A non-textual approach to modelling expressive speech

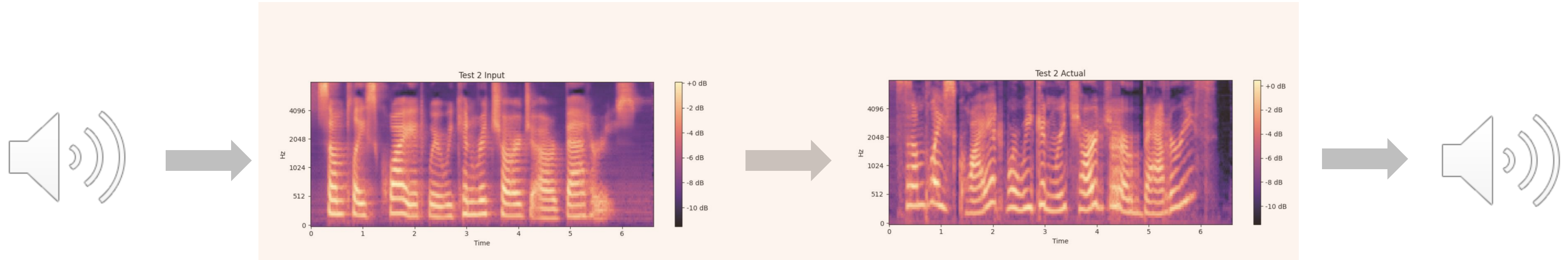
Jianzhi Wang

Agenda

- Introduction
- Methodology
- Results
- Discussion

Problem Statement

- Given a context and a Melspectrogram corresponding to monotonous audio, say $(M_t^A)_t$, label, want to produce a Melspectrogram $\left(\widehat{M}_t^H\right)_t$ corresponding to expressive audio



$(A_t^A)_t$, "happy" $(M_t^A)_t$, "happy" $\left(\widehat{M}_t^H\right)_t$ $\left(\hat{A}_t^H\right)_t$

Motivation

- See how the Mel spectrogram changes with respect to contexts
  - Monotonous audio provides a form of standardised speech to observe this effect
- The label is a “control knob” for the generated audio
- No textual data during training
  - Inductive bias that humans do not need text to decipher emotions
  - Avoids an alignment problem as well as the multimodality problem

Previous Work

- Audiobox, VALL-E
  - All of these models use some form of textual elements (e.g. natural language prompts, transcripts)
- CycleGAN-VC
  - Focus is on voice-conversion

