M_t^A)_t$ and $(M_t^H)_t$. Hence, soft-DTW loss is introduced to allow the model to adjust to possibly better alignments during training. The loss function is a weighted combination of feature loss and the soft-DTW loss, similar to the previous section. α is a hyperparameter.

$$\mathcal{L}(M^H, M^A) = \left\|M^H - \hat{M}^A\right\|_F^2 + \alpha\ \text{sdtw}(M^H, \hat{M}^A)$$

3.6 CNN Layers and Template Kernels

To inject image locality inductive bias into the naive transformer architecture, three Conv2D layers were added between the positional embedding layer and the linear layer. Given an input image of dimensions $1 \times T \times 80$, the output image of the Conv2D layers is of dimension $64 \times \lfloor \frac{T}{8} \rfloor \times 10$ with each element having an 8×8 receptive field.

At the output, a vector quantisation layer [12] was used with 64 learnable template 8×8 kernels. The loss used remains the MSE. The templates aid the visualisation of the learnt Mel spectrogram image kernels. Another intention was that the discrete representation enables the use of cross entropy loss as an intermediate loss, which can possibly improve the quality of the predicted Mel spectrograms. However, this was not pursued because it necessitates a two-phase training, which is not possible with the limited computation resources.

3.7 SPARC [13]

To improve the quality of the predicted Mel spectrograms, a decision was made to pivot to using articulatory features instead. The SPARC architecture depended on WavLM layers ([13], [14]), which is pre-trained on unlabelled speech datasets. Hence, this pivot still satisfies the project goal of keeping the pipeline in non-textual space.

The same transformer encoder architecture can be used. The adapted architecture (Figure 8) has $52,263$ parameters and remains the same except for a

few changes. Firstly, the input and output are time series of reduced dimension $\mathbb{R}^{T\times 14}$. The 14 dimensions include 12 EMA articulatory features, a pitch feature and a loudness feature. Secondly, the SPARC decoder requires a speaker embedding as an input. As such, one can either forward the monotonous speaker embedding or the expressive speaker embedding to the SPARC decoder. A decision was made to forward the expressive speaker embedding to ensure the consistency of the speaker. Later experiments reveal that the speaker embedding is relatively uncorrelated with the emotion label, justifying this decision.

Besides the availability of a speaker embedding vector, another motivation for using SPARC features is that articulatory features are more standardised across individuals due to physical constraints. A preliminary exploration of a single SPARC features time series (Figure 9) reveals that the first 12 EMA articulatory features are already approximately standardised. The pitch feature is approximately normally distributed but not standardised. The loudness feature appears exponentially distributed. As such, both the pitch and loudness features are z-scored and an additional log-transform is applied to the loudness feature.

The loss remains as the MSE on the transformed SPARC features: $\mathcal{L}(\hat{S}^H, S^H) = \left\| S^H - \hat{S}^H \right\|_F^2$.

3.8 Subjective Evaluation

Since the SPARC-transformer model already attained an acceptable level of intelligibility and naturalness, a subjective evaluation is favoured over an automated evaluation.

Based on the applications of this model, two types of subjective evaluations were conducted. Due to time constraint, there were only $N = 3$ evaluators.

3.8.1 Type I Evaluation

The purpose of Type I evaluation is to compare the quality of SPARC-transformer model predictions with the original expressive audio. Evaluators are informed of the true label (e.g. "sad") and are presented with both the predicted audio and the expressive audio. The order of presentation is randomly shuffled. Scores of both audio are obtained and their relative ranks are aggregated.

3.8.2 Type II Evaluation

The purpose of Type II evaluation is to measure the fidelity of a SPARC-transformer model prediction with respect to its label l, without any comparison or ground truth. In this evaluation, an unseen speaker embedding is used. For each label l in the set $\{\text{happy}, \text{sad}, \text{laughing}, \text{confused}\}$, the SPARC-transformer model generates an audio $\hat{A}^{(l)}$. The evaluator is asked to infer the label l.

4 Results

4.1 Baseline

During training, it took around 30 epochs for the baseline model to reach saturation (Figure 10).

There is a significant blurring effect on the predicted Mel spectrogram (Figure 11). The content of the audio is very unnatural and barely intelligible. The conclusion is that the naive transformer model with MSE loss produces low-quality Mel spectrogram and hence low-quality speech. This is not alleviated by increasing or decreasing the number of transformer encoder layers d.

4.2 Duration Model

During training, the duration model took about 15 epochs to reach saturation (Figure 12).

However, the training loop for the main model is computationally unfeasible. The dynamic program has a time complexity of $O(20 \times T \times T')$, but the constant factor might be larger due to tensor manipulations. As such, the computation for a single batch took > 20 minutes. Thus, despite significant progress, this approach was abandoned.

4.3 GAN Loss

Since the feature loss is on the order of 10^4, α was initially picked to be around $60,000$. This training approach is not successful because the discriminator is often too strong. The generator takes more epochs for training (Figure 13) and the resultant loss remains the same order of magnitude as the baseline model. The resultant prediction was still a blurry Mel spectrogram (Figure 14).

In an effort to tackle the "strong discriminator problem", the training pipeline was modified so that the optimiser of the generator is stepped through more times than the optimiser of the discriminator. However, that did not alleviate the problem, as the generator remained stuck above a certain threshold while the discriminator quickly converged to near 0 loss.

A few other values of α are also tried, but they generally produce the same result.

