Quantifying Emotional Landscape of Music in Three Dimensions

Kirtana Sunil Phatnani, Prof. Hemant A. Patil

Presented by: Prof. Hemant A. Patil

We all listen to Music

We all listen to Music

BUT WHAT HAPPENS INSIDE US WHEN WE DO?

We feel Emotions

We feel Emotions

What are emotions?

Damasio (1999) describes an emotion as neural object (or internal emotional state) as an (non-conscious) neural reaction to a certain stimulus, realised by a complex ensemble of neural activations in the brain.

We feel Emotions

What are emotions?

Damasio (1999) describes an emotion as neural object (or internal emotional state) as an (non-conscious) neural reaction to a certain stimulus, realised by a complex ensemble of neural activations in the brain.

Emotions evolved to help us form bonds and relationships.

Gómez, C. C. (2000). Damasio, Antonio (1999). The feeling of what happens Body and emotion in the making of consciousness. New York: Harcourt Brace & Company. 386 pp. *Persona: Revista de la Facultad de Psicología*, (3), 188-192.

Bosse, T., Jonker, C. M., & Treur, J. (2008). Formalisation of Damasio's theory of emotion, feeling and core consciousness. *Consciousness and cognition*, *17*(1), 94-113.

Why is it **important** to understand emotions?

INTELLIGENCE FOR IMAGINATION

Why is it **important** to understand emotions?

Emotions center our motivations, actions and decisions [1].

Why is it **important** to understand emotions?

Emotions center our motivations, actions and decisions [1].

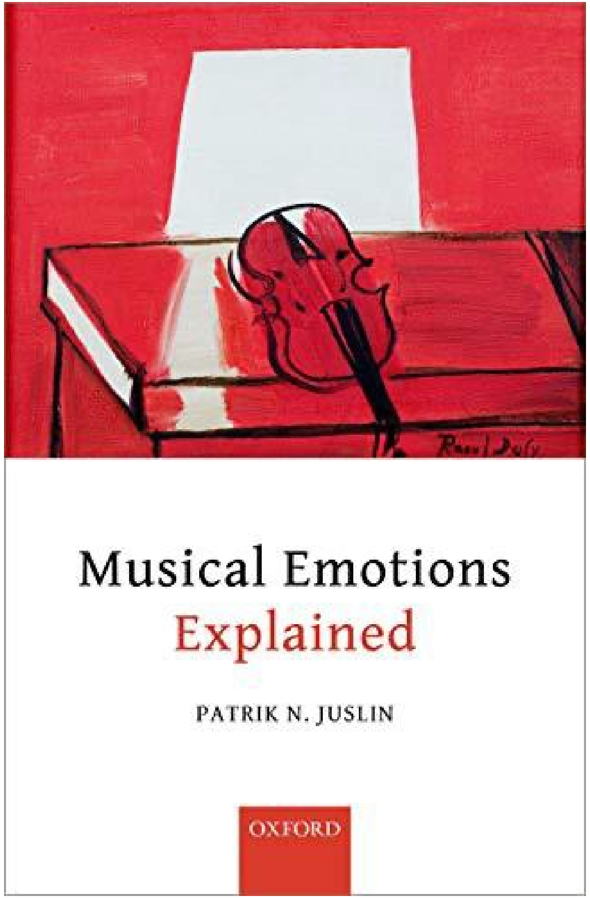
Mental health issues are a growing concern [2].

REFERENCES:

[1] DAMASIO, A. The Strange Order of Things: Life, Feeling, and the Making of Cultures New York: Pantheon Books 2018, 336 s. *Filozofia*, *73*(6), 481.

[2] American Psychological Association. (2019). Mental health issues increased significantly in young adults over last decade. *Retrieved December*, *12*, 2022.

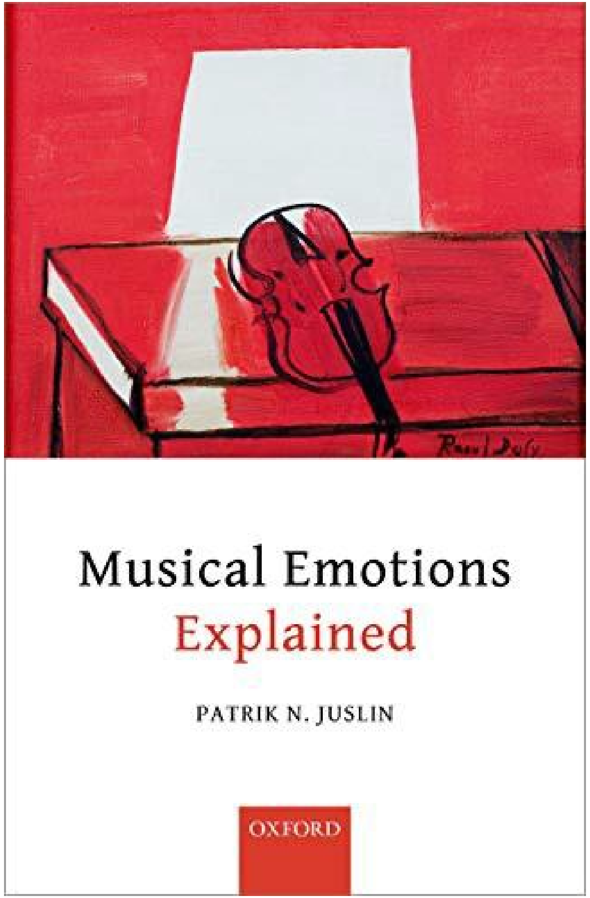
Musical Emotions Explained



Juslin, P. N. (2019). *Musical emotions explained: Unlocking the secrets of musical affect*. Oxford University Press, USA.

Musical Emotions Explained

Juslin's research was particularly instrumental in developing a comprehensive understanding of musical emotions, which identified seven key phenomena:

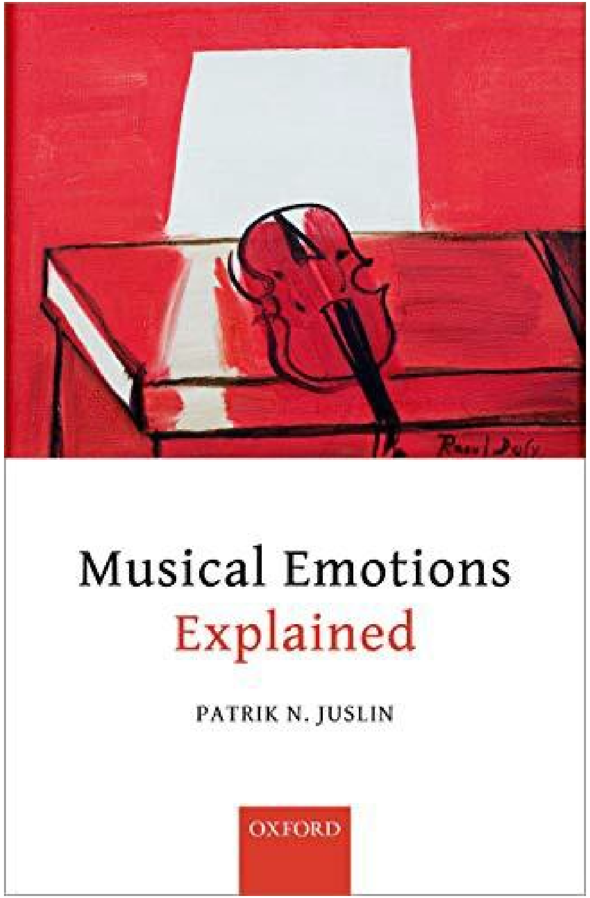


Juslin, P. N. (2019). *Musical emotions explained: Unlocking the secrets of musical affect*. Oxford University Press, USA.

