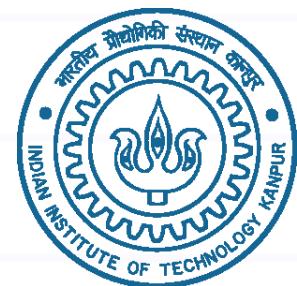


# Uncertainty Estimation for Trustworthy AI

---

Part 2: Real World Applications

Vipul Arora



# Speech

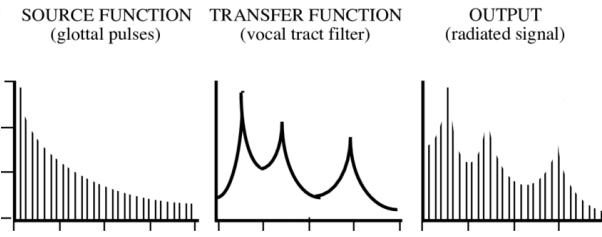
---



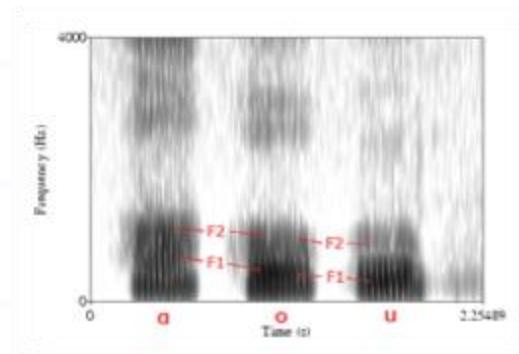
[Image from Amazon Science]

# Evolution of Speech Technologies

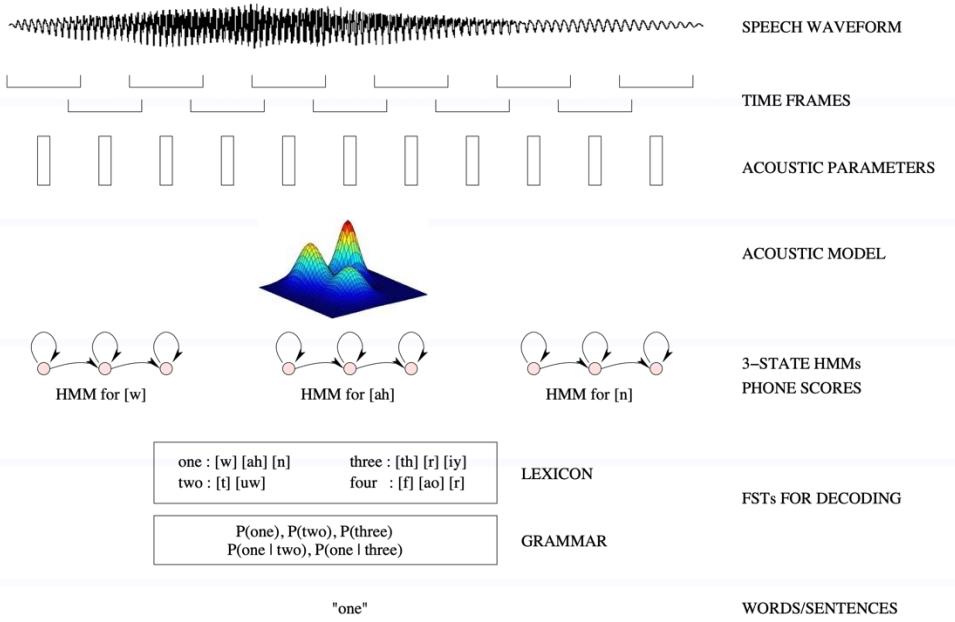
- Linguistic rules and signal processing
- GMM-HMM
- DNN-HMM



[Image from Jacques Koreman, 1995]



[Image from Ken Stevens, 2000]



# Machine Learning: end-to-end



Image from <https://distill.pub/2017/ctc/>



Image from <https://doi.org/10.1038/s41598-022-12260-y>

Hello, I am giving  
this talk

# Speech Technologies

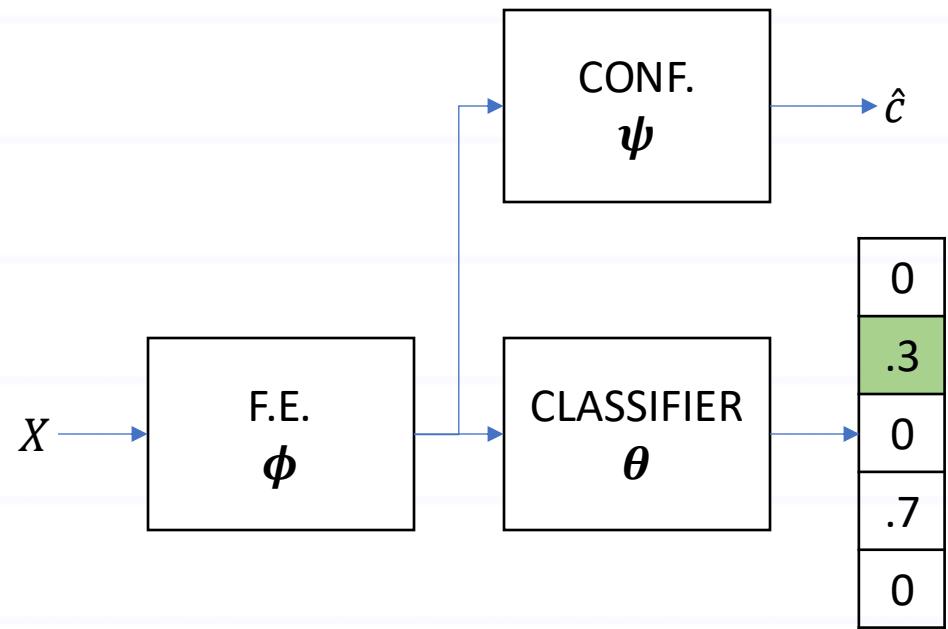
---

- Speech Recognition
- Emotion Recognition
- Clinical Diagnosis
- ...