4.4 Soft-DTW Loss

Initially, the magnitude of the soft-DTW loss is very similar to the feature loss. However, after a few epochs, the soft-DTW loss back-propagation resulted in the collapse of all frames except for a few, giving a "stretched" predicted Mel spectrogram (Figure 15). This was not alleviated by using other values of α.

4.5 CNN Layers and Template Kernels

The CNN-transformer had similar blurry predicted Mel spectrograms (Figure 16).

The convolution kernels (Figure 17) suggests that most of them might be sparsely used, with less than 10 of them reminiscent of Mel spectrogram image portions.

4.6 SPARC

The SPARC-transformer model predictions resulted in vastly improved Mel spectrograms (Figure 18). The produced audio is natural, intelligible and correlates with the emotion label.

4.7 Subjective Evaluation

4.7.1 Quality and Intelligibility

Out of all the trained models, only the SPARC-transformer model achieved a reasonable level of intelligibility and naturalness (Table 1). As such, evaluation efforts are focused towards it only.

Model	Description	Prediction quality
1	Transformer (baseline)	×
2	Duration + Transformer	−
3	Transformer + GAN Loss	×
4	Transformer + Soft-DTW Loss	−
5	CNN-Transformer	×
6	SPARC-Transformer	✓

Table 1: Summary of models

4.7.2 Type I Evaluation

The result of Type I evaluation (Figure 19) shows that the SPARC-transformer model predictions fall slightly short of human parity when injecting emotions. They are close but rank less than the expressive audio.

4.7.3 Type II Evaluation

The result of Type II evaluation shows that the SPARC-transformer model does extremely well in generating "sad" audio compared to the rest of the labels. During the evaluation process, many evaluators predicted "sad" for most of the samples, regardless of the true label.

5 Discussion

5.1 Summary

The objectives of the research project are achieved, but the quality of the predicted audio can be better. Nonetheless, a pipeline to inject expressivity into monotonous audio without the need for textual input is established.

5.2 Why is SPARC-transformer better?

Among the models, only the SPARC-transformer model predictions were reasonably intelligible and natural.

The most probable explanation is due to the fact that the SPARC-transformer model leveraged a pre-trained encoder and decoder. Furthermore, the SPARC speaker embedding vector is also forwarded to the output. These features were not present in any of the other models.

Another hypothesis is that the SPARC transformer model performed better due to more standardised features, as opposed to the varying energy levels of the Mel spectrogram, which resulted in better training. To verify this, the baseline model was rerun with Mel spectrograms standardised per channel across the entire training set. However, the result was still a blurry Mel spectrogram predictions (Figure 21). This is some evidence against the hypothesis. However, one must also note that SPARC features are articulatory and hence have physical meanings, so it is more proper for them to be standardised as opposed to the energy levels of the Mel spectrogram.

Another hypothesis offered at early stages of the project was that the naive transformer model is more suitable for time-series-to-time-series conversion, rather than image-to-image conversion, due to the absence of built-in locality inductive bias. However, the CNN-transformer experiments are evidence against this hypothesis.

5.3 Validity of forwarding SPARC speaker encoding

The result from Type II subjective evaluations (Figure 20) suggests that there might be a correlation between the SPARC speaker embedding and the emotion label. In particular, it might be the unseen speaker embedding that results in all predicted expressive audio to be perceived as “sad”.

To test the validity of this hypothesis, the speaker embeddings are permuted. For example, the speaker embedding corresponding to a true emotion label of “happy” may be forwarded to decode a prediction that has emotion label “sad”.

It turned out that there is perceptually no difference between audio generated by the true speaker embedding versus permuted speaker embeddings. The conclusion is that SPARC speaker embeddings are relatively uncorrelated with emotion labels.

5.4 Emotion switching experiment

The trained SPARC-transformer model enables the switching of emotions. As an experiment, given a pair of Mel spectrograms $\left(M^A, M^H_{l^*}\right)$ where l^* is the true emotion label of the expressive Mel spectrogram, all predictions $\hat{M}^H_{l'}$ are produced for other labels $l' \neq l^*$. This allows for comparison across emotions and one would expect the audio produced to be perceptually different. This was indeed verified to be the case.

5.5 Future Directions

The model can be improved in several ways: (1) have a better architecture, such as one that disables dropout on the first element or one that adds label embedding directly to each sequence elements, which force the model to depend more on the label; (2) scale up training data and leverage pre-trained models; (3) use cleverer ways of aligning the Mel spectrogram pairs.

To use the model, one also needs to tackle the inverse problem of "standardising" speech: given an expressive audio $(A^H_t)_t$, can its expressiveness be purged out to extract a monotonous audio $(A^A_t)_t$? The methodology in this project currently still relies on generating the monotonous audio from a transcript via a TTS system.

Several downstream applications of this project includes: (1) data augmentation; (2) speech classification without the need for natural language annotations.

6 Appendix

6.1 Individual Contributions

All code, research infrastructure, analysis, surveys, and report writing are done by the author.

I would like to thank a few individuals for their kind advice and suggestions: Cheol Jun Cho for the numerous office hour discussions and for providing guidance along the way; Xinghong Fu (MIT) who suggested several image-to-image translation techniques and images losses [15] to try; Rahul Shah (Berkeley) for a discussion on which schedulers to use for training GANs; Professor Gopala for the semester on speech processing.

I sincerely thank my group of friends who agreed to participate in the subjective evaluation portion of my project.