Musical Emotions Explained

Juslin's research was particularly instrumental in developing a comprehensive understanding of musical emotions, which identified seven key phenomena:

- brain stem reflex,

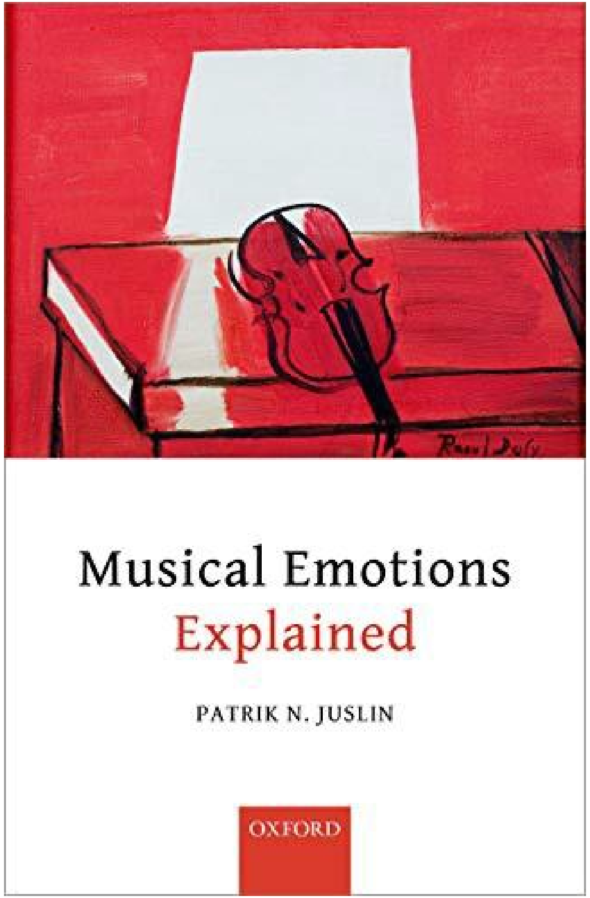


Juslin, P. N. (2019). *Musical emotions explained: Unlocking the secrets of musical affect*. Oxford University Press, USA.

Musical Emotions Explained

Juslin's research was particularly instrumental in developing a comprehensive understanding of musical emotions, which identified seven key phenomena:

- brain stem reflex,
- rhythmic entertainment,

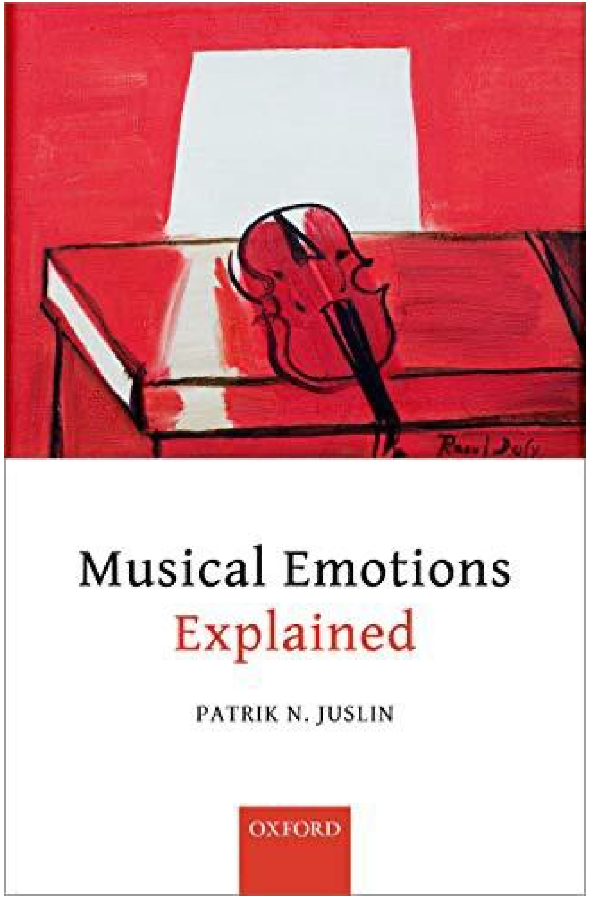


Juslin, P. N. (2019). *Musical emotions explained: Unlocking the secrets of musical affect*. Oxford University Press, USA.

Musical Emotions Explained

Juslin's research was particularly instrumental in developing a comprehensive understanding of musical emotions, which identified seven key phenomena:

- brain stem reflex,
- rhythmic entertainment,
- emotional contagion,



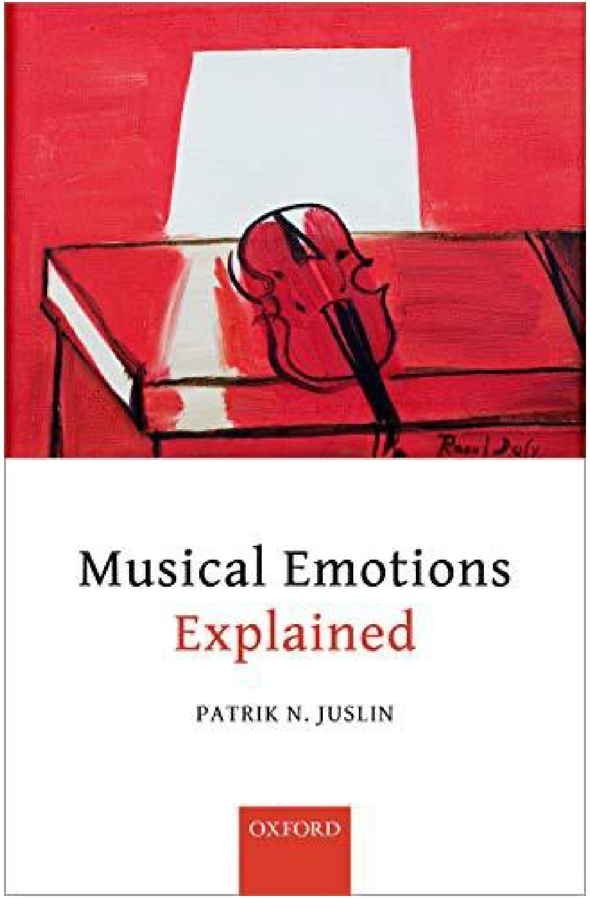
Juslin, P. N. (2019). *Musical emotions explained: Unlocking the secrets of musical affect*. Oxford University Press, USA.

fractal
INTELLIGENCE FOR IMAGINATION

Musical Emotions Explained

Juslin's research was particularly instrumental in developing a comprehensive understanding of musical emotions, which identified seven key phenomena:

- brain stem reflex,
- rhythmic entertainment,
- emotional contagion,
- evaluative conditioning,

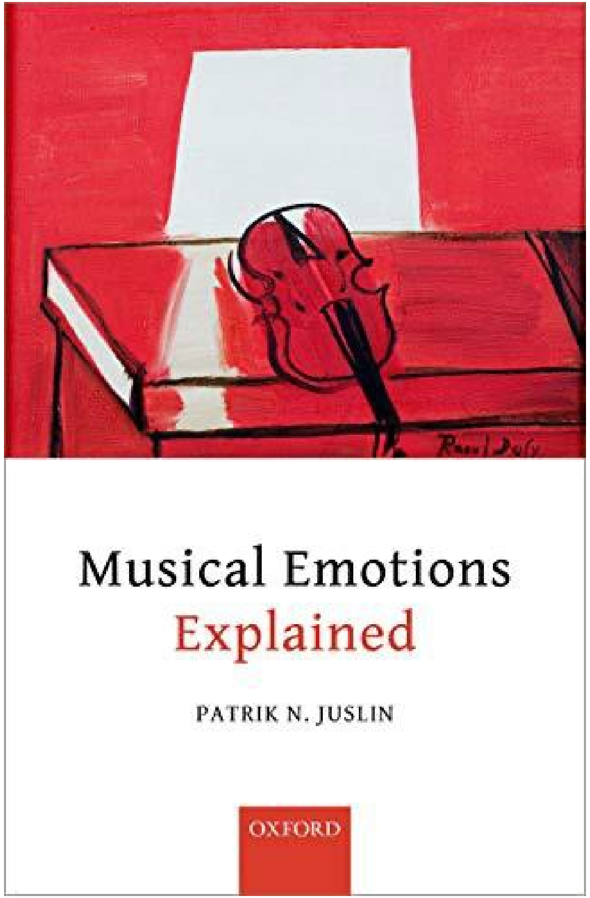


Juslin, P. N. (2019). *Musical emotions explained: Unlocking the secrets of musical affect*. Oxford University Press, USA.