Vyas et al, 2023; Wang et al, 2023; Kaneko et al, 2017

Data

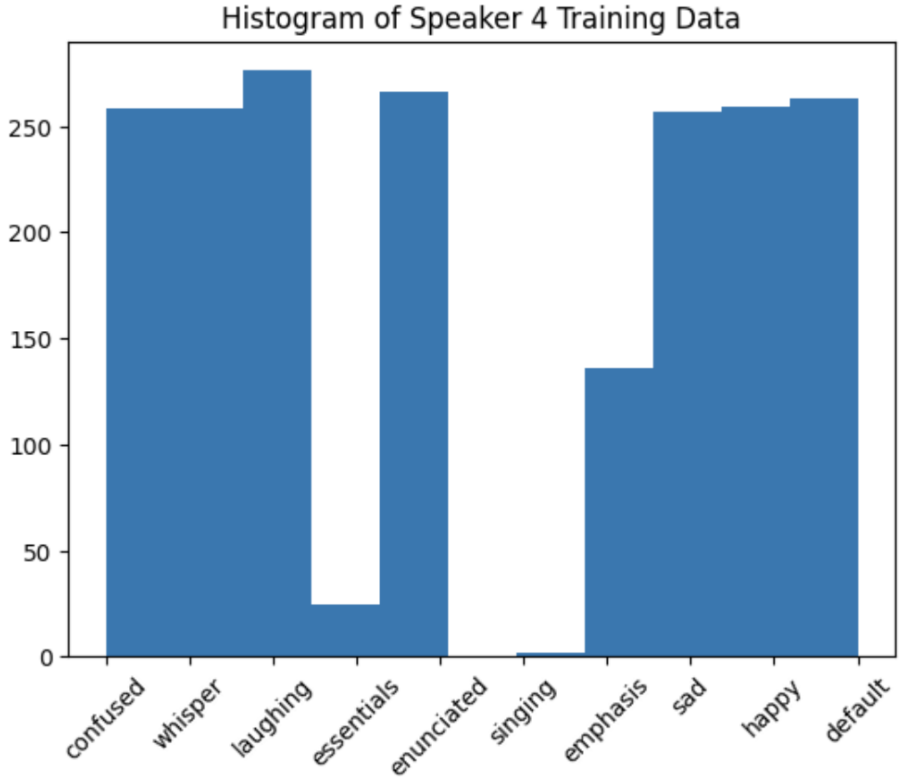
$(M_t^A)_t$ $(M_t^H)_t$

Data - Dataset

$(M_t^A)_t$ $(M_t^H)_t$

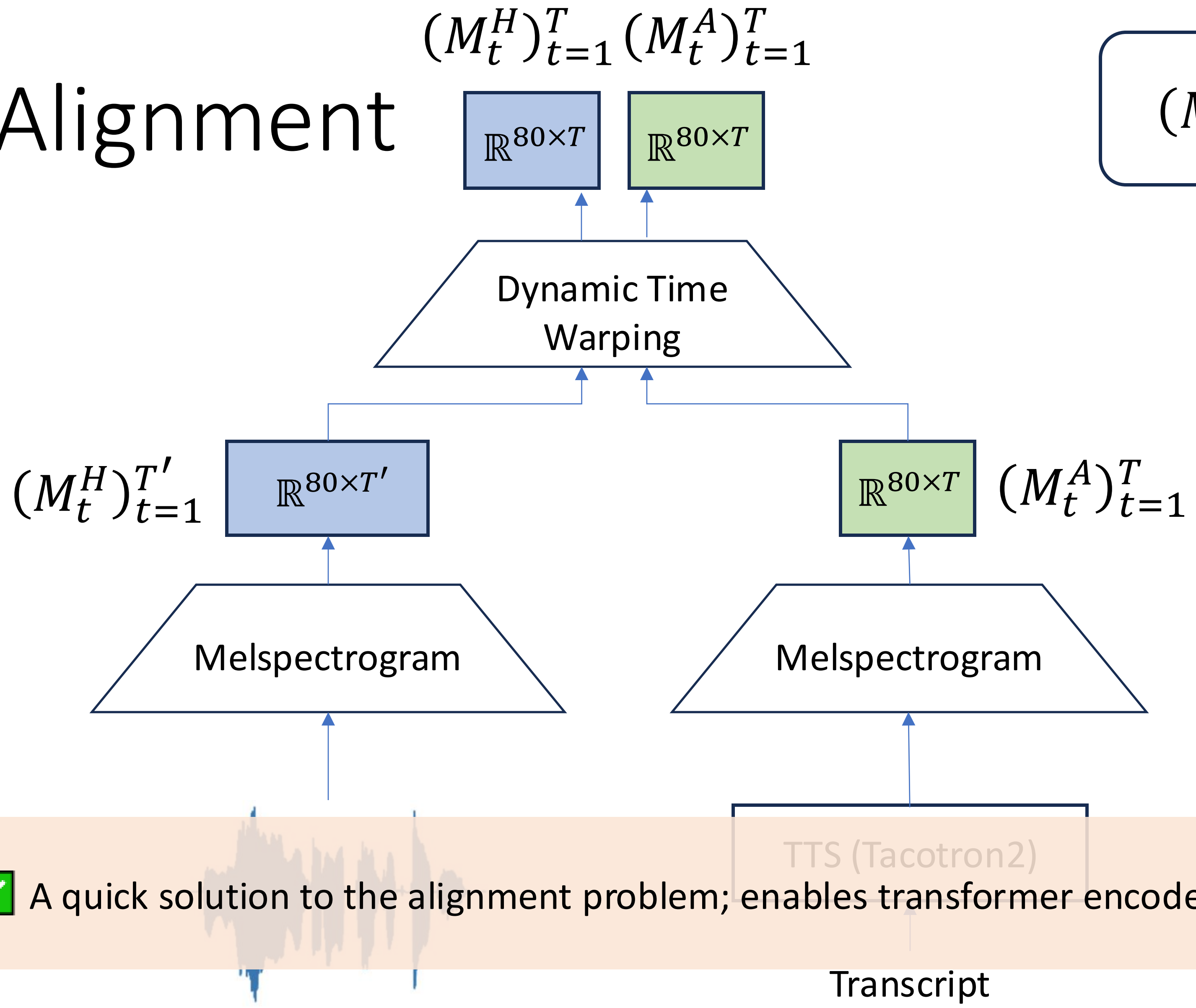
- Used EXPRESSO dataset, focused on speaker 4 (female)
- Train-valid-test split: (2000, 450, 453)

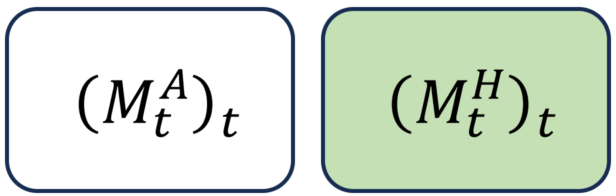
```
{'confused': 258,
 'whisper': 258,
 'laughing': 276,
 'essentials': 25,
 'enunciated': 266,
 'singing': 2,
 'emphasis': 136,
 'sad': 257,
 'happy': 259,
 'default': 263}
```