# Uncertainty Estimation for Speech Technologies

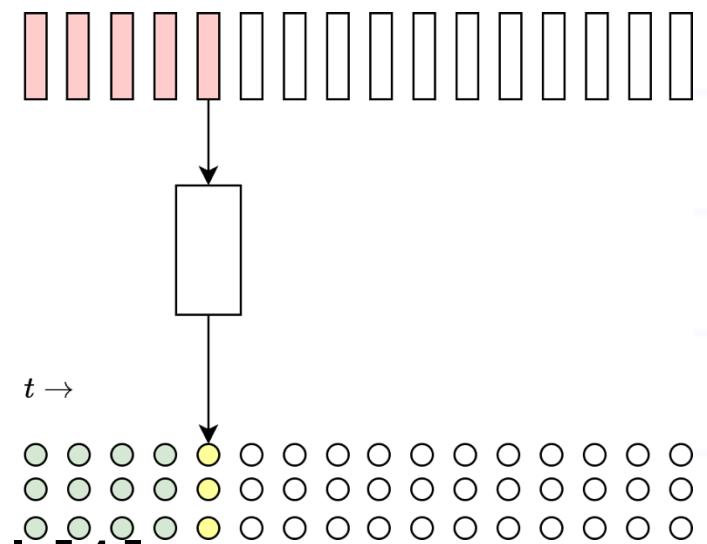
# Confidence Estimation Model (CEM)



- What should be the target for training Confidence Model?
- Binary:  $\mathbb{I}(\hat{Y}^* = Y)$
- True class probability:  
 $P(Y)$

# CEM for ASR

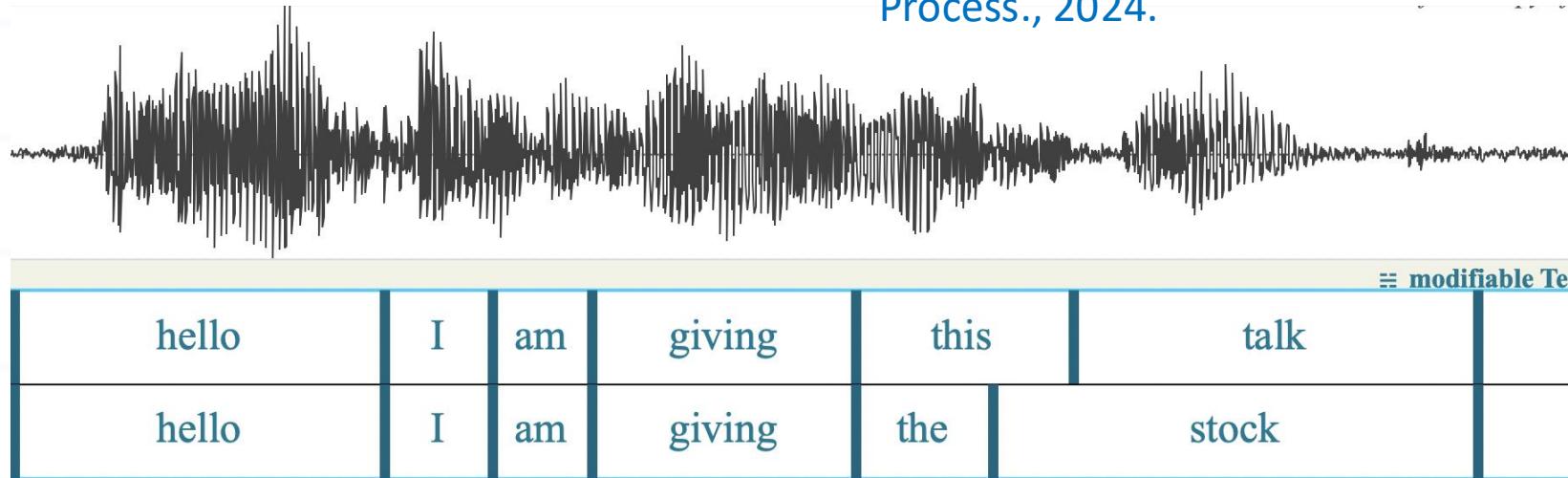
- Speech is a **time series**
- Word-level confidence estimation model [1]
  - Hello, I am giving **the stock**
  - 1, 1, 1, 1, 0, 0
  - Minimize binary cross entropy w.r.t. 0/1 target



1. Q. Li et al., "Confidence Estimation for ... Speech Recognition," ICASSP 2021

# TeLeS

N. Ravi, Tishyan R. T, and V. Arora, "TeLeS: Temporal Lexeme Similarity Score to Estimate Confidence in End-to-End ASR." IEEE Trans. Audio Speech Lang Process., 2024.