My thanks to fellow ELENG 225D classmates for listening to my presentation and their subsequent feedback. It prompted me to further investigate several aspects of the trained model, including the possible correlation between the SPARC speaker embedding and the emotion label.

6.2 Limitations

All of the work are done on the Kaggle environment and notebook. As such, there is a limitation of 30 hours of GPU usage per week and on the memory usage: 16GB of CPU and 16GB of GPU. Processing the EXPRESSO dataset to get Mel spectrograms and SPARC features often requires several rounds of notebook restarts due to the memory limit.

Using the Kaggle environment also made the development of research infrastructure more difficult. It is not possible to import libraries to Kaggle unless one sets up a mirror Git repository. Due to the convenience of iteration, I kept my codebase in Kaggle and updated the corresponding Git repository incrementally, instead of pulling the Git repository into Kaggle.

References

[1] 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, 2024.

[2] Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. Neural codec language models are zero-shot text to speech synthesizers, 2023.

[3] Apoorv Vyas, Bowen Shi, Matthew Le, Andros Tjandra, Yi-Chiao Wu, Baishan Guo, Jiemin Zhang, Xinyue Zhang, Robert Adkins, William Ngan, Jeff Wang, Ivan Cruz, Bapi Akula, Akinniyi Akinyemi, Brian Ellis, Rashel Moritz, Yael Yungster, Alice Rakotoarison, Liang Tan, Chris Summers, Carleigh Wood, Joshua Lane, Mary Williamson, and Wei-Ning Hsu. Audiobox: Unified audio generation with natural language prompts, 2023.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[5] Takuhiro Kaneko and Hirokazu Kameoka. Cyclegan-vc: Non-parallel voice conversion using cycle-consistent adversarial networks. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 2100–2104, 2018.

[6] Tu Anh Nguyen, Wei-Ning Hsu, Antony D'Avirro, Bowen Shi, Itai Gat, Maryam Fazel-Zarani, Tal Remez, Jade Copet, Gabriel Synnaeve, Michael Hassid, Felix Kreuk, Yossi Adi, and Emmanuel Dupoux. Expresso: A benchmark and analysis of discrete expressive speech resynthesis, 2023.

[7] Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris,

Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio. SpeechBrain: A general-purpose speech toolkit, 2021. arXiv:2106.04624.

[8] Keon Lee. Soft-dtw-loss, 2021.

[9] Sefik Emre Eskimez, Xiaofei Wang, Manthan Thakker, Chung-Hsien Tsai, Canrun Li, Zhen Xiao, Hemin Yang, Zirun Zhu, Min Tang, Jinyu Li, Sheng Zhao, and Naoyuki Kanda. Total-duration-aware duration modeling for text-to-speech systems, 2024.

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

[11] Marco Cuturi and Mathieu Blondel. Soft-dtw: a differentiable loss function for time-series, 2018.

[12] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning, 2018.

[13] Cheol Jun Cho, Peter Wu, Tejas S. Prabhune, Dhruv Agarwal, and Gopala K. Anumanchipalli. Coding speech through vocal tract kinematics, 2024.

[14] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, October 2022.

[15] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric, 2018.