Nguyen et al., 2023

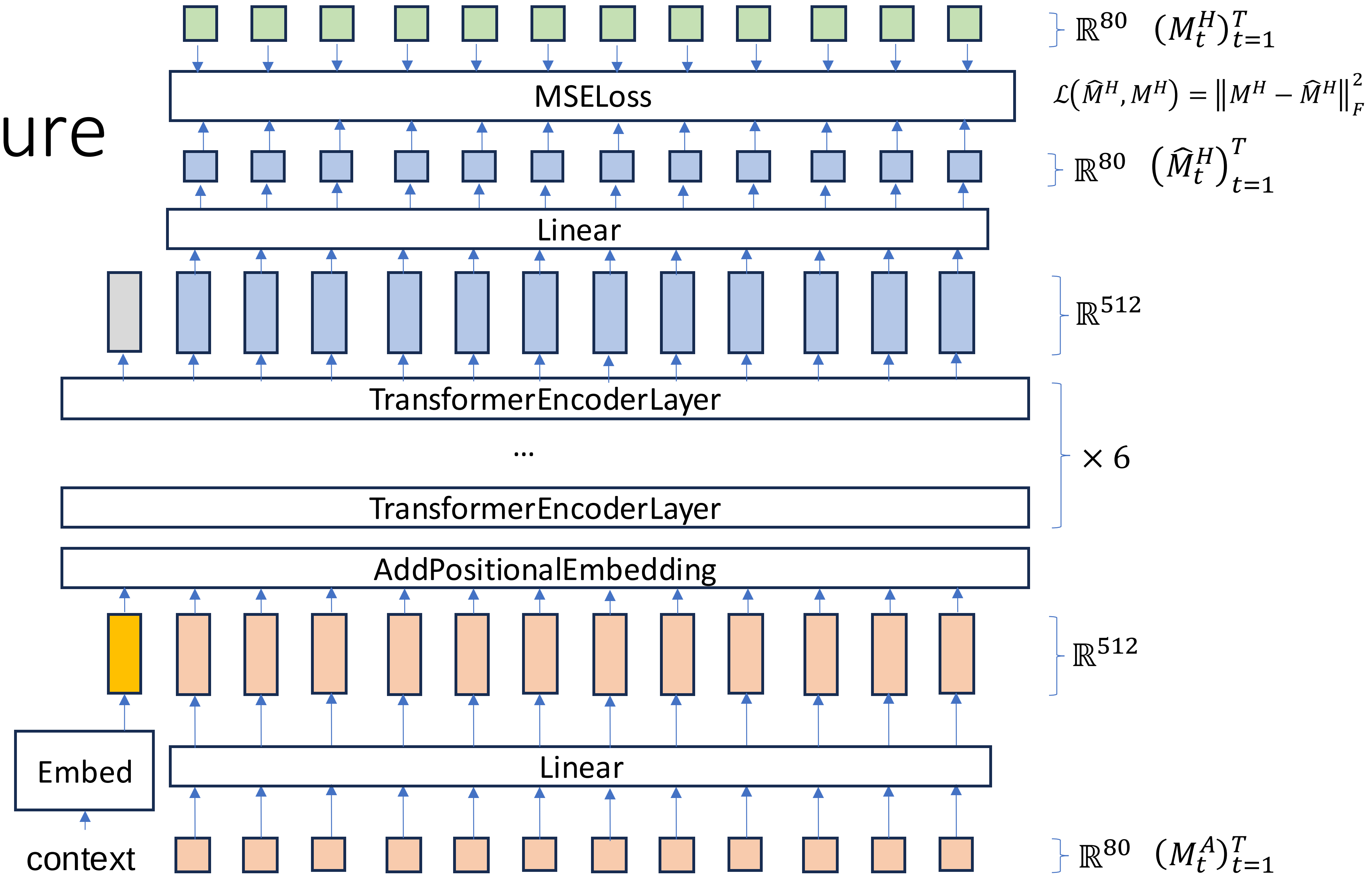
Data - Alignment





✅ A quick solution to the alignment problem; enables transformer encoder model

Architecture

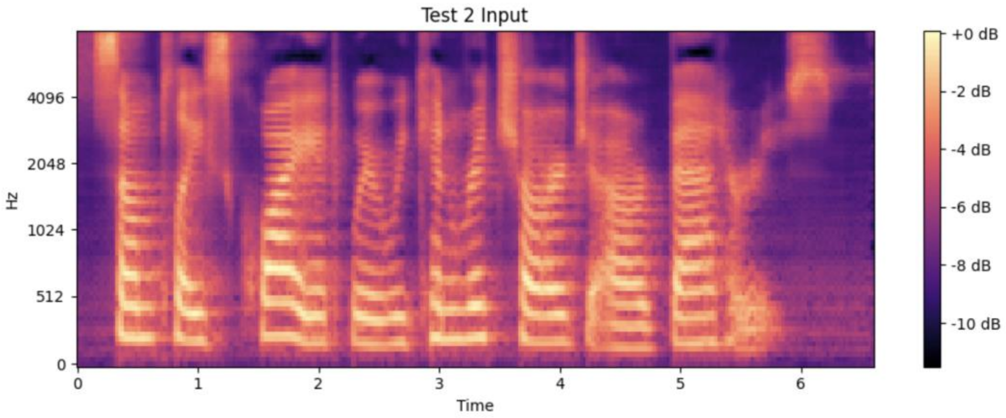


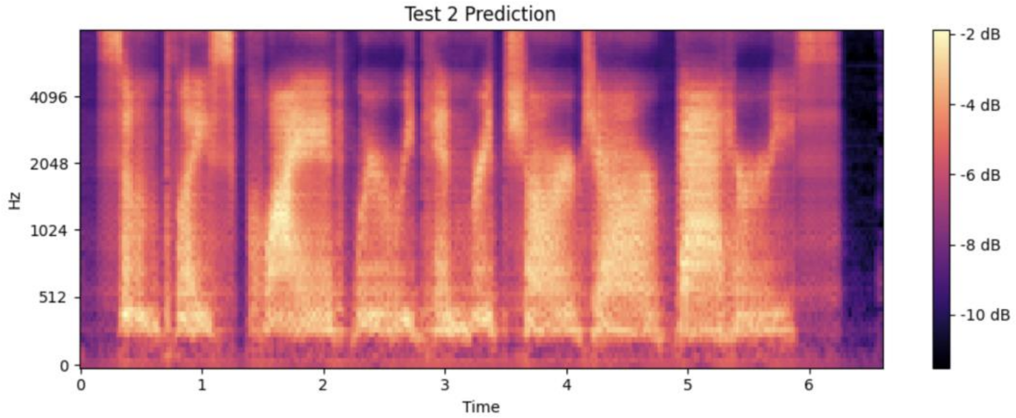
Results - Baseline

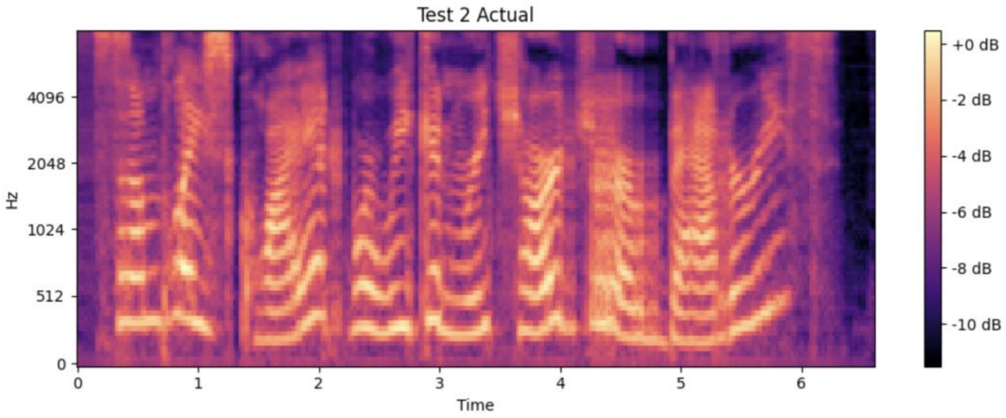
- Transcript:

"Someplace quiet where we can recharge our batteries?"