Musical Emotions Explained

Juslin's research was particularly instrumental in developing a comprehensive understanding of musical emotions, which identified seven key phenomena:

- brain stem reflex,
- rhythmic entertainment,
- emotional contagion,
- evaluative conditioning,
- episodic memory,



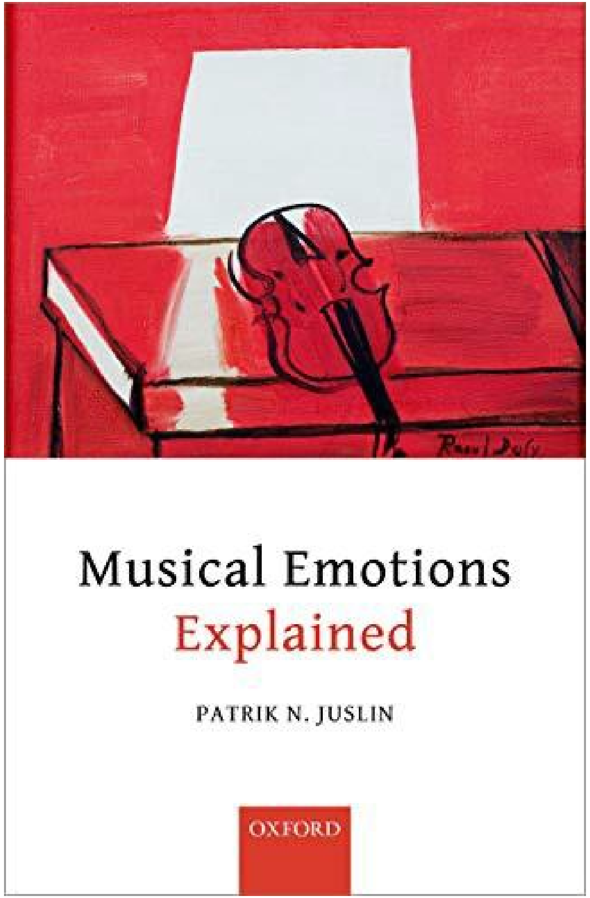
Juslin, P. N. (2019). *Musical emotions explained: Unlocking the secrets of musical affect*. Oxford University Press, USA.

fractal
INTELLIGENCE FOR IMAGINATION

Musical Emotions Explained

Juslin's research was particularly instrumental in developing a comprehensive understanding of musical emotions, which identified seven key phenomena:

- brain stem reflex,
- rhythmic entertainment,
- emotional contagion,
- evaluative conditioning,
- episodic memory,
- mental visual imagery, and

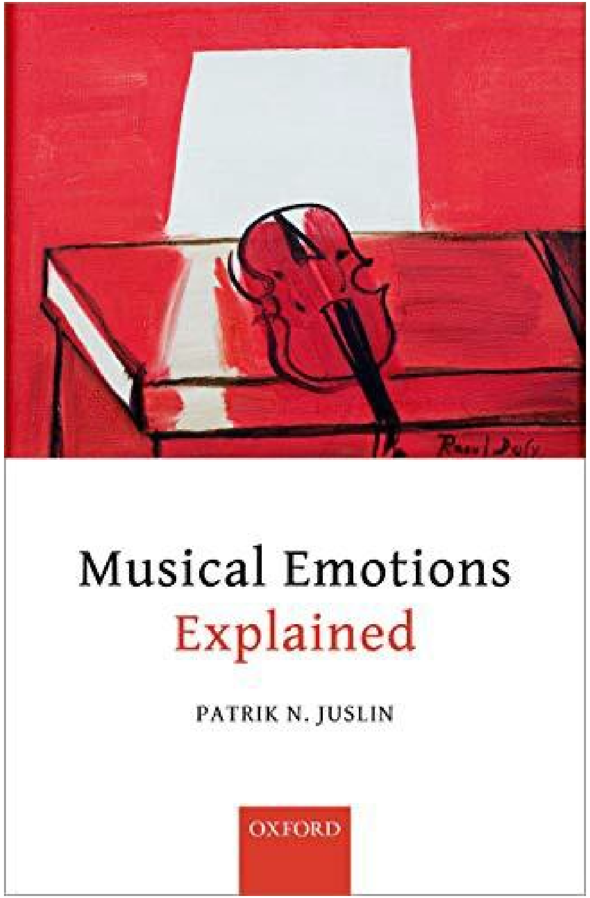


Juslin, P. N. (2019). *Musical emotions explained: Unlocking the secrets of musical affect*. Oxford University Press, USA.

Musical Emotions Explained

Juslin's research was particularly instrumental in developing a comprehensive understanding of musical emotions, which identified seven key phenomena:

- brain stem reflex,
- rhythmic entertainment,
- emotional contagion,
- evaluative conditioning,
- episodic memory,
- mental visual imagery, and
- musical expectancy



Juslin, P. N. (2019). *Musical emotions explained: Unlocking the secrets of musical affect*. Oxford University Press, USA.

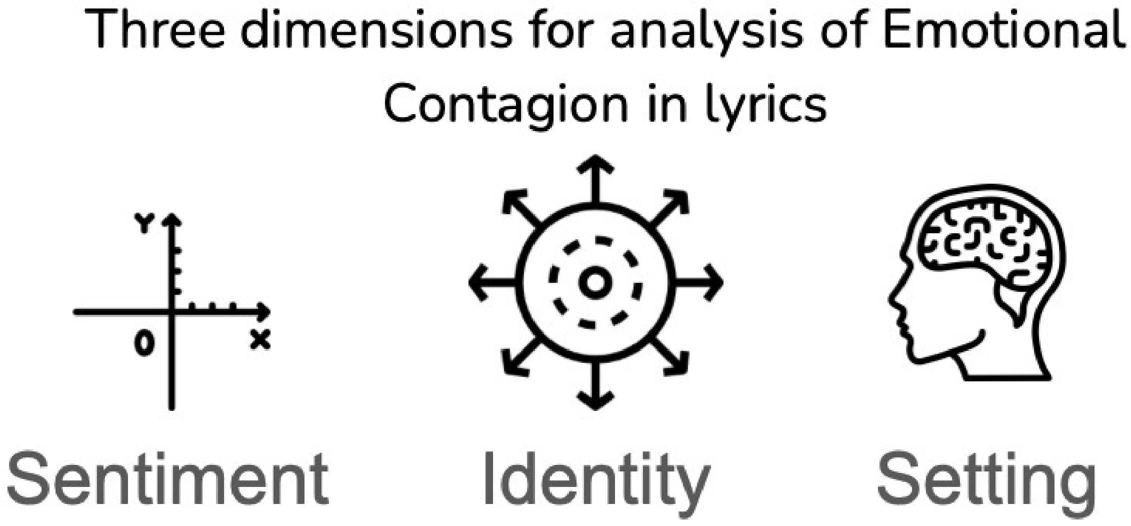
Music Emotion Recognition: Existing Paradigm Limitations

Most studies in the field of MER label an entire music piece with one emotional label [1].

Emotions from a lyrical music piece arise from the story between its characters.

REFERENCES:
[1] Yang, X., Dong, Y., Li, J.: Review of data features-based music emotion recognition methods. Multimedia Systems 24 (4), 365–389 (2018)

Brain Inspired Music Emotion Recognition



fractal
INTELLIGENCE FOR IMAGINATION

Capturing the **Emotional Landscape** of lyrics in a Song

To learn more please visit our poster

WiSSAP 2023

Interference Reduction in Music Signals

Rajesh R, Padmanabhan Rajan
Indian Institute of Technology Mandi

Recordings from Live Concerts

https://images.app.goo.gl/g9MPV2bNE5faJz4M7

- Live recordings lacks acoustic shielding
- Microphone intended to pick specific source picks up the other sources as well

Why and How?