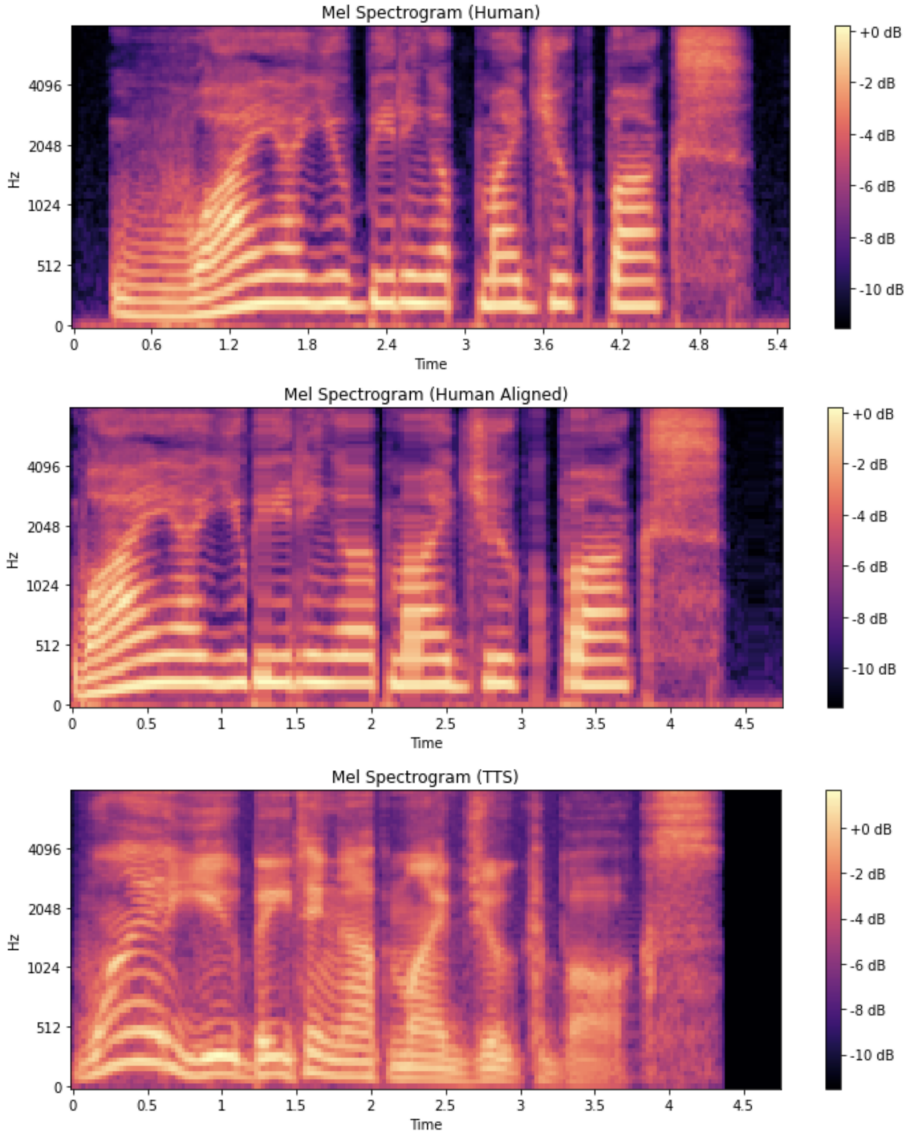

Figure 4: Samples from DTW, where the expressive Mel spectrogram $((M_t^H)_{t=1}^{T'}$; top) is aligned to monotonous Mel spectrogram $((M_t^A)_{t=1}^{T}$; bottom) to produce the aligned expressive Mel spectrogram $((M_t^H)_{t=1}^{T}$; middle).

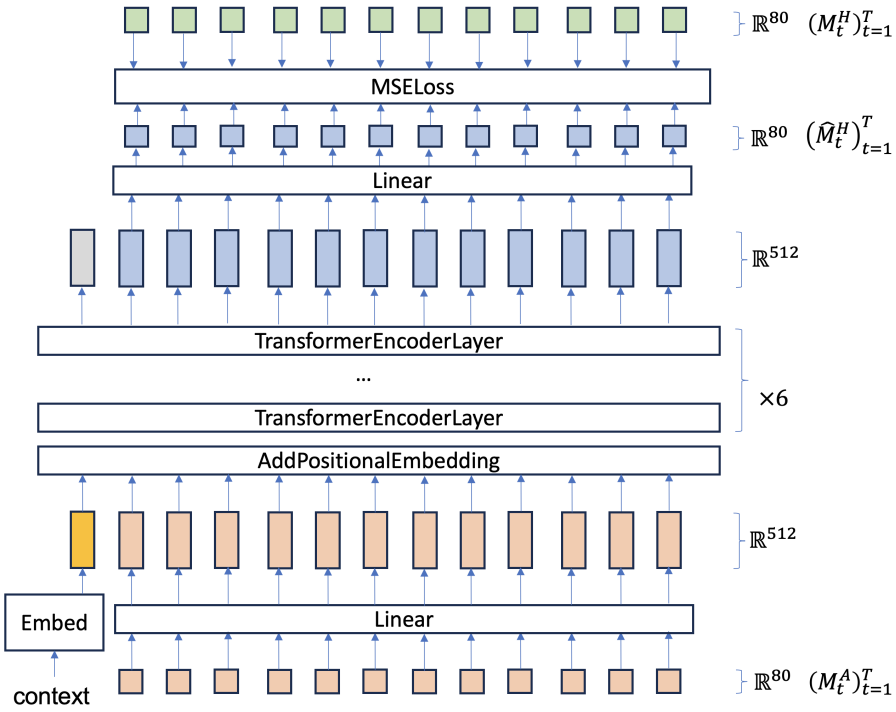

Figure 5: The baseline transformer encoder architecture. The input is the monotonous Mel spectrogram $(M_t^A)_{t=1}^T$. The target is the expressive Mel spectrogram post DTW $(M_t^H)_{t=1}^T$.

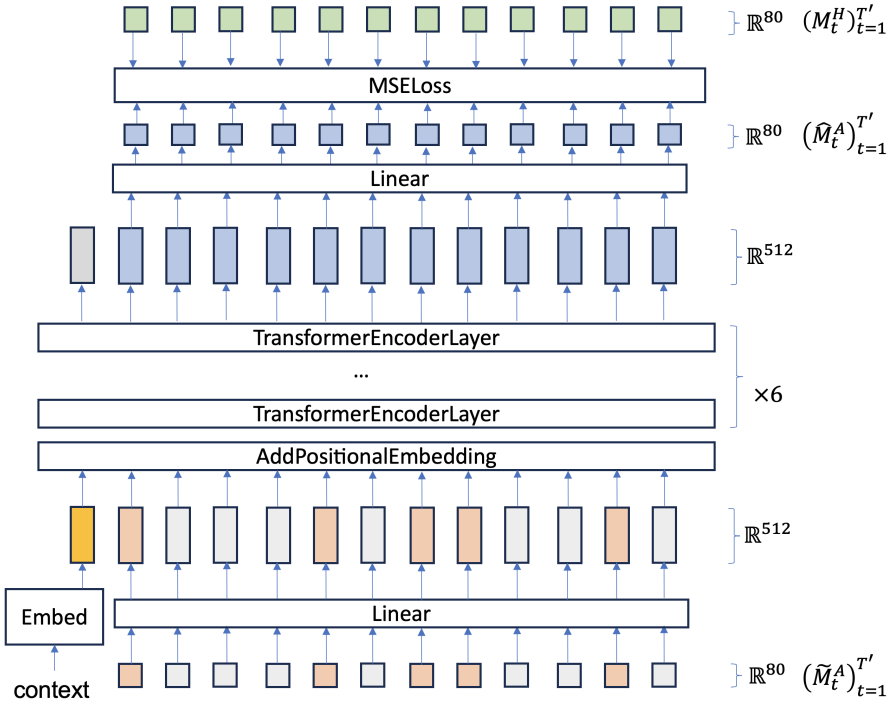

Figure 6: The transformer encoder architecture adapted for the duration model. The input vectors in grey are padded vectors due to the duration model. The target expressive Mel spectrogram $(M_t^H)_{t=1}^{T'}$ is pre-DTW.