- Context: Happy





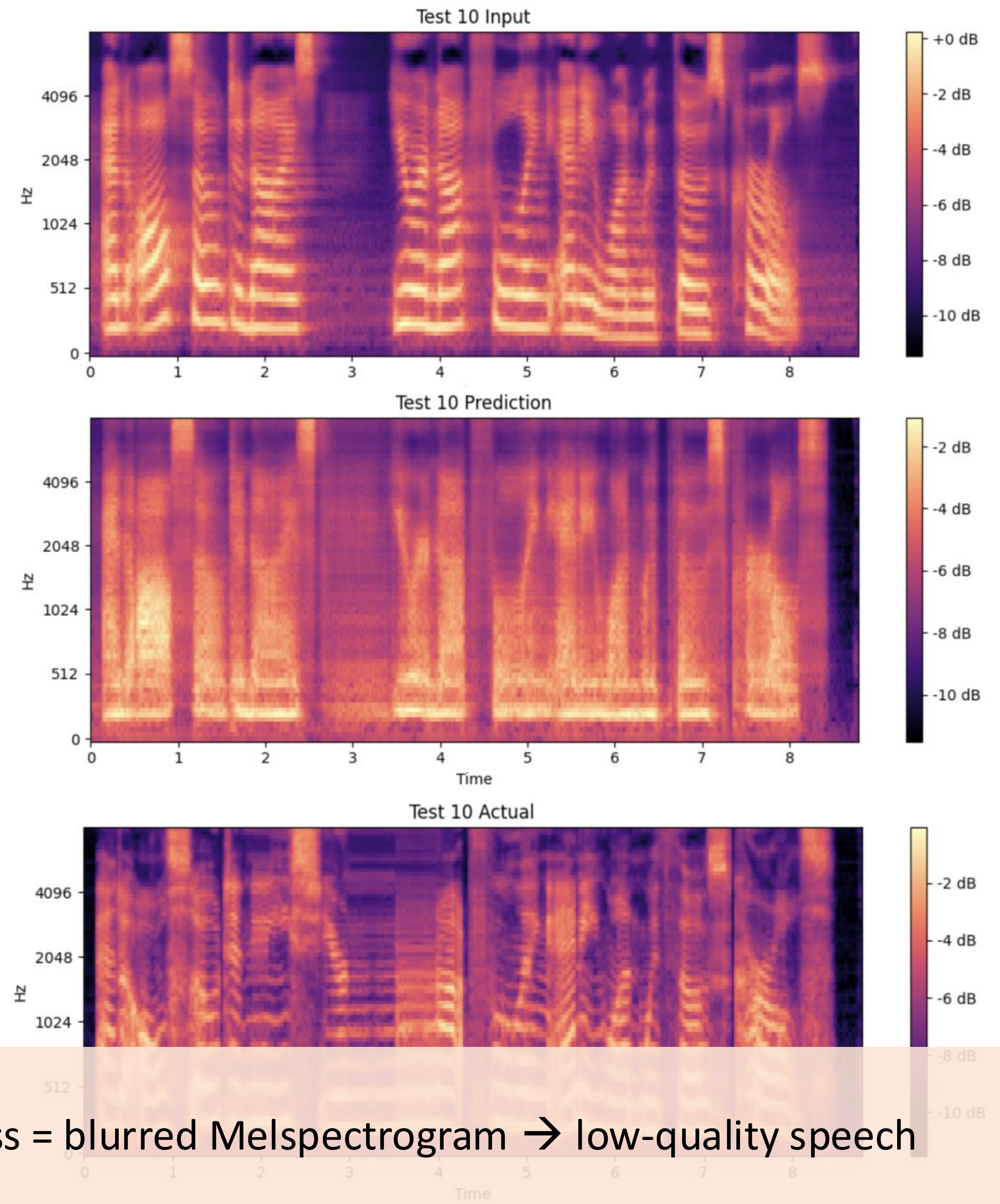


Results - Baseline

- Transcript:

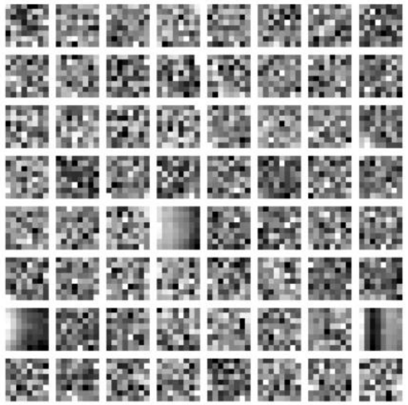
"And also, Denise, you're not fooling anyone with those curls."

- Context: Sad



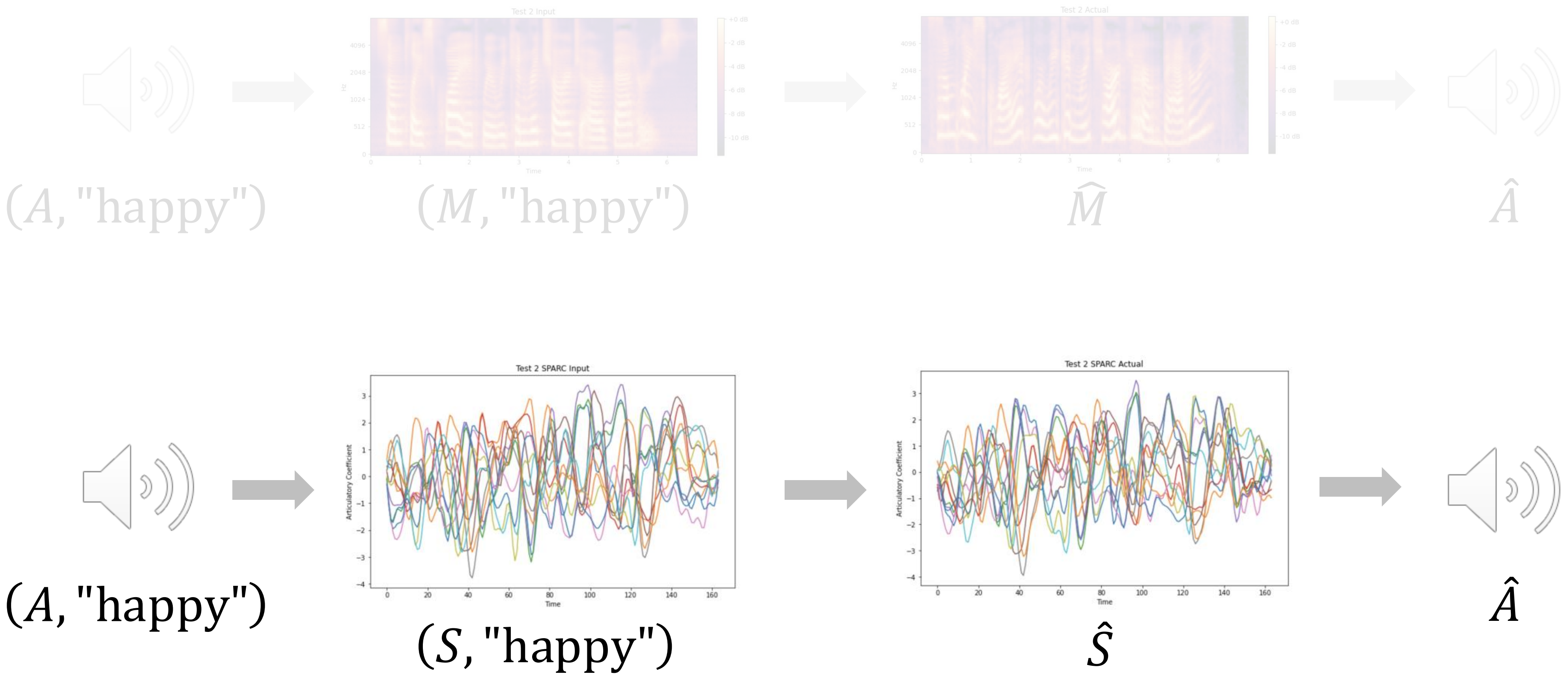
⛔ Transformer encoder model + MSE loss = blurred Melspectrogram → low-quality speech

Architecture Variants



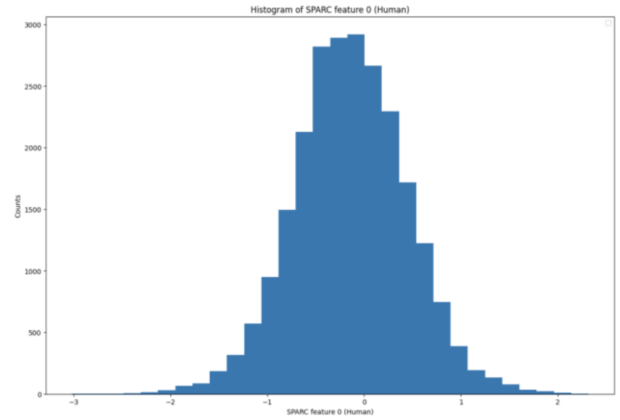
- Adding intermediate CNN layers
  - (+) Adds in inductive bias of locality of spectrogram image features
  - (-) Same blurry Melspectrogram and quality of speech
- MSE + GAN-Loss
  - (-) Combining with MSE loss results in stretching out of spectrogram
- Soft DTW loss
  - (-) Combining with MSE loss results in stretching out of spectrogram
- Duration models
  - (+) An alternative to soft DTW
  - (-) Computationally unfeasible to complete before project deadline