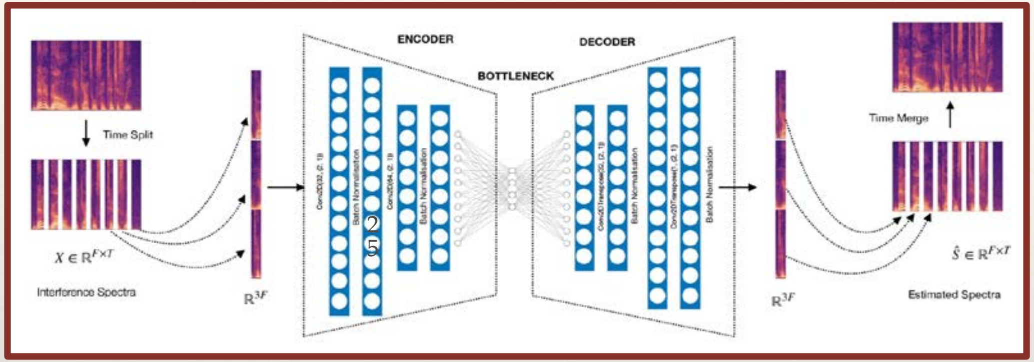
Needs

- Creating rich datasets for supervised source separation
- Music Information Retrieval (MIR) tasks

Goals

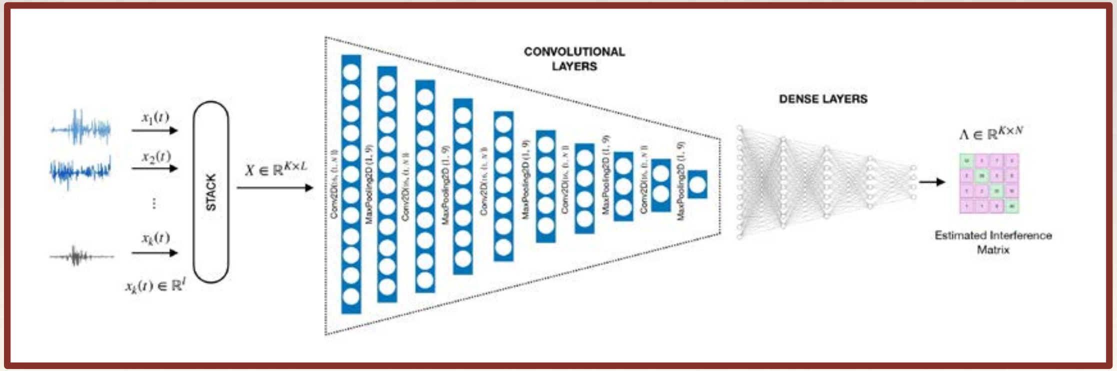
- Data independent models
- Faster, simpler, and efficient for live recordings

Learning based Frameworks



- **CAEs**: TF domain
- Treats interference as noise

- **t-UNet**: Waveform domain
- Estimates interference strength and uses that information to reduce bleed



Catch me at the poster session!