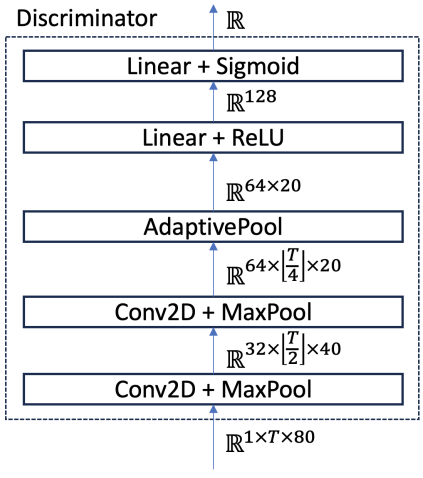

Figure 7: A simple discriminator architecture for GAN loss experiment, featuring an adaptive pool layer.

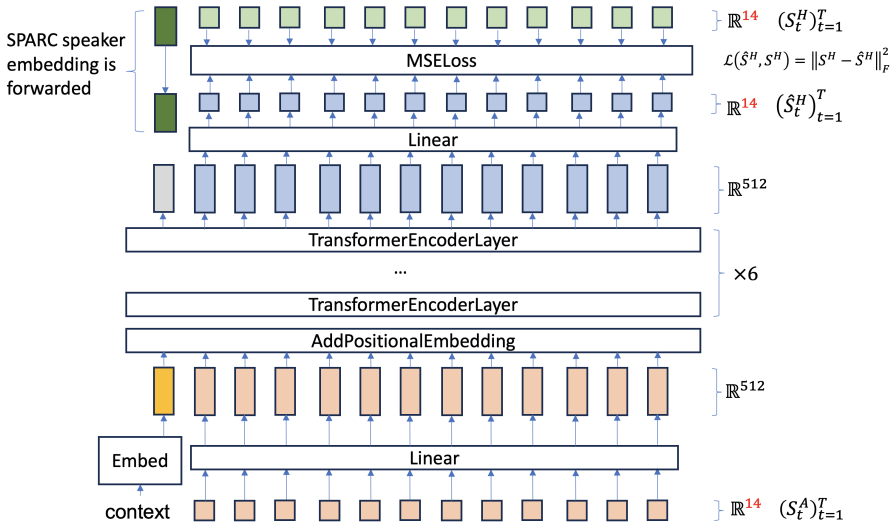

Figure 8: The SPARC-transformer architecture.

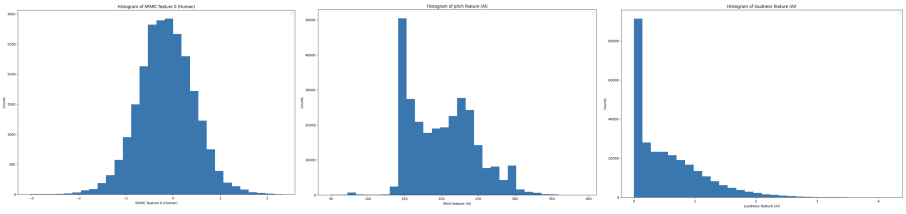

Figure 9: Histograms of SPARC feature 0 (L), pitch feature (M), loudness feature before log-transform (R).

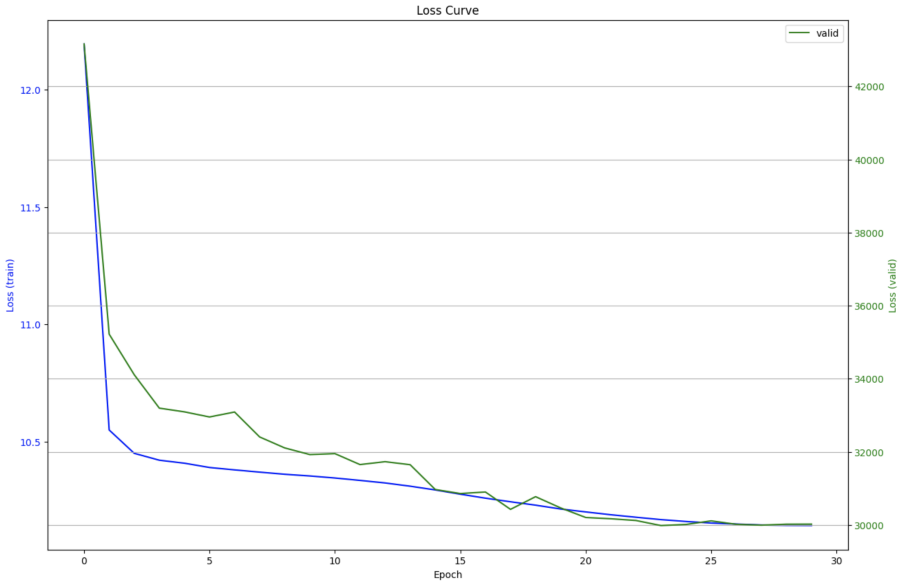

Figure 10: The loss curve of the baseline architecture. Blue: train (log); Green: validation.