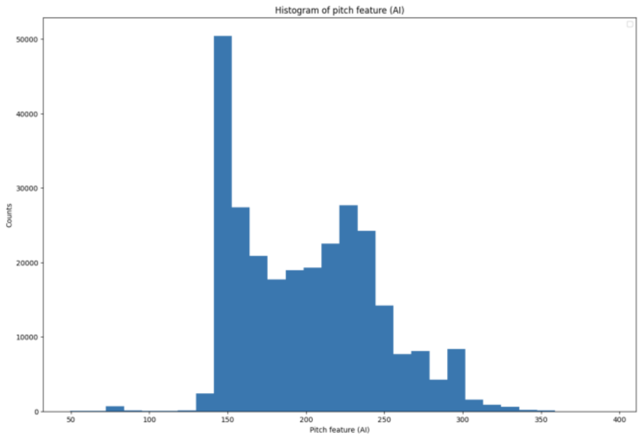
Methodology - SPARC

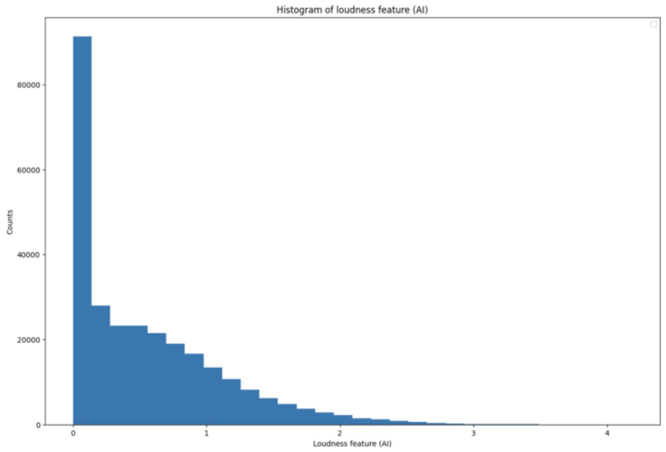


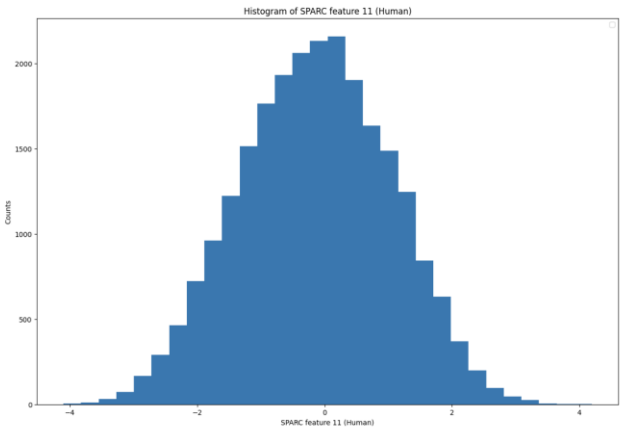
Cho et al., 2023

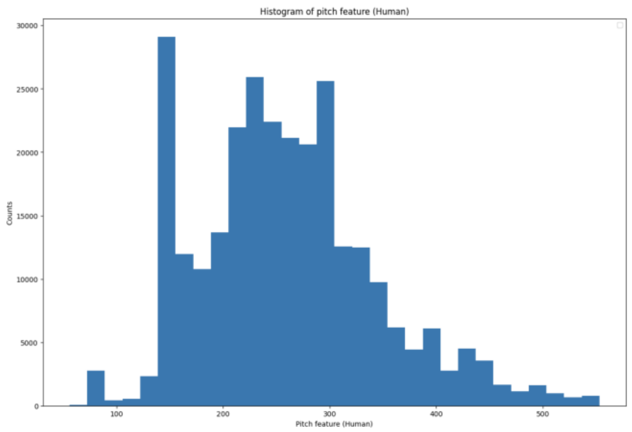
Data - SPARC

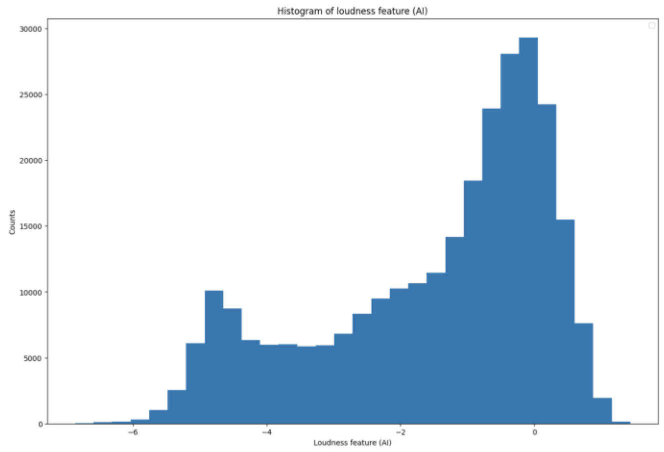

Histogram of SPARC feature 0 (Human)
Counts
SPARC feature 0 (Human)


Histogram of pitch feature (AI)
Counts
Pitch feature (AI)