DIRECT SPEECH TO SPEECH TRANSLATION WITH VOICE INTERPOLATION

WiSSAP 2023

Industrial Engineering and Operation Research

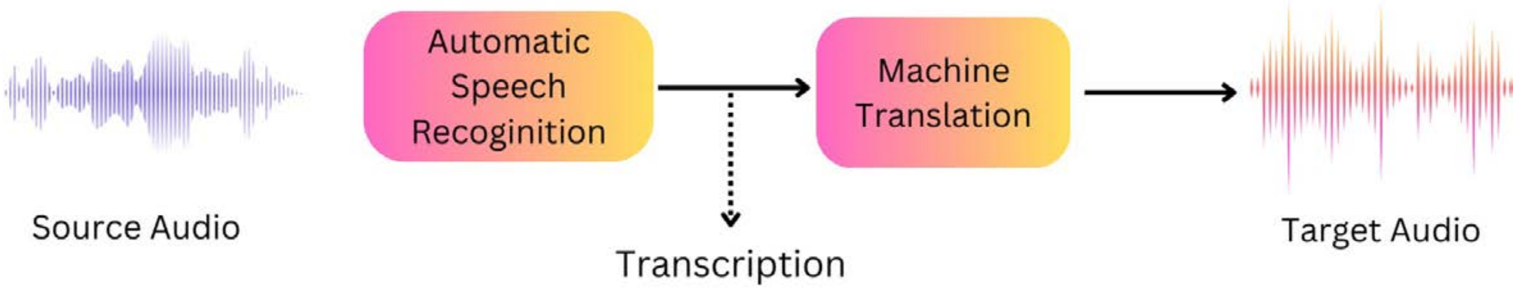
IIT Bombay

December 18, 2023

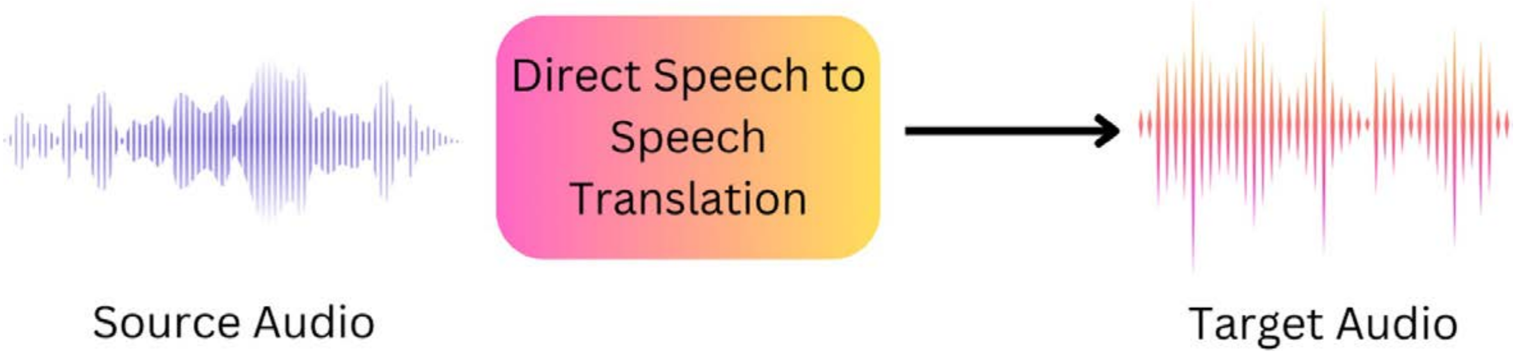
Presentation Outline

IIT Bombay

Problem Description

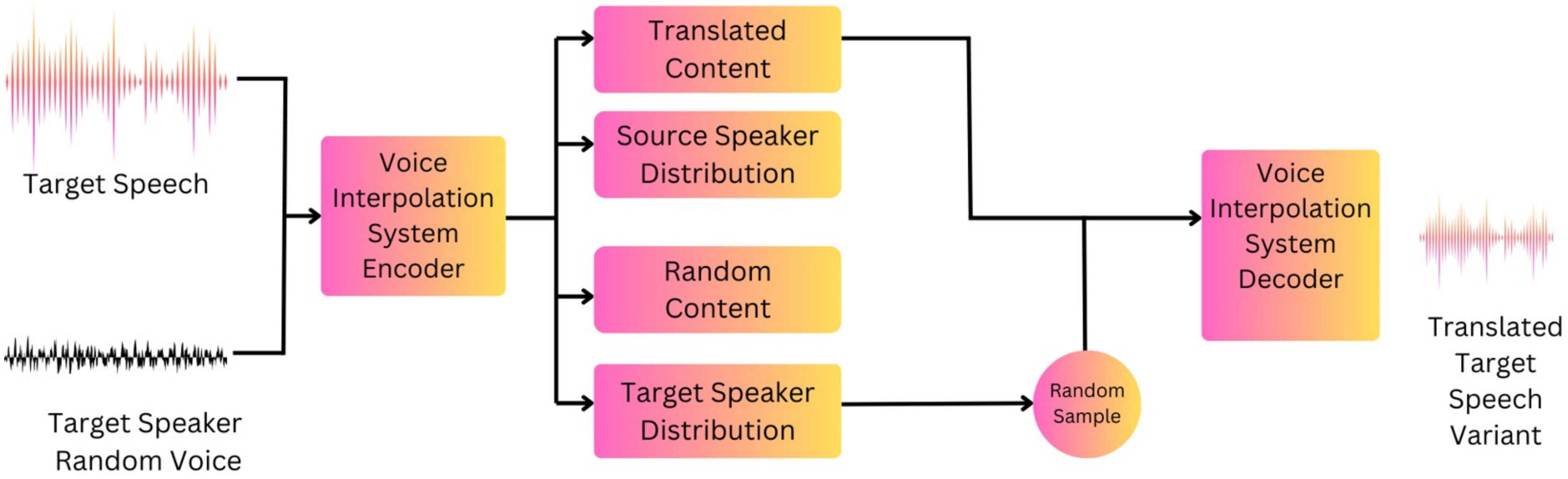


Conventional Speech-to-Speech approach



Direct Speech-to-Speech approach

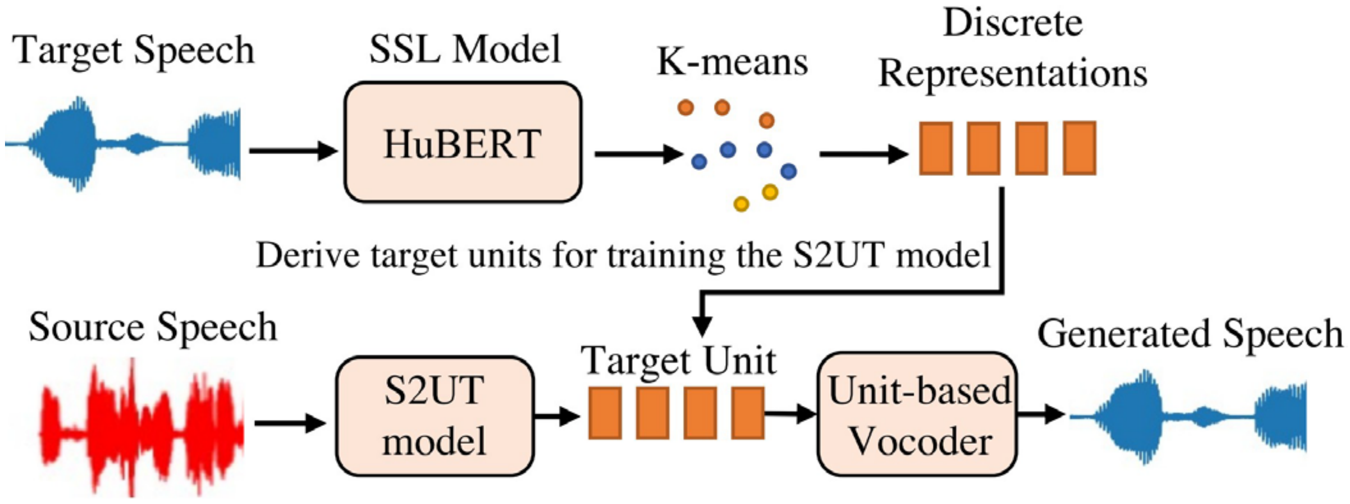
Problem Description



Speech-to-Speech translation and voice interpolation framework

Methodology

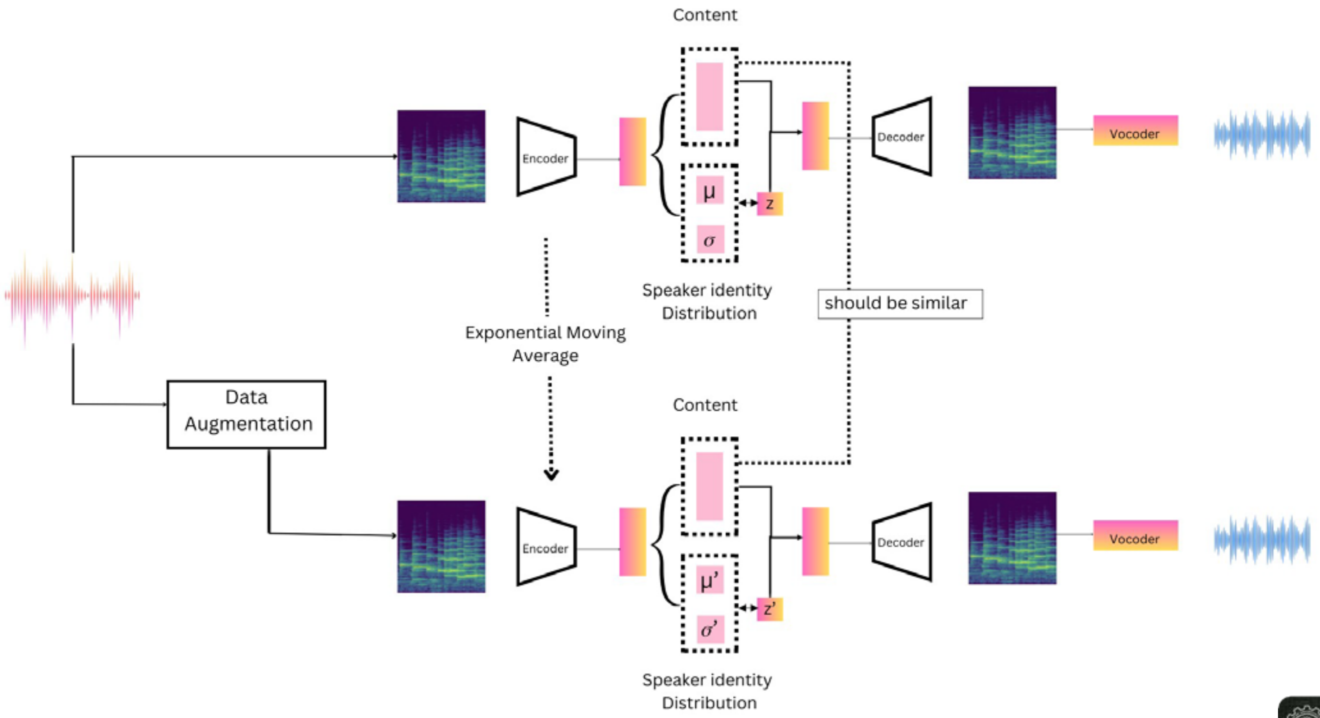
We fine-tuned the speech-to-speech translation model proposed by Lee et al. [1] for English-to-German translation.



Direct Speech-to-Speech Translation with Discrete Units

Methodology

We propose model architecture inspired by Grill et al. [2].



Results

- We have qualitative results for Speech-to-speech translation but have yet to work on quantitative results and comparisons.
- Voice interpolation work still in progress. We are using the Crema-d dataset, which contains 7442 clips of 91 actors for experiments.
- We faced a mode collapsing issue using previous implementations while training, where we used triplet loss for contrastive learning. The encoder learned to generate the same embedding for content.

Conclusion

- Finetuned a direct Speech-to-Speech translation system [3] that directly converts source speech to target speech, bypassing traditional pipelines.
- The method employs a pretrained HuBERT model trained with self-supervised learning and K-Means to create discrete unit representation.
- Our voice interpolation framework describes a novel approach to generate multiple speech variations of a speaker.

Future Work

- Analysis of the voice characteristics space by incorporating the output of the Speech-to-Speech translation system has to be performed.
- Performing quantitative analysis of described Speech-to-Speech translation and voice interpolation method.
- Currently, our Speech-to-Speech translation and voice interpolation framework incorporates a two-stage pipeline that can be incorporated into a single end-to-end network.

References

[1] R. Huang, J. Liu, H. Liu, Y. Ren, L. Zhang, J. He, and Z. Zhao, "Transpeech: Speech-to-speech translation with bilateral perturbation," *arXiv preprint arXiv:2205.12523*, 2022.

[2] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, B. Piot, k. kavukcuoglu, R. Munos, and M. Valko, "Bootstrap your own latent - a new approach to self-supervised learning," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 21271–21284, Curran Associates, Inc., 2020.

[3] A. Lee, P.-J. Chen, C. Wang, J. Gu, S. Popuri, X. Ma, A. Polyak, Y. Adi, Q. He, Y. Tang, *et al.*, "Direct speech-to-speech translation with discrete units," *arXiv preprint arXiv:2107.05604*, 2021.

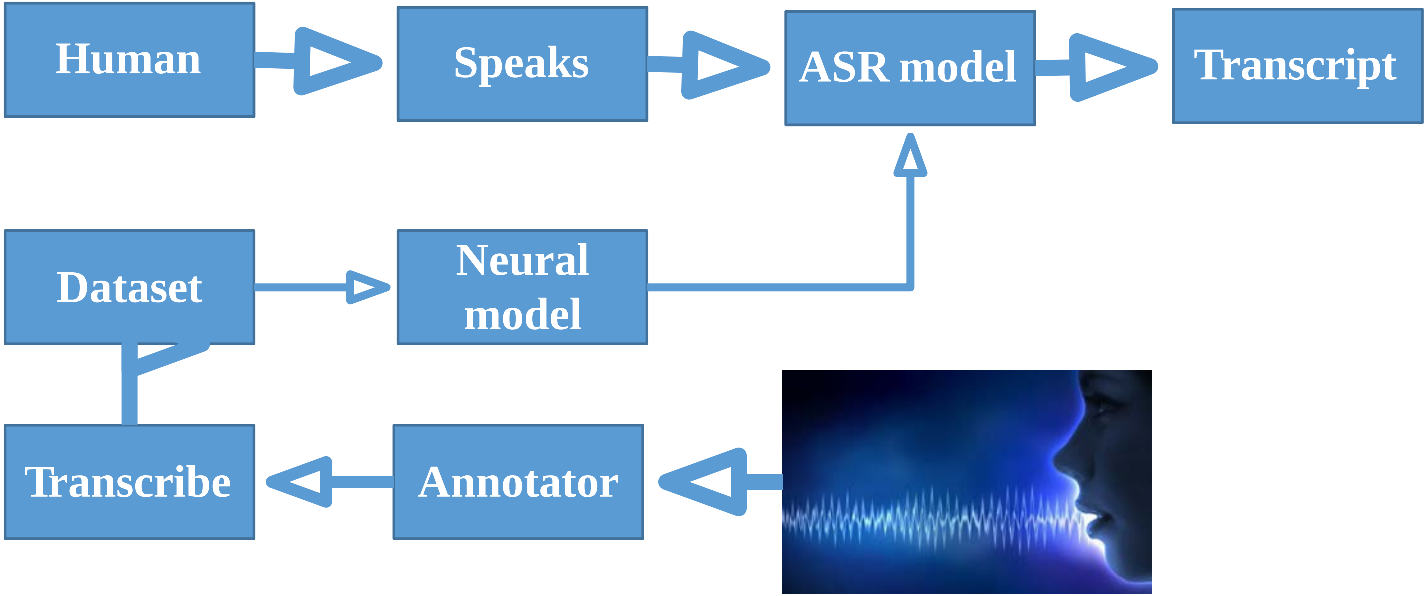
Thank You!