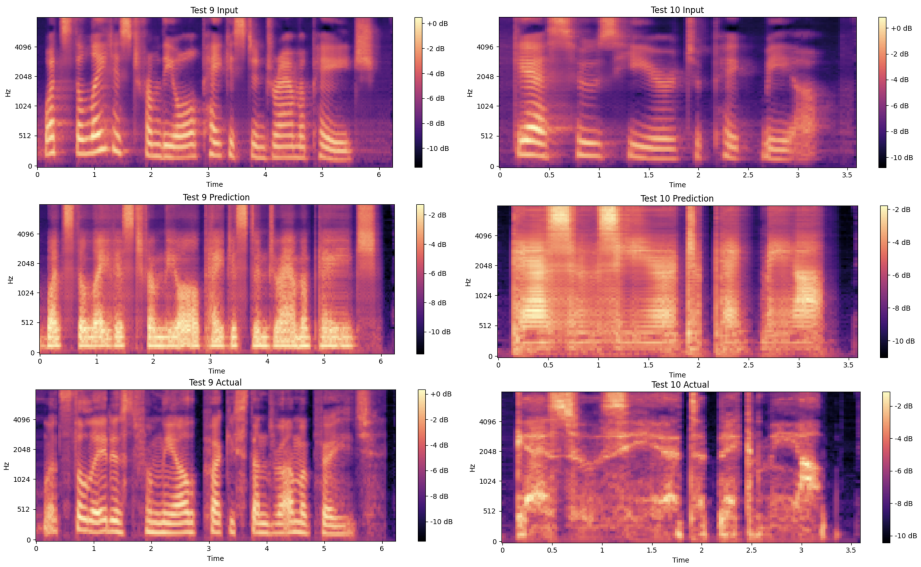

Figure 11: Samples of Mel spectrograms from the baseline model. Top: the monotonous Mel spectrogram $(M_t^A)_{t=1}^T$, which is the input; Bottom: the expressive Mel spectrogram $(M_t^H)_{t=1}^T$, which is the target; Middle: the baseline model prediction $(\hat{M}_t^H)_{t=1}^T$.

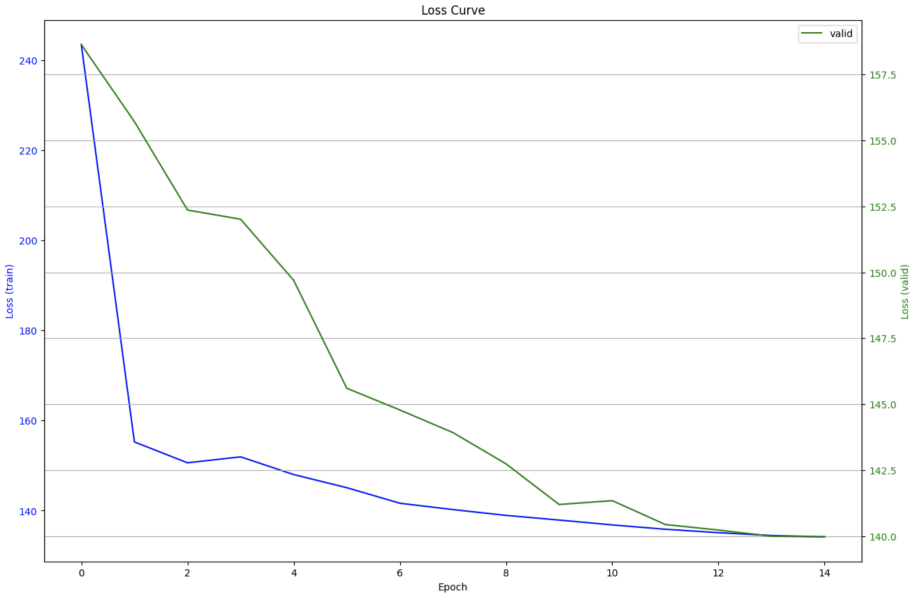

Figure 12: The loss curves for duration model. Blue: train; Green: validation.

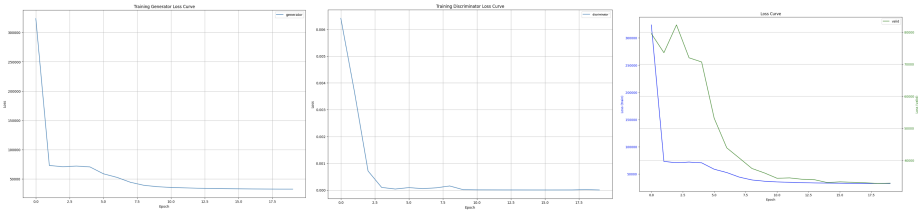

Figure 13: The GAN loss curves. L: generator (training); M: discriminator (training); R: generator (training, validation).