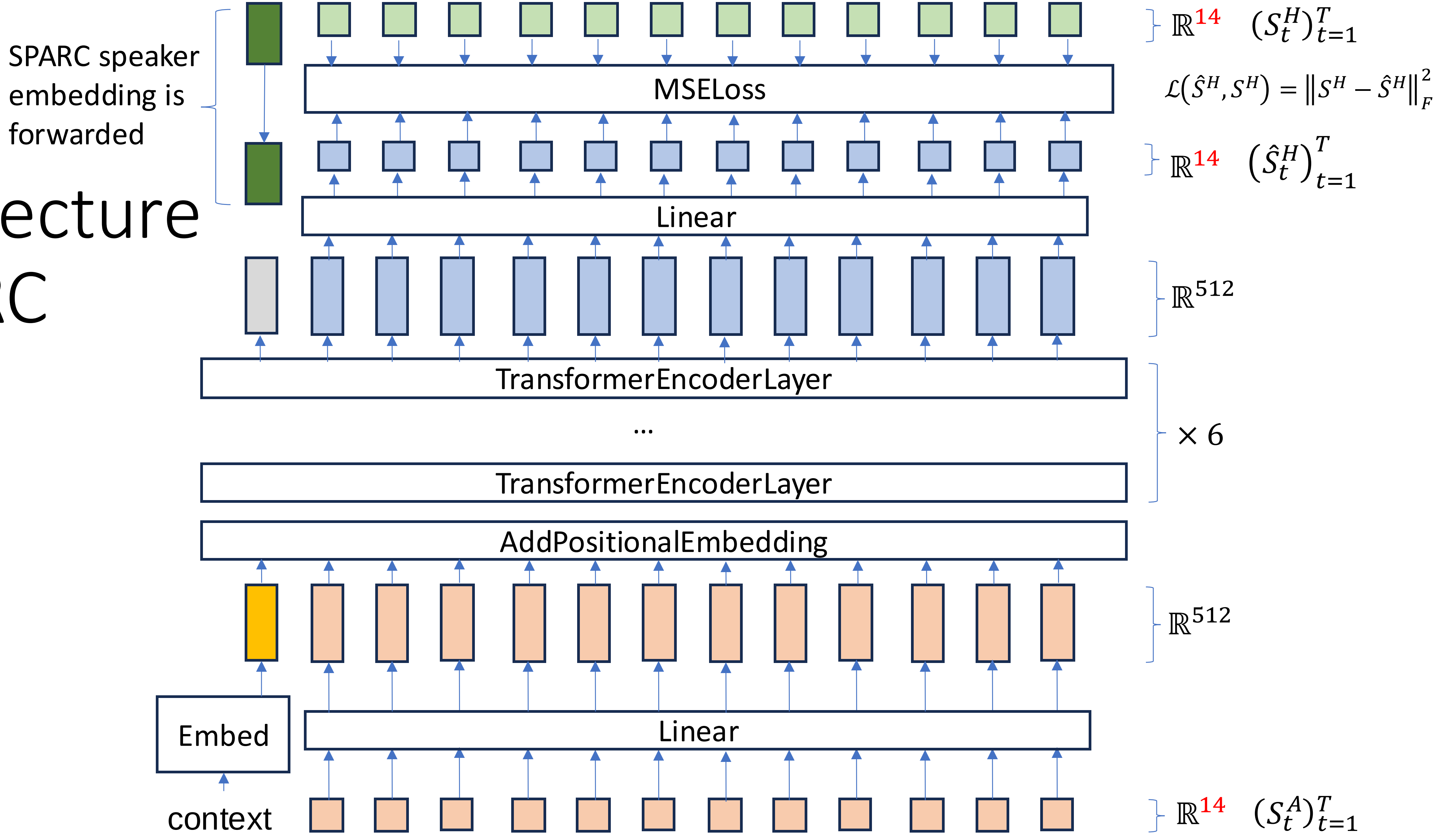

Histogram of loudness feature (AI)
Counts
Loudness feature (AI)


Histogram of SPARC feature 11 (Human)
Counts
SPARC feature 11 (Human)


Histogram of pitch feature (Human)
Counts
Pitch feature (Human)


Histogram of loudness feature (AI)
Counts
Loudness feature (AI)

Architecture - SPARC

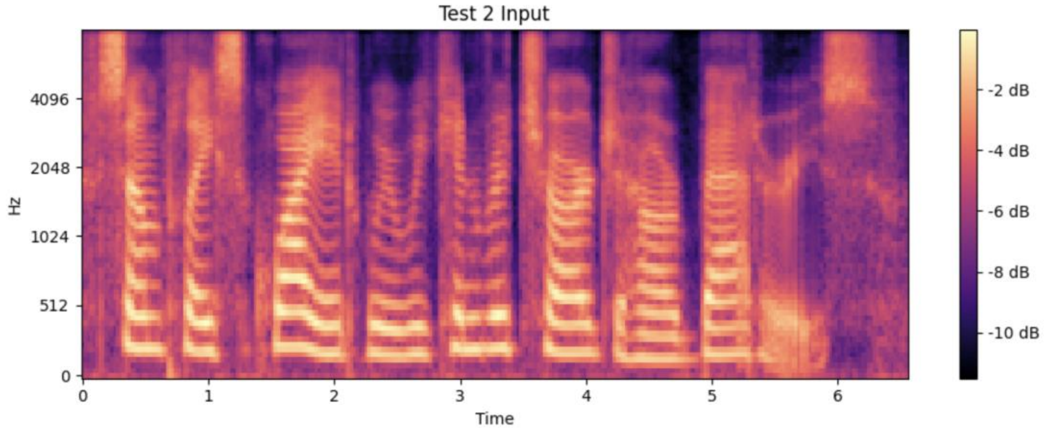


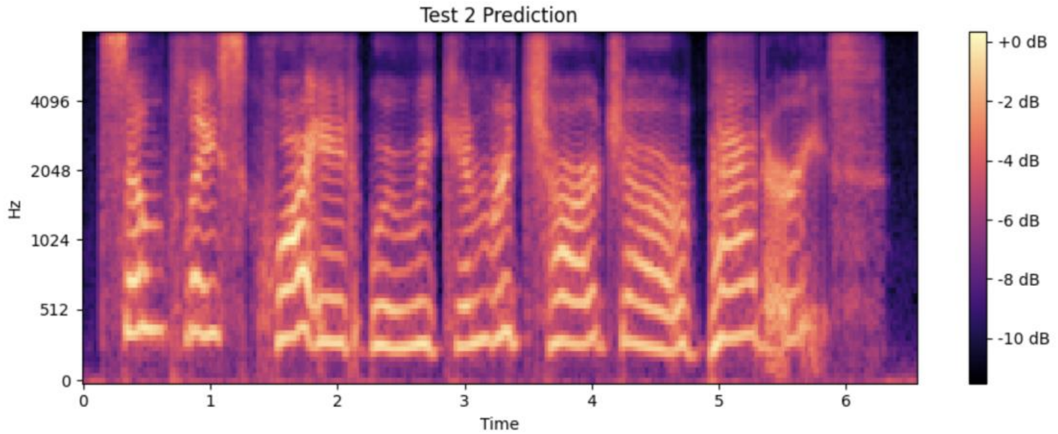
Results - SPARC

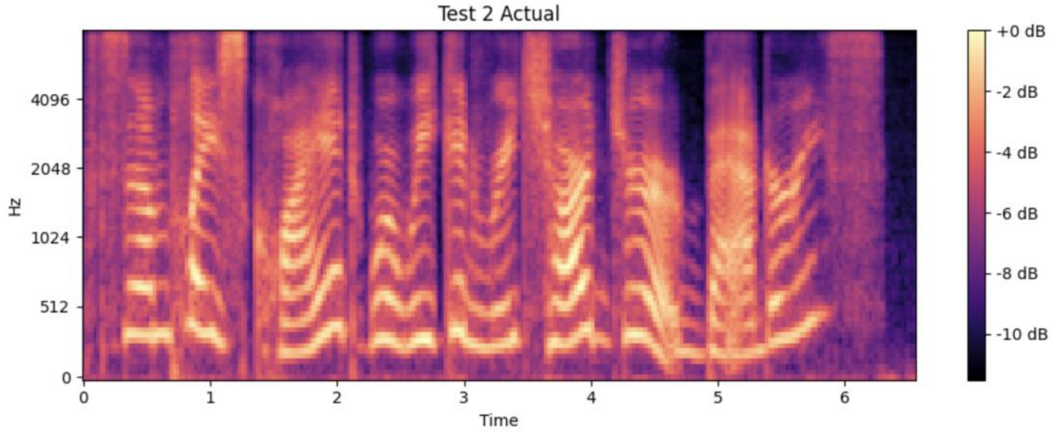
- Transcript:

"Someplace quiet where we can recharge our batteries?"