- Match word boundaries

$$c^T = \begin{cases} \max \left( 0, 1 - \frac{|w_i^{ST} - \hat{w}_j^{ST}| + |w_i^{ET} - \hat{w}_j^{ET}|}{|w_i^{ET} - w_j^{ST}|} \right) & \text{if } g_p = C, S \\ 0 & \text{if } g_p = I, D \end{cases}$$

# TeLeS

---

- Treat time series as bag-of-words
- $|\{this\} \cap \{\text{the}\}| = 2$
- $|\{this\} \cup \{\text{the}\}| = 5$

$$c^L = \begin{cases} \frac{|w_i \cap \hat{w}_j|}{|w_i \cup \hat{w}_j|} & \text{if } g_p = C, S \\ 0 & \text{if } g_p = I, D \end{cases}$$

- $\frac{c^T + c^L}{2}$  is the target for CEM

# Results

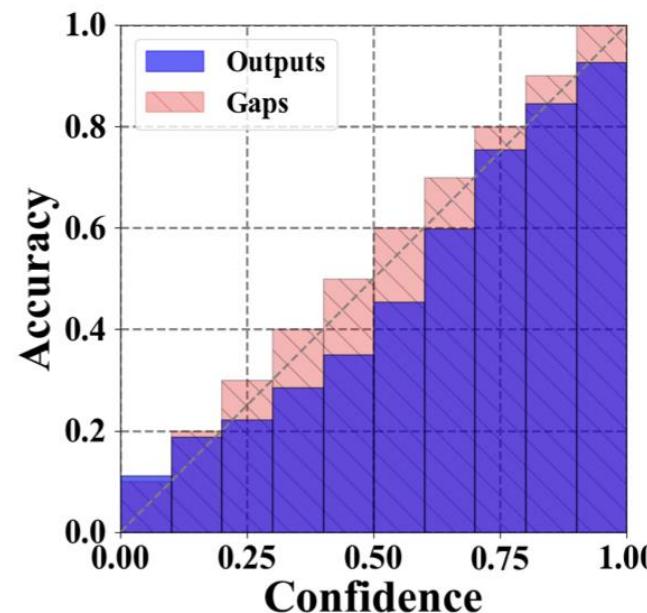


TABLE V: TeLeS-WLC and SOTA - Hindi PB Testset

Metrics	Class-Prob	Entropy	Binary	TeLeS
MAE ↓	0.5026	0.3998	0.2650	<b>0.1817</b>
KLD ↓	0.7912	0.6333	1.1799	<b>0.1785</b>
JSD ↓	0.2313	0.1818	0.1920	<b>0.0467</b>
RMSE-WCR ↓	0.2883	0.2684	0.2593	<b>0.1457</b>
NCE ↑	-0.2641	-0.0100	-0.0055	<b>0.1363</b>
ECE ↓	0.2472	0.2253	0.2627	<b>0.0260</b>
MCE ↓	0.3904	0.3663	0.4294	<b>0.1011</b>

Acquired Data Proportion	Path Probability (WER/CER ↓)	SMCA (WER/CER ↓)	LMC (WER/CER ↓)	TeLeS (WER/CER ↓)
1/10	43.53/15.22	59.97/23.01	92.55/53.16	<b>32.96/11.09</b>
1/7	31.71/10.59	33.09/11.00	36.93/12.42	<b>28.78/9.51</b>

# Impact

- Active Learning
- 10x faster annotation workflows

The screenshot shows a digital interface for audio annotation. At the top is a pink waveform visualization of an audio recording. Below it is a control bar with icons for play, pause, stop, and volume, along with a search function and settings. The main area displays three segments of transcribed text with corresponding timestamps and a list of changes.

**Segments:**

- 00:00:00,001 --> 00:00:15,000:  
प्रसार भी अभिनिखगाकीप्रस्थि बाहर सुधरे खजाना ऊंची लंबे तकरीबन रामगांधमीआकपहने  
प्रसार भी अभिनिखगाकीप्रस्थि बाहर सुधरे खजाना ऊंची लंबे तकरीबन रामगांधमीआकपहने
- 00:00:26,001 --> 00:00:40,000:  
या से जमू कश्मीर की चंडगिनिचुनीशक्सितों में से एक है शहर की अग्नि समाजी मई किलों में  
आप उन्हें अक्सर देख सकते हैं शास्त्रीक
- 00:00:40,001 --> 00:00:52,000:  
जमू में ही हुआ वरसकेहोचुकेहैंगर की एक ऐसी जानी पहचानी शख्सियत जिन्हें देखकर अधिक  
सर झुक जाता है
- 00:00:52,001 --> 00:01:07,000:

**Changes Log:**

- make changes is 00:11:00,001 --> 00:11:15,000  
स्कूल से कॉलेज की तरफ आना और मैं जानता हूं कि जमू में जब हम...
- make changes is 00:21:14,001 --> 00:21:26,000  
आपने हमें अपने पढ़ाने के बारे में तो बताया अपनी नौकरी के बारे में बताया...
- make changes is 00:01:48,001 --> 00:02:02,000  
इनके जीवन के बारे में उस समय के जमू के बारे में इनके बचपन इनकी...
- make changes is 00:01:36,001 --> 00:01:48,000  
लेवल तक पहुंची है उस सफर के और उसके साक्षी प्रोफेसर रामनाथ शा...
- make changes is 00:29:16,001 --> 00:29:30,000  
सभी आप सब अच्छी घटना की आजादी मिल गई दुर्घटना के एक देश के ...
- make changes is 00:29:57,001 --> 00:30:09,000

## project5

The screenshot shows a digital audio workstation (DAW) interface with a purple header bar containing the text "AUDACITY" and "Audacity". Below the header is a toolbar with various icons. The main area features a large pink waveform representing an audio recording. A horizontal timeline at the bottom has numerical markers from 0.00 to 1.00. Several text annotations in Hindi are overlaid on the waveform:

- A pink box highlights a section of the waveform from approximately 0.00 to 0.50. The text inside reads:

मार्गों पर चढ़ गीर तक यही भा यहां है कहुं पर तुम मां है देखनी में यही अगले के  
लिए जो लावी धिक्कत ही नहीं है अपने जन्म से जन्मायी बहु देखे बहु और कल जन्माय  
देखी बहास का इस्तर जगते धिक्कत यहां है ऐसे निष्ठाक बद दूसर जन्म में उभार अब यही  
है जो जन्मती है कि दूसरे जन्म से जन्माये हो जाते हैं आजी पुरा जन्माया फैसल इस  
प्रथा दूसरा हुआ जन्म नहीं यह गुरु है
- A white box highlights a section from 0.50 to 0.70. The text inside reads:

ही जन्माया की जन्मती अपने दूसरे रही है जहां यहा जन्माया बहिर्य के बाद यही भर रहा है  
जिसे दूसरों में यही जन्मा देता
- A white box highlights a section from 0.70 to 0.85. The text inside reads:

जन्मे की जन्मती में लोटों के भूतों में यही भरा है लोटों की दृश्यते में यही भर रहा है लोटों  
को जीवे जीवे में धिक्कत ही नहीं है इतनीक जाप लोरी जहां के लिए जिस जन्मती की गई है
- A white box highlights a section from 0.85 to 1.00. The text inside reads:

जन्मती जीवे जीवे की जन्मती में अनुष्ठ बहिर्य जन्मती जीवे जीवे

# Active Learning for different languages

---

Acquired Data Proportion	Path Probability (WER/CER ↓)	SMCA (WER/CER ↓)	LMC (WER/CER ↓)	TeLeS (WER/CER ↓)
1/10	43.53/15.22	59.97/23.01	92.55/53.16	<b>32.96/11.09</b>
1/7	31.71/10.59	33.09/11.00	36.93/12.42	<b>28.78/9.51</b>

## Hindi

Acquired Data Proportion	Path Probability (WER/CER ↓)	SMCA (WER/CER ↓)	LMC (WER/CER ↓)	TeLeS (WER/CER ↓)
1/10	50.05/10.52	49.63/10.43	50.64/10.71	<b>47.38/9.83</b>
1/7	45.62/9.28	45.11/9.13	45.49/9.19	<b>44.35/8.87</b>

## Tamil

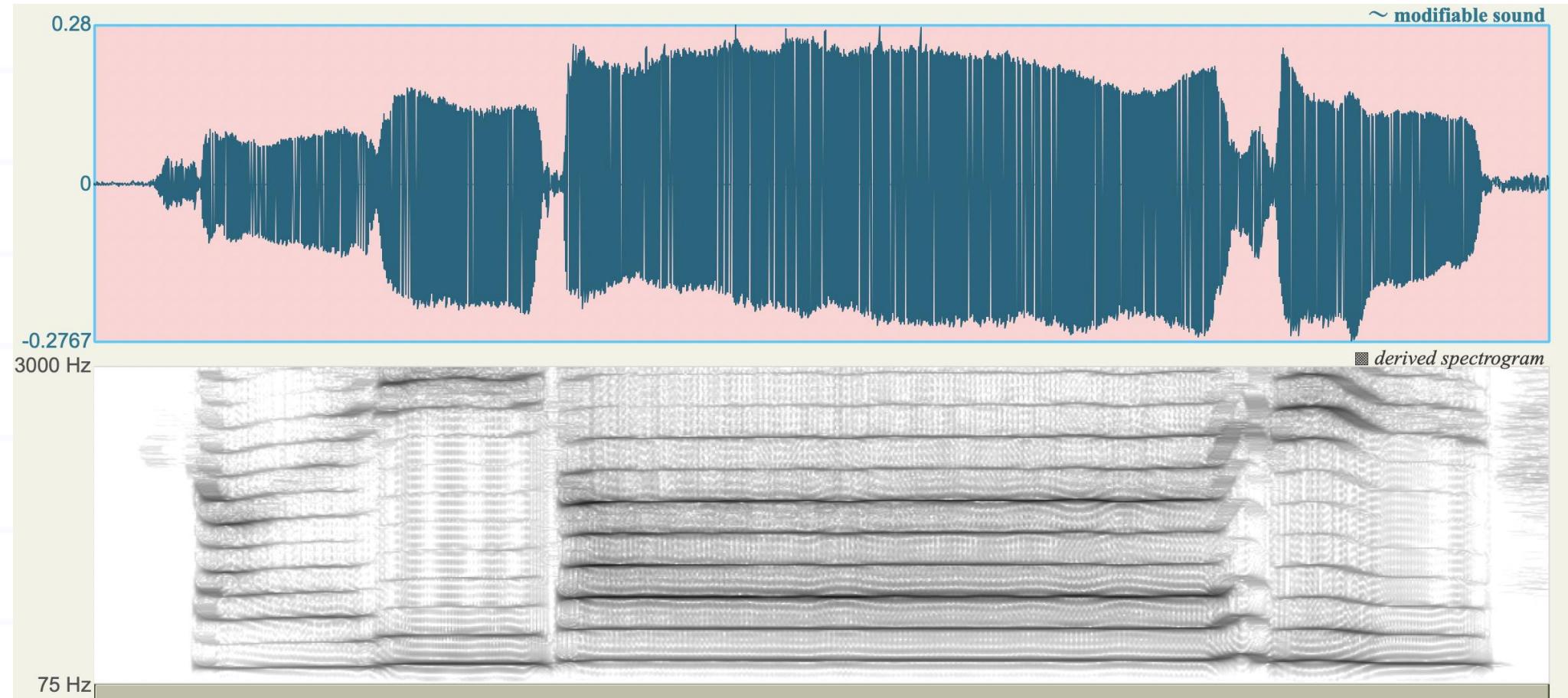
Acquired Data Proportion	Path Probability (WER/CER ↓)	SMCA (WER/CER ↓)	LMC (WER/CER ↓)	TeLeS (WER/CER ↓)
1/10	54.29/12.22	99.97/85.58	99.95/93.97	<b>47.98/10.71</b>
1/7	43.75/9.57	60.87/13.67	74.47/17.69	<b>41.91/9.15</b>