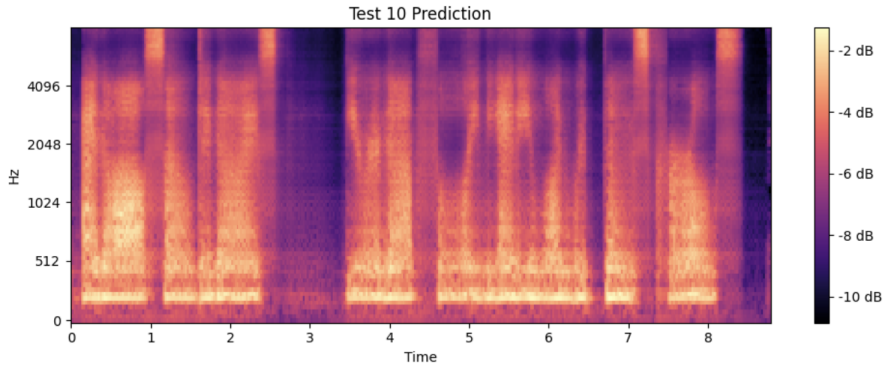

Figure 14: A sample prediction of the transformer model with GAN loss.

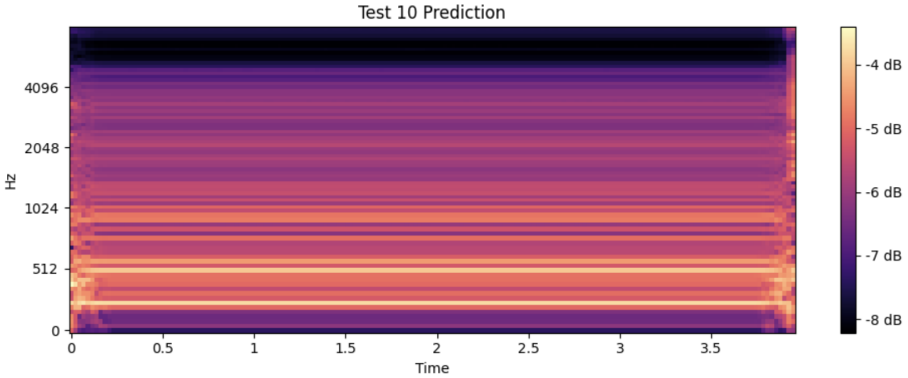

Figure 15: A sample prediction of the transformer model with soft-DTW loss.

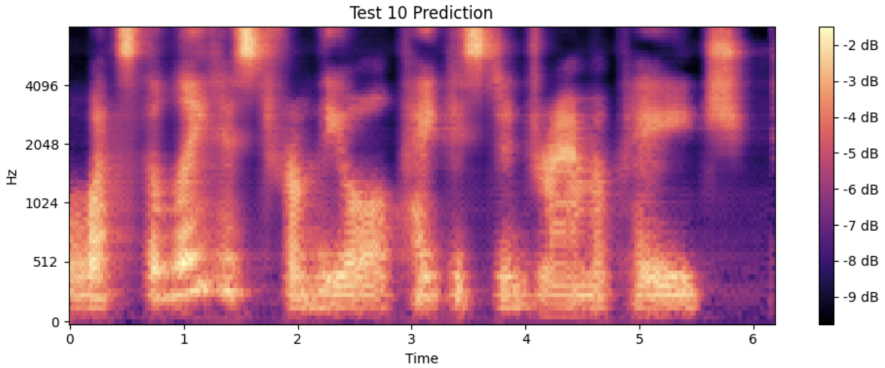

Figure 16: A sample prediction of the CNN-transformer model.

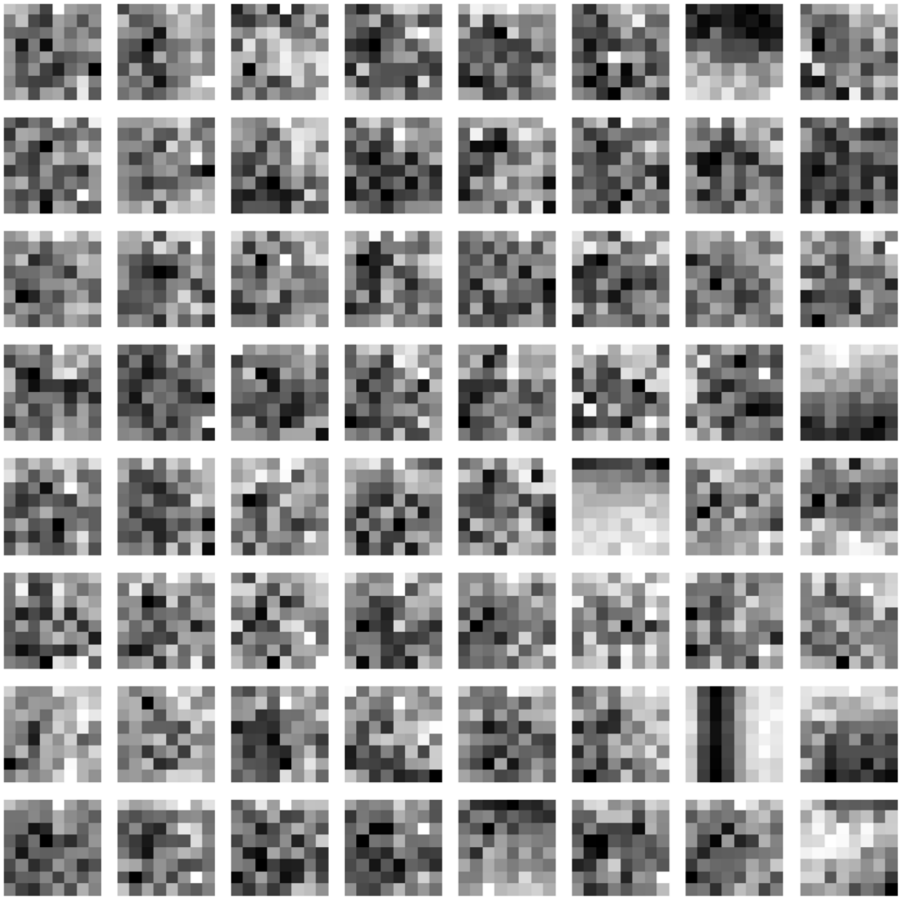

Figure 17: A visualisation of the 64 8×8 template kernels in the CNN-transformer model.