- Context: Happy





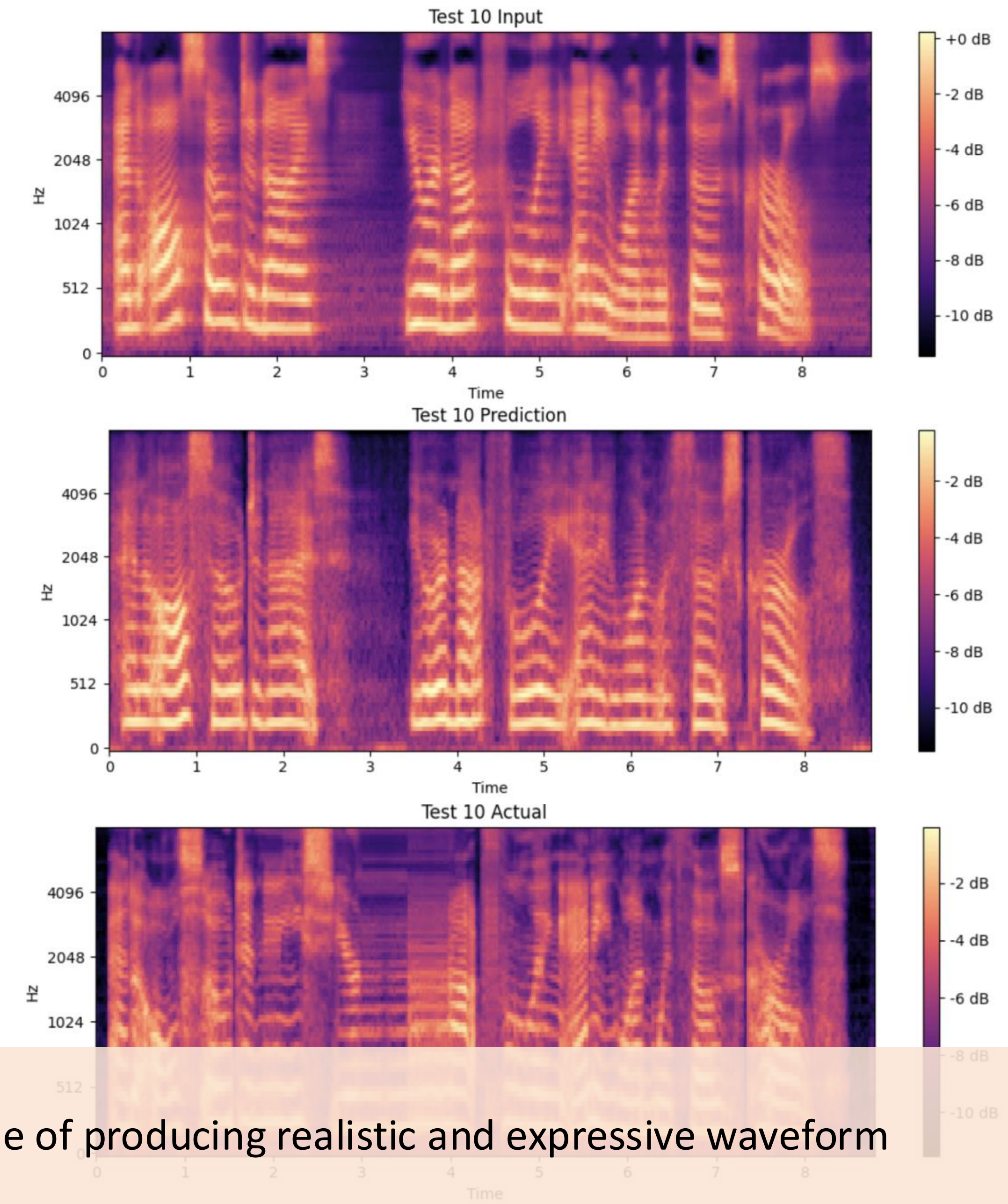


Results - Baseline

- Transcript:

"And also, Denise, you're not fooling anyone with those curls."

- Context: Sad



✅ SPARC-transformer model is capable of producing realistic and expressive waveform

Evaluations

- Type 1:
  - Tell evaluator the true label, then with paired $\left(A_i, \hat{A}_i, \mathrm{label}\right)_i$, randomly show evaluator one followed by the other
  - Obtain score of how well the context is expressed
  - Rank which one is better
- Type 2:
  - Generate $\hat{A}_i$ based on a label in {*happy*, *sad*, *laughing*, *confused*}, ask evaluator to infer the label