ASR Annotation Tool

Nagarathna R

Thishyan Raj T

Raja Ravi Teja Chaganti

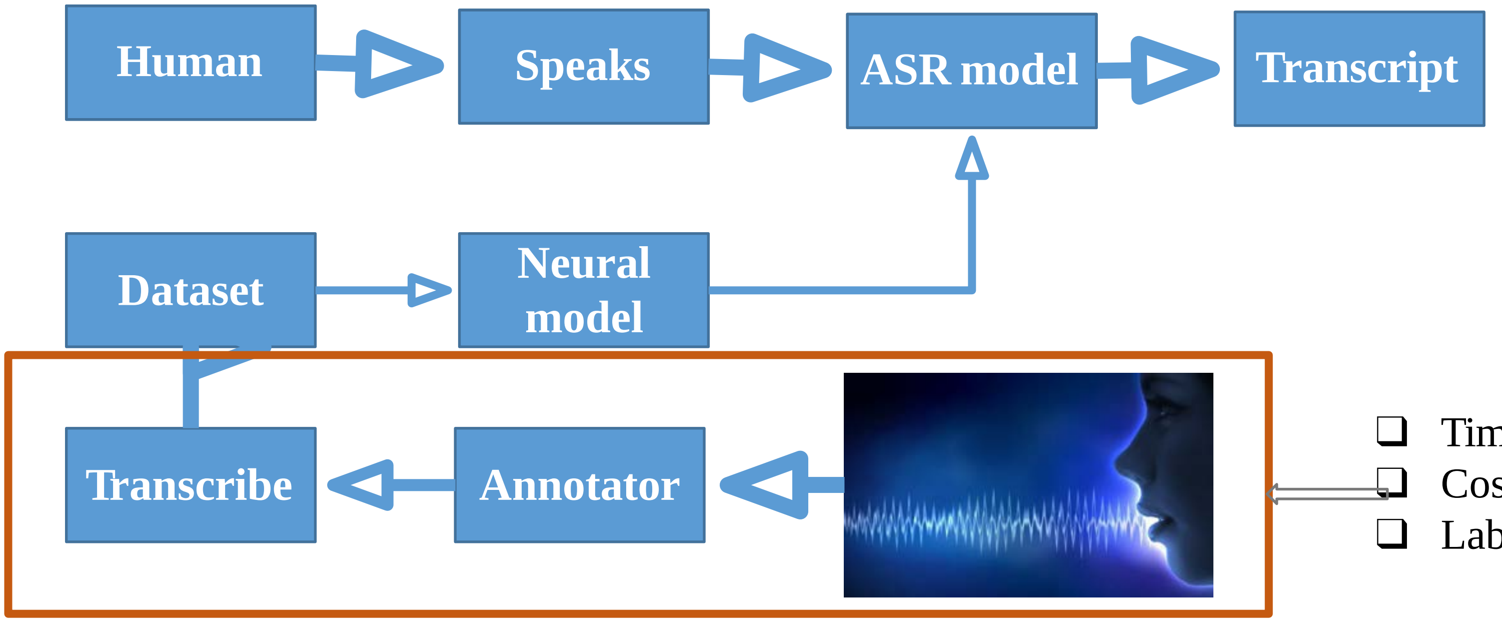


ASR Annotation Tool

Nagarathna R

Thishyan Raj T

Raja Ravi Teja Chaganti



- Time Consuming
- Costly
- Labor-intensive

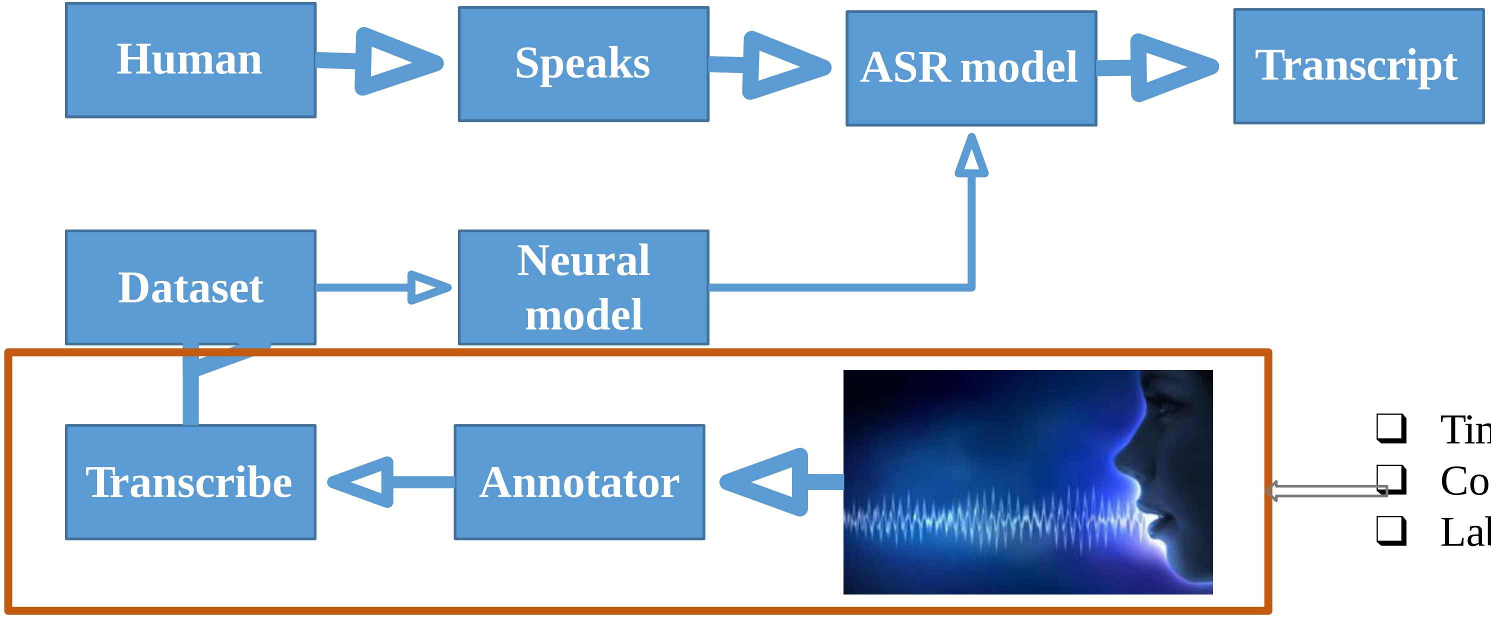
ASR Annotation Tool

Solution

Nagarathna R

Thishyan Raj T

Raja Ravi Teja Chaganti



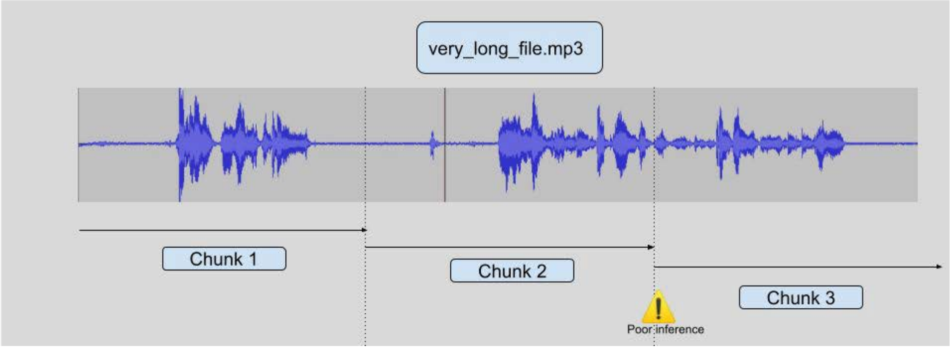
- Time Consuming
- Costly
- Labor-intensive