## Kannada

# Uncertainty Estimation for Music Analysis

Melody Estimation

# Singing



MADHAV

# Polyphonic Music

---



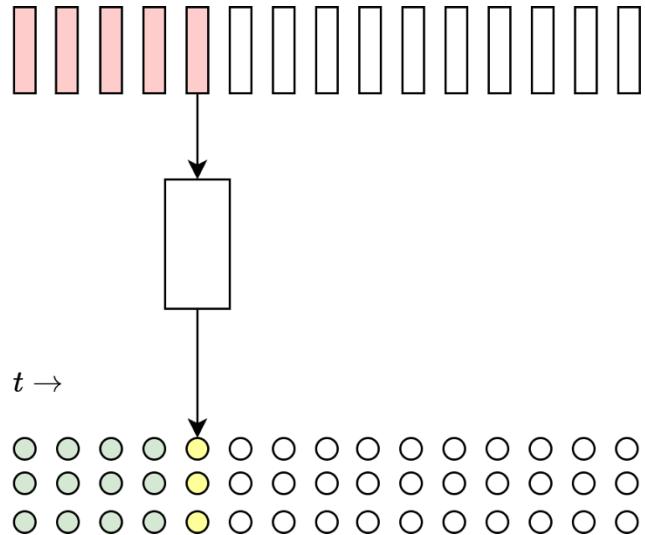
[This Photo](#) by Unknown Author is licensed  
under [CC BY-NC-ND](#)



[This Photo](#) by Unknown Author is licensed  
under [CC BY-NC](#)

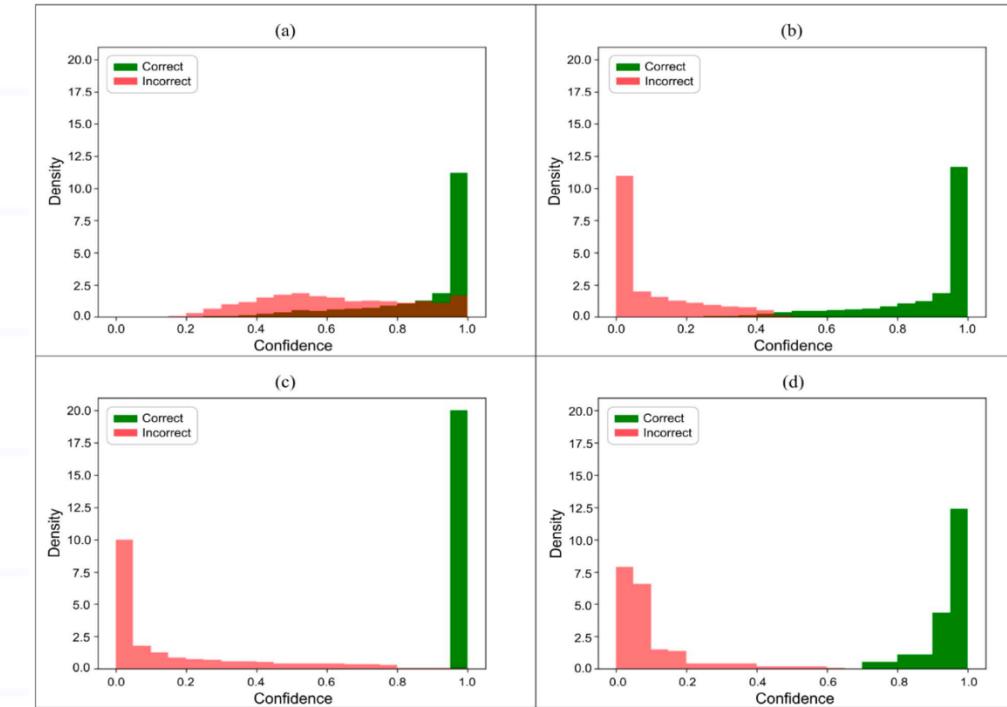
# Melody Estimation

K. R. Saxena and V. Arora, "Interactive Singing Melody Extraction Based on Active Adaptation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 32, pp. 2729–2738, 2024, doi: [10.1109/TASLP.2024.3399614](https://doi.org/10.1109/TASLP.2024.3399614).



max class prob.

Normalized  
true class prob.