Evaluation – Samples

Confused

Actual Prediction

Laughing

Actual Prediction

Unseen

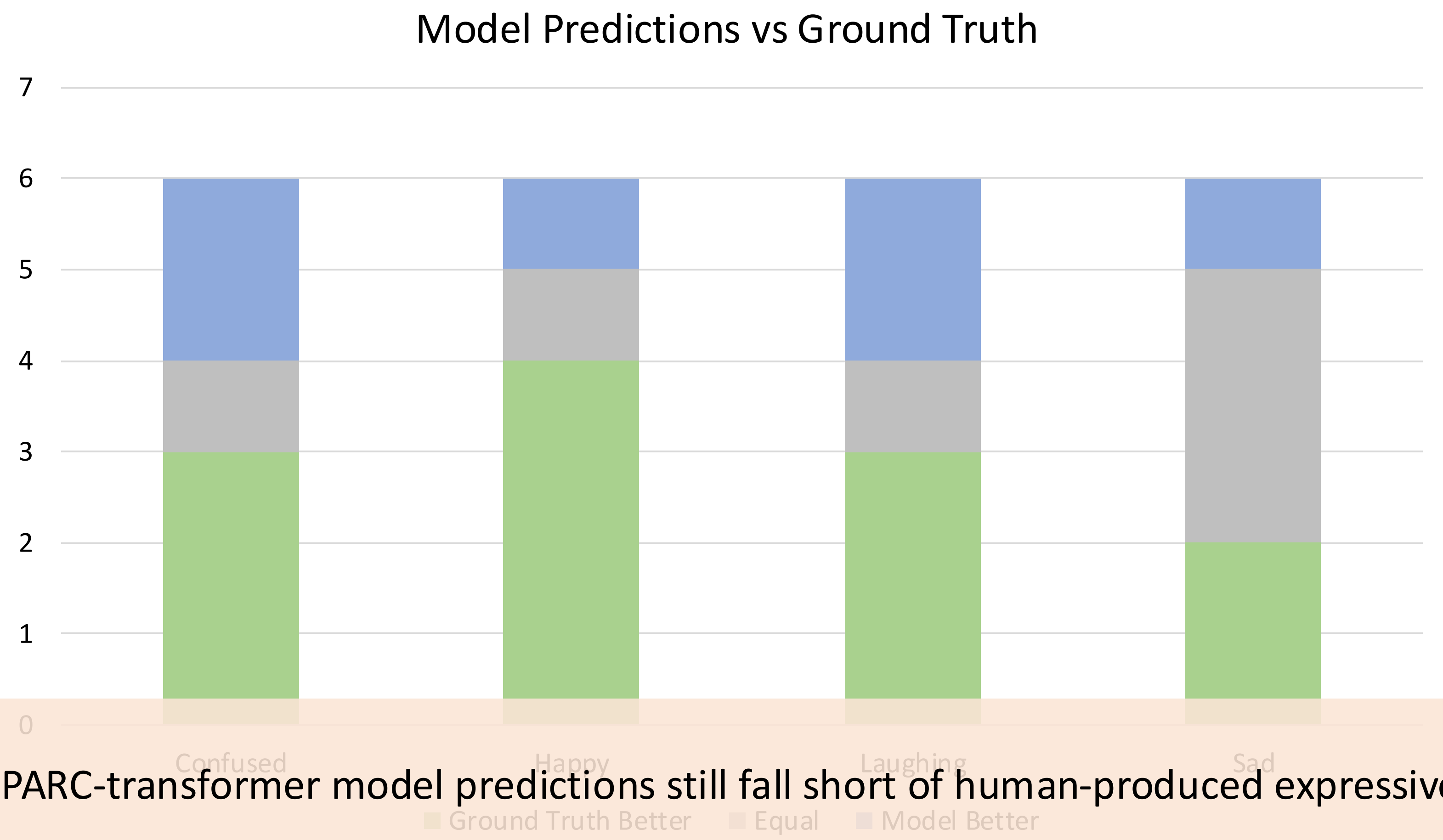
Happy

Actual Prediction

Sad

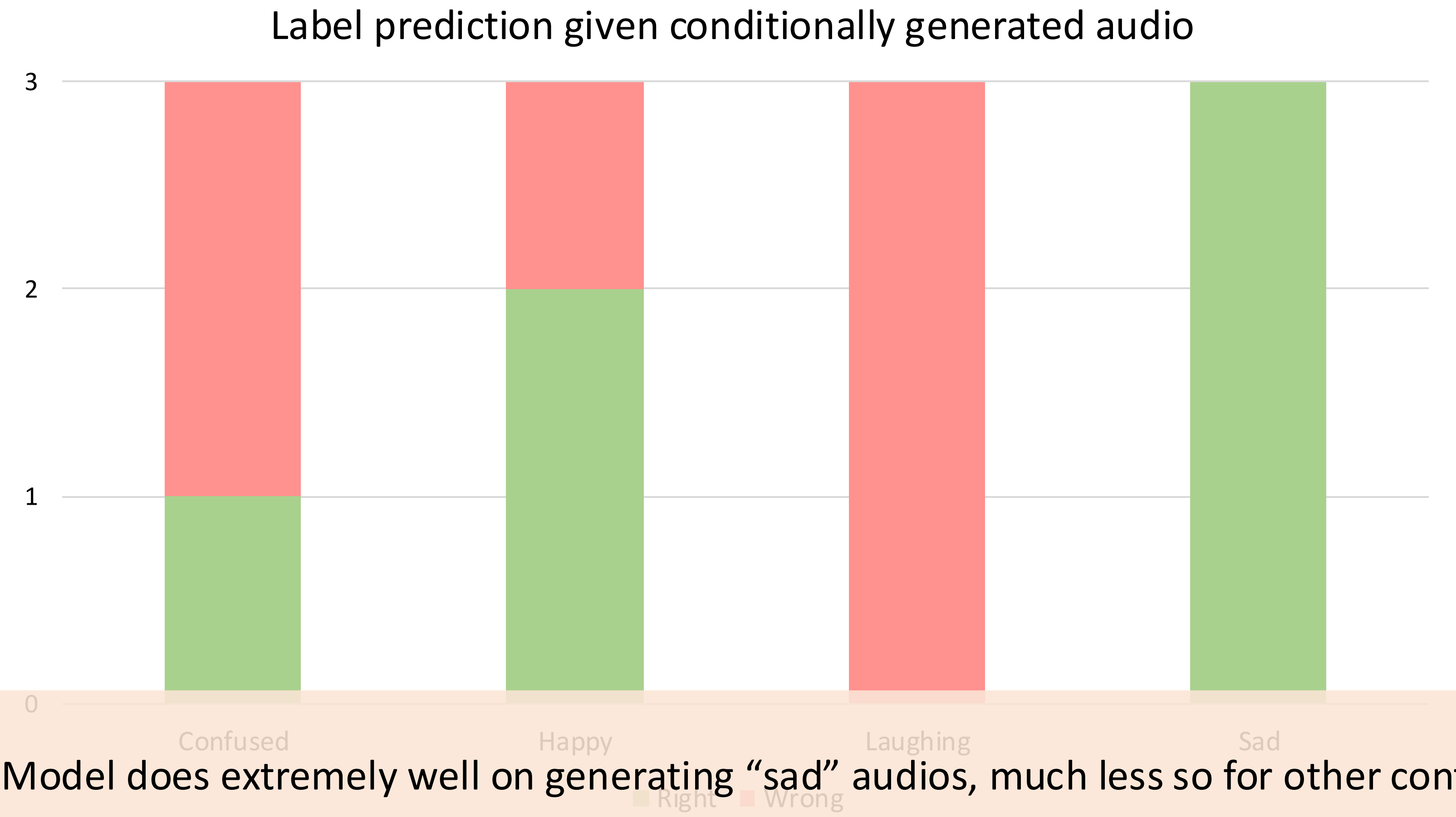
Actual Prediction

Evaluations – Type 1



❗SPARC-transformer model predictions still fall short of human-produced expressive speech

Evaluations – Type 2



❗ Model does extremely well on generating "sad" audios, much less so for other contexts…

Discussions – Intuitive Explanations

- SPARC features are more standardised as opposed to the energy levels of the Melspectrogram
- Time series problem rather than an image problem
- Forwarding of the SPARC speaker embedding vector
- Involved a pre-trained model, which inevitably involved textual elements (WavLM)

Discussions - Future Directions

- Could SPARC speaker embeddings have encoded some contextual information?
  - Type 2 subjective evaluations support this hypothesis
- The inverse problem of “standardising” speech – purging out the expressiveness in an audio to obtain a monotonous speech
- Scaling up training data
- Cleverer ways to deal with alignment problems

Applications

- Producing speech with various emotions
  - Data augmentation, adversarial models
- Removes the need for natural language annotations

Thank you!