Aspects to look into while building a tool

- Automatic chunking of long audio
- ASR system
- Confidence estimation
- Recommendation system
- Interface

Long audios

- Raw audios are generally long audios
- It takes time for an annotator to chunk the audios by listening to the audios
- Long audio – Use VAD/AED to chunk audio at non-speech regions

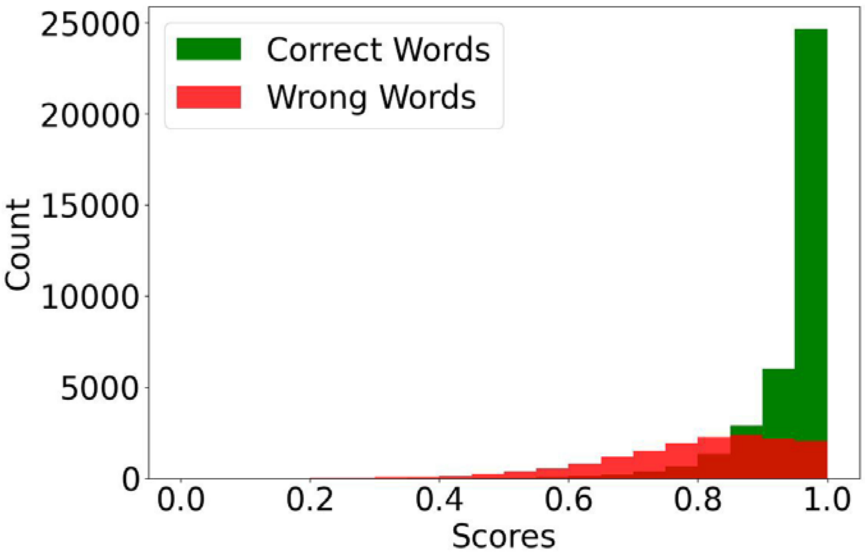


ASR Model

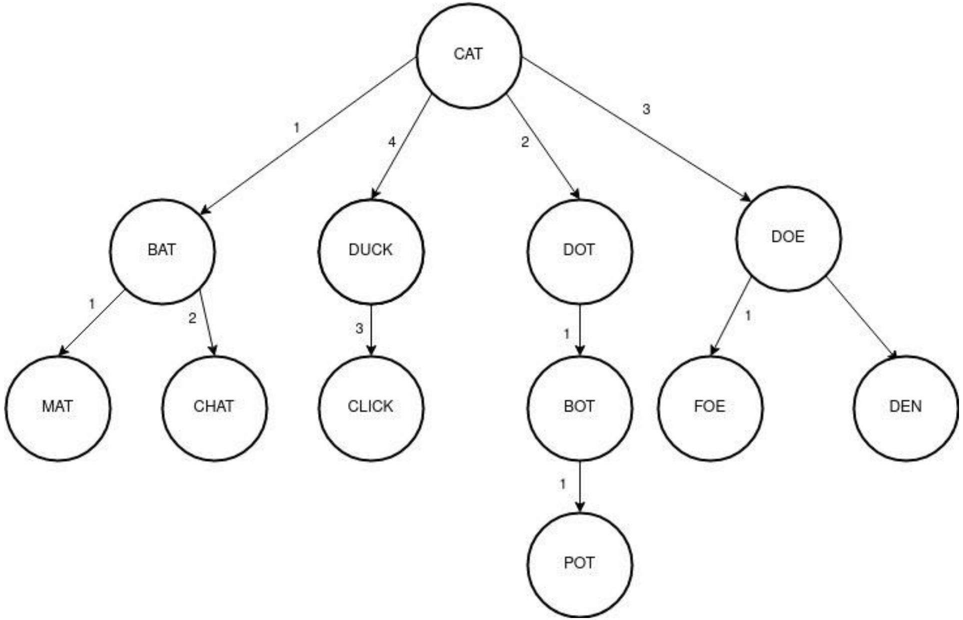
- Step 1 - Check for pre-trained models for the desired language
- Step 2 - If pre-trained model not available, find open source ASR datasets.
- Step 3 - If no dataset is available, manually transcribe a few hours of data.
- Step 4 - Take a pre-trained model. Use transfer learning to build an ASR model.

Confidence Estimation

- Prediction from neural network is over confident.
- A method is required to estimate the correctness of the predictions.
- Maximum class probability is usually high even for incorrect predictions
- There are various methods to estimate the confidence on the predictions:
    - Temperature scaling of the logits
    - Auxiliary model
    - Ensemble



Recommendation System



- Find alternative words of low confident words.
- Recommendation system
- Add words in a dictionary to a tree based on similarity metric. Find close matching words for less confident words.
- Train an auxiliary model to find the correct word based on context.

Interface

- Transcripts highlights as audio plays
- To correct the transcript, play the corresponding audio segment
- Highlight least confidence words to make quick corrections.
- Generate final transcript along with timestamps, convenient for users to chunk the long audio and create chunks for training the ASR system.

THANK YOU

Interactive singing melody estimation using active adaptation

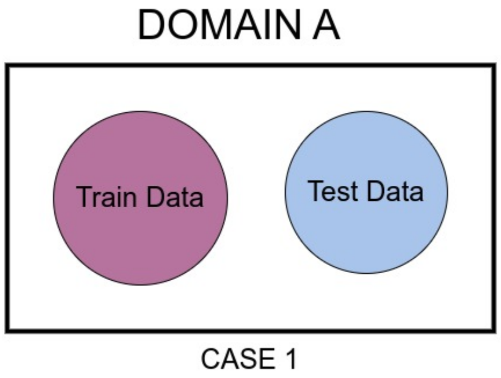
Kavya Ranjan Saxena

IIT Kanpur

Scenarios of Supervised Learning

There are different cases:

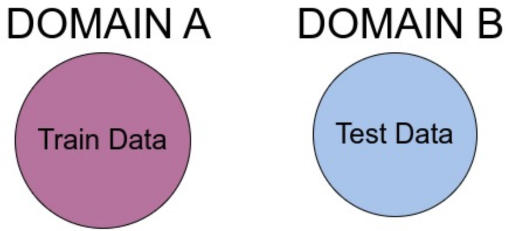
- CASE 1 - Train on one domain and test on the same domain



Scenarios of Supervised Learning

There are different cases:

- CASE 1 - Train on one domain and test on the same domain
- CASE 2 - Train on one domain – test on another domain
  - Different feature space
  - Same label space, no label shift



CASE 2

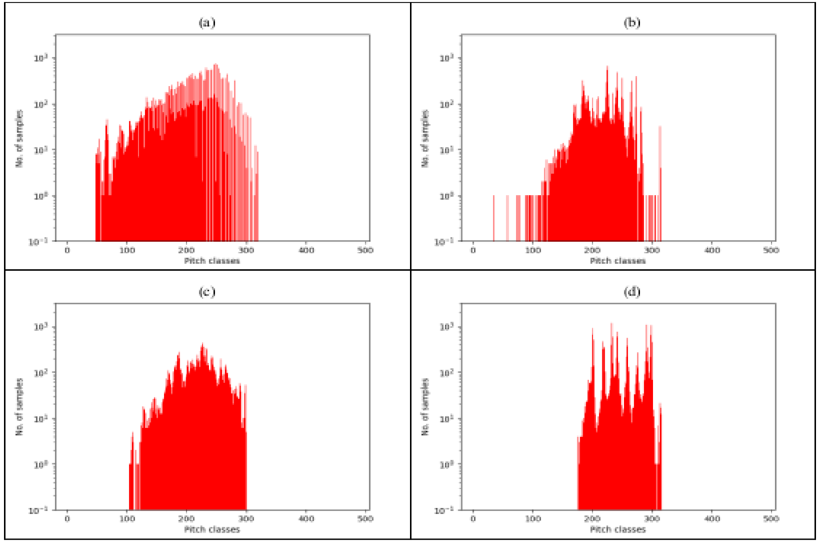
Scenarios of Supervised Learning

- CASE 3 – Train on one domain – test on another domain
  - Different feature space
  - Different label space and label shift

Scenarios of Supervised Learning

- CASE 3 – Train on one domain – test on another domain
  - Different feature space
  - Different label space and label shift

MELODY ESTIMATION!!



<u>Datasets</u>

a) MIR1K
b) ADC2004
c) MIREX05
d) HAR

Solution?

- Active adaptation.
- Train Data: MIR1K[1]
- Test Data: ADC2004[2], MIREX05[2], HAR[3]

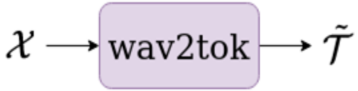
[1] https://sites.google.com/site/unvoicedsoundseparation/mir-1k
[2] http://labrosa.ee.columbia.edu/projects/melody/
[3] https://zenodo.org/record/8252222

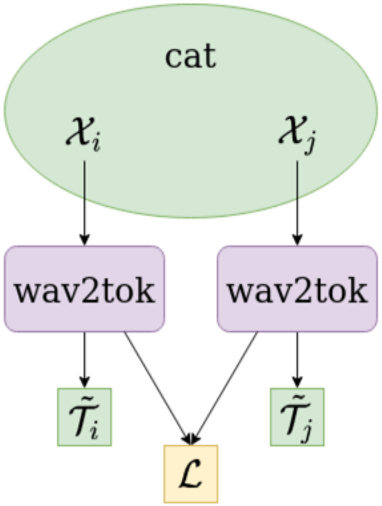
Slides for Today

Adhiraj Banerjee

wav2tok: Deep Sequence Tokenizer for Audio Retrieval

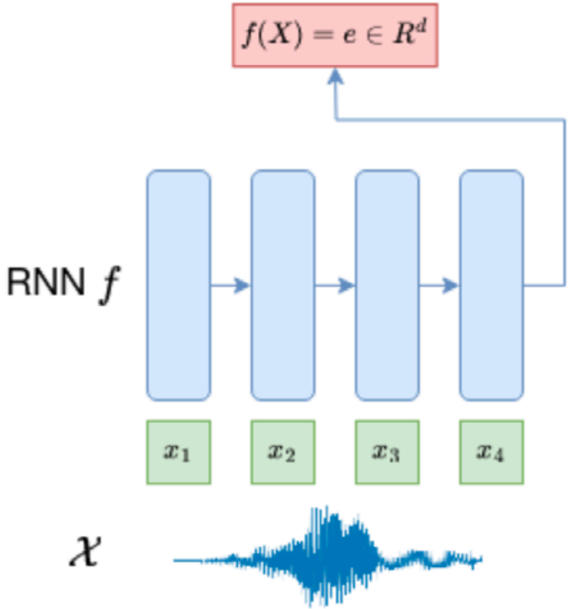
A model mapping audio $\mathcal{X}$ to discrete tokens $\tilde{\mathcal{T}}$





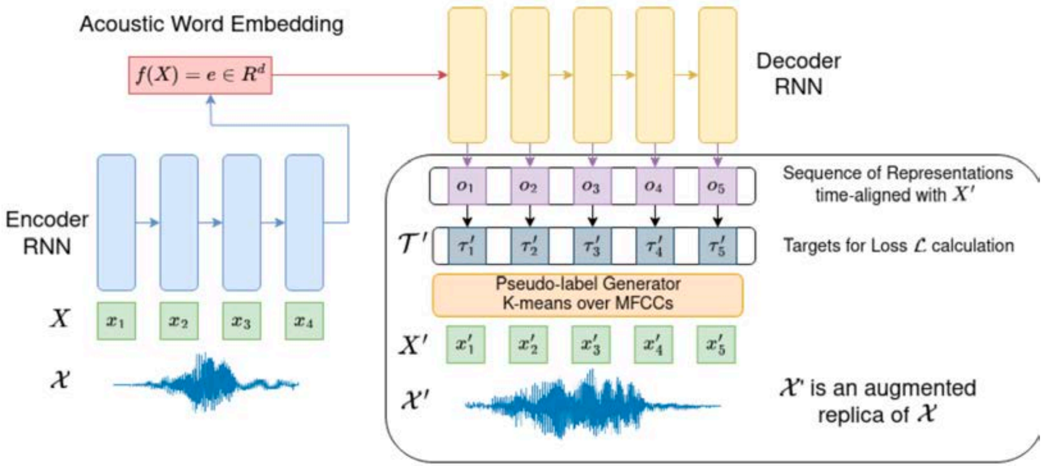
Model learns the tokens un-supervised from pairs of similar audio

Enc-Dec RNN Acoustic Word Embeddings learning via Pairwise Prediction



- NAWE models encode variable length acoustic feature sequences to a fixed dimensional embedding.
- Improves search time as two acoustic segments can be compared via calculation of cosine similarity between their embeddings.
- Allows us to consider a flexible set of features.

Enc-Dec RNN Acoustic Word Embeddings learning via Pairwise Prediction



Loss: $\mathcal{L} = \sum_{t=1}^{5} \log \frac{\exp(cos(Wo_t, e_{\tau'_t}))}{\sum_{k \in [K]} \exp(cos(Wo_t, e_k))}$

where $e_{\tau'_t}$ is an embedding corresponding to pseudo-label τ'_t

Thank You

UNSUPERVISED DOMAIN ADAPTATION FOR SOUND EVENT DETECTION IN MUSIC APPLICATIONS (ISMIR 2022 LBD)

Arkaprava Biswas

MS-R Student

IIT Kanpur

Sound Event Detection for non-overlapping audios (K-class):

- Use Synthetic Audio, and no labels of real audio.
- Learn class boundaries with labelled synthetic audio:

$$\min_{F,C_1,C_2} [L_{CE}(h_1(X_S),Y_S) + L_{CE}(h_2(X_S),Y_S)]$$

- Push class boundaries for real audio towards synthetic audio:

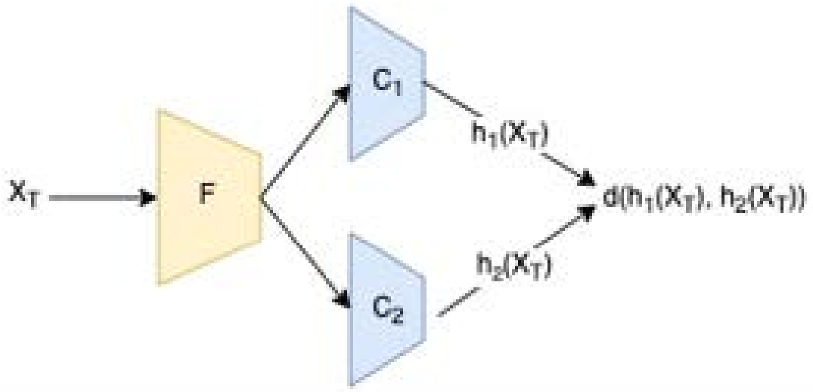
$$\min_{C_1,C_2}[L_{CE}(h_1(X_S),Y_S) + L_{CE}(h_2(X_S),Y_S) - L_{disc}(X_T)]$$

$$L_{disc}(X_T) = E_{x\sim X_T}[d(h_1(x),h_2(x))]$$

$$d(h_1(x),h_2(x)) = \frac{1}{K}\sum_K |h_1(x) - h_2(x)|$$

- Generate new features for real audio within newly formed class boundary:

$$\min_F L_{disc}(X_T)$$



Experiments for 10 classes

Table 1: Accuracy and F1 score obtained for the method and the baseline

Train data	Test data	Accuracy	F1
FSD+US	FSD+US	95.65%	87.8%
Audioset	Audioset	45.645%	38.8%
Without Adaptation			
FSD+US	Audioset	24.705%	21.35%
With Adaptation			
FSD+US	Audioset	40.55%	36.21%

Audio Search

Akshay Raina, Sagar Dutta

Acoustic Event Detection

Akshay Raina, Sayeedul I Sheikh, Vipul Arora

Automatic Detection and Analysis of Singing Mistakes for Music Pedagogy

Vipul Arora, Suraj Jaiswal, Akshay Raina, Sumit Kumar

Narottam: A Smart Platform for Music Education

Suraj Jaiswal, Vipul Arora

HarMIDI: Sensor System To Read MIDI from Indian Harmoniums

Suraj Jaiswal, Vipul Arora

See you for posters and demos!