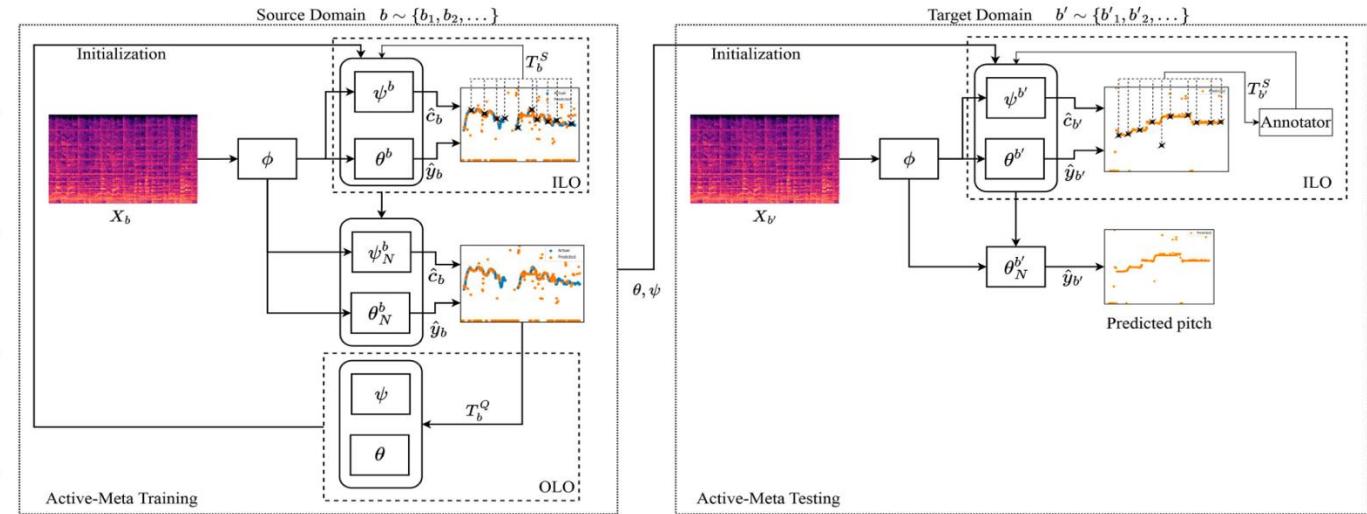


true class prob.

CEM trained with  
Normalized  
true class prob.

# Active Learning with CEM

- Actively acquire samples to update the model with Meta-learning



# Dealing with heavy class imbalance

---

- **Dynamic class weights** in the cross-entropy loss during meta learning
- These weights change from episode to episode

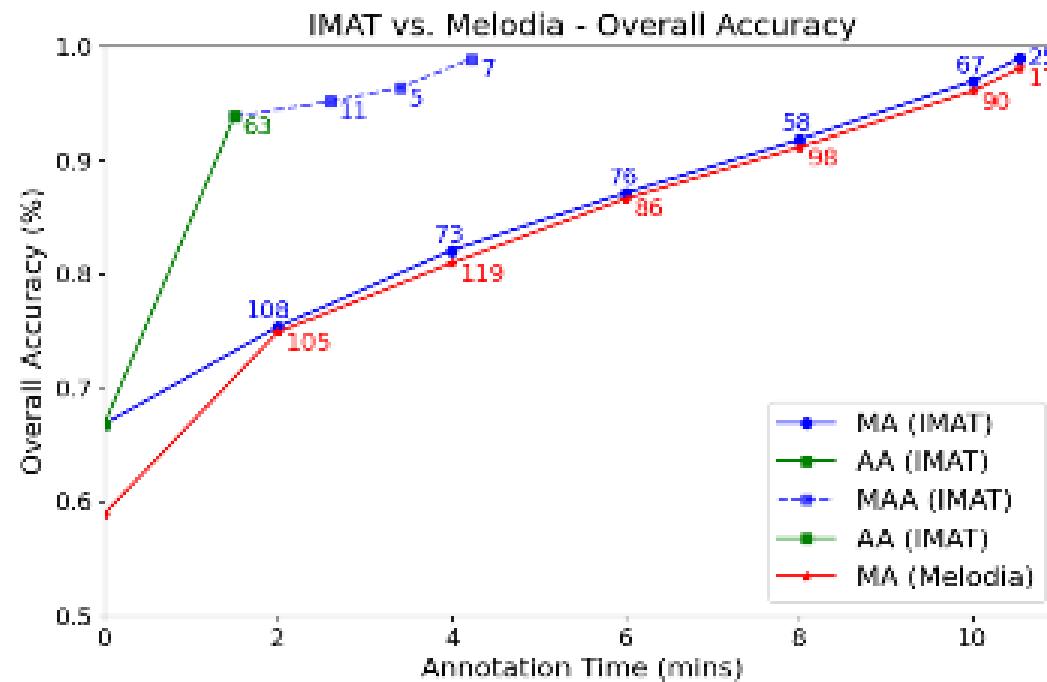
# Comparison with other adaptation methods

---

Experiments				ADC2004			MIREX05			HAR		
Method	MW	AA	RA	RPA	RCA	OA	RPA	RCA	OA	RPA	RCA	OA
FT-RA	-	X	✓	80.34	81.45	80.98	81.16	81.98	82.10	76.45	77.10	76.88
FT-AA	-	✓	X	81.55	81.59	81.66	81.80	81.78	81.85	76.95	77.34	76.17
MAML-RA	X	X	✓	81.10	82.56	81.41	83.16	84.57	83.28	77.70	78.12	78.10
MAML-AA	X	✓	X	81.32	82.56	81.99	82.12	83.88	81.80	75.78	76.55	75.98
w-AML(Ours)	✓	✓	X	<b>86.40</b>	<b>87.01</b>	<b>86.15</b>	<b>87.23</b>	<b>88.15</b>	<b>87.80</b>	<b>80.60</b>	<b>80.99</b>	<b>81.45</b>

# Experiments

---





# Code Repository and Tutorial

- Saxena and Arora, “Interactive singing melody extraction based on active adaptation.” IEEE TASLP 2024.
- Saxena and Arora, “IMAT”, submitted.
- [https://github.com/madhavlab/2024\\_imat](https://github.com/madhavlab/2024_imat)
- <https://youtu.be/1SSQOE8CIHE?si=ktSEyapmJvTaWsZ5>



[Browse](#) [Annotate](#) [Team](#) [MadhavLab](#)

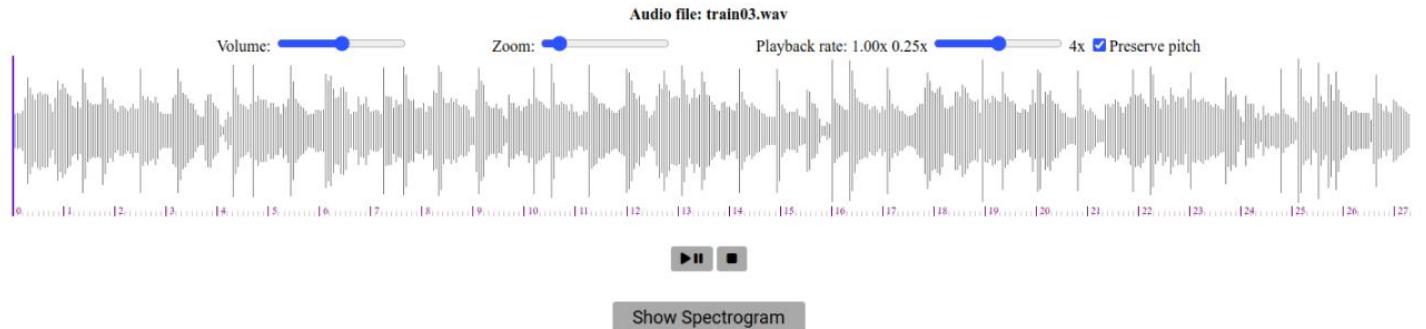
Interactive Interface for Singing Melody Extraction and Annotation

Browse audio file that you wish to annotate

[Upload audio file \(.wav\)](#)



(a)



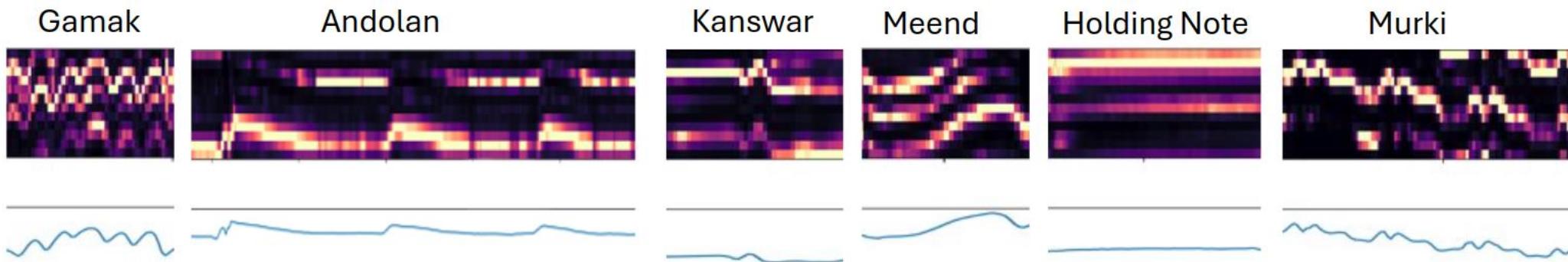
(b)

# Uncertainty Estimation for Music Analysis

Ornaments and Raga

# Ornamentation Detection

submitted to TASLP



## Recognizing Ornamentations in Singing Voices in Indian Art Music

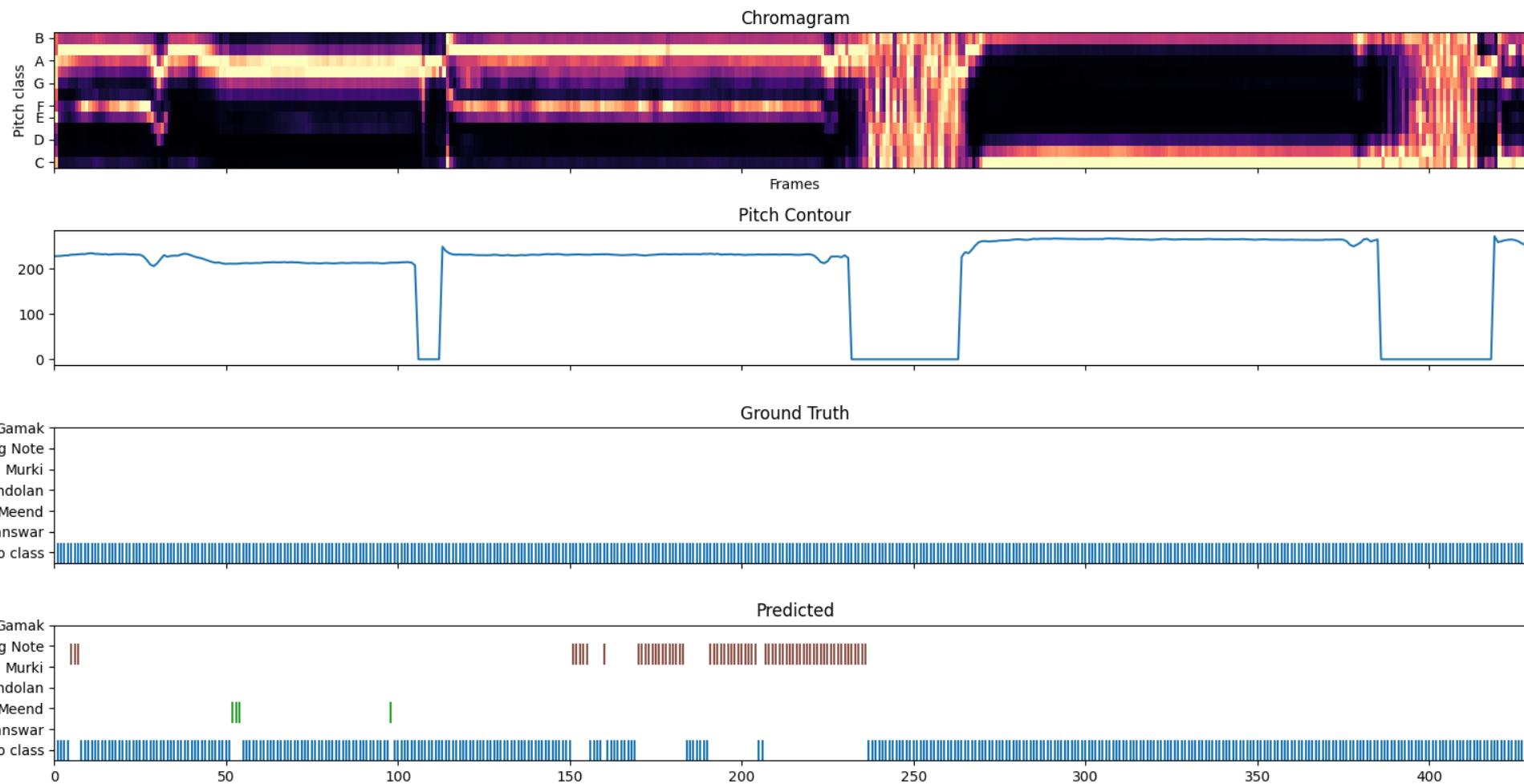
Sumit Kumar and Vipul Arora

*Department of Electrical Engineering, Indian Institute of Technology, Kanpur, India<sup>a</sup>*

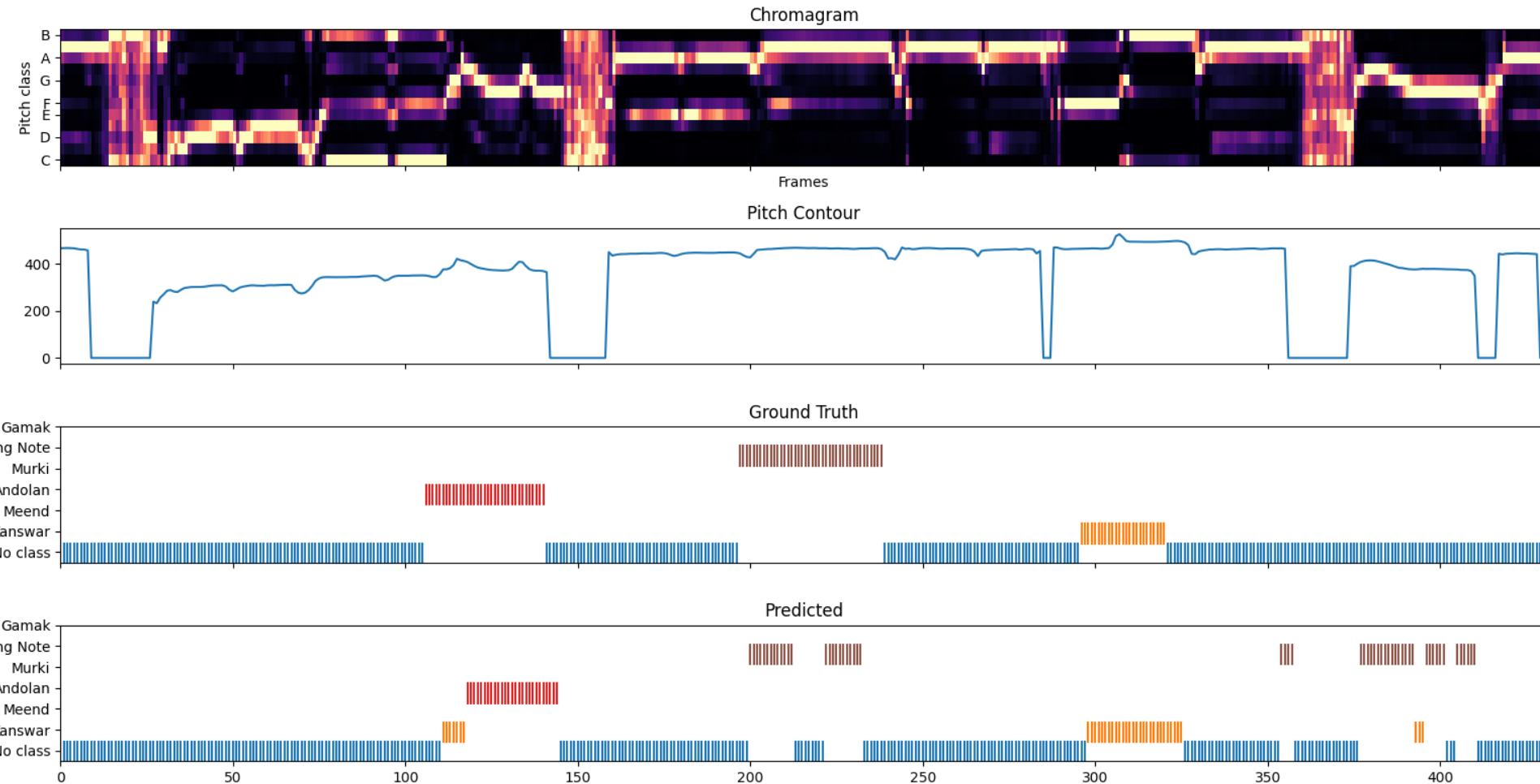
(Dated: 15 November 2024)

In this work, we study ornamentation or singing techniques in Indian art music. The identification of ornamentations not only strengthens the understanding of music but also opens avenues for applications such as singer identification, genre classification, controlled singing voice generation, and music teaching. We build a dataset named ROD: *Rāga* Ornamentation Detection dataset comprising audio recordings of singing with various ornamentations along with manual annotations (strong labels). Next, we propose a machine learning method to detect ornaments in audio using time-dilated convolution networks. During audio chunking,

# Holding note