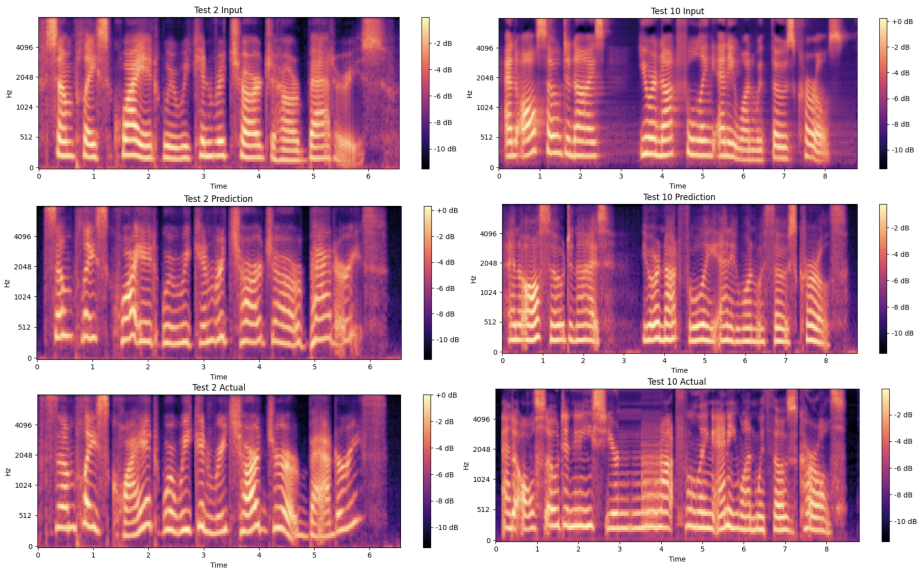

Figure 18: Samples of Mel spectrograms from SPARC-transformer model. L: Prediction for "happy" context; R: Prediction for "sad" context.

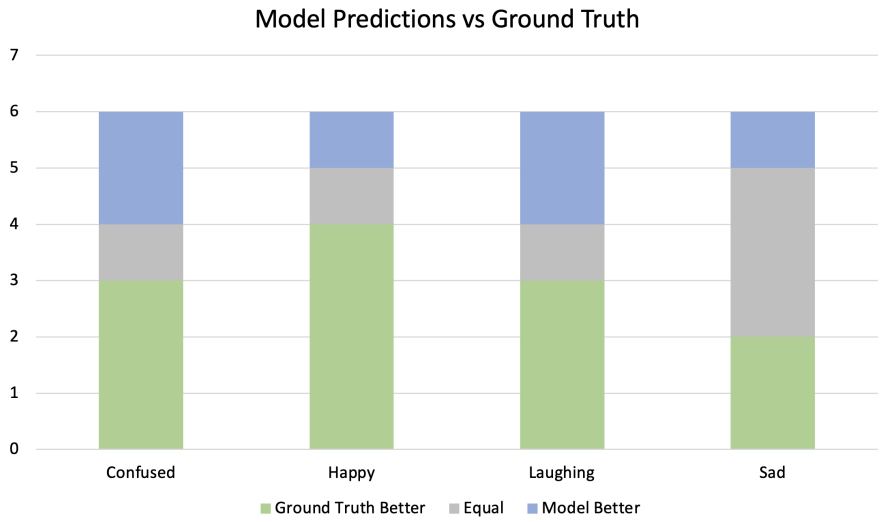

Figure 19: The result of Type I evaluation, stratified by label.

Figure 20: The result of Type II evaluation, stratified by label.

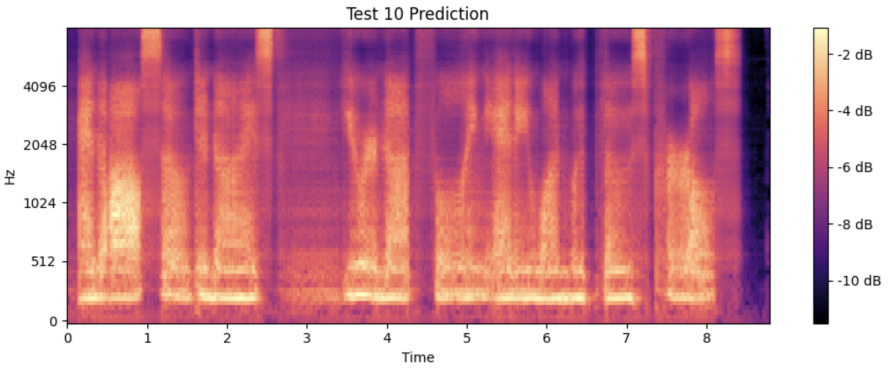

Figure 21: A sample prediction produced by the baseline model with per-channel standardised Mel spectrogram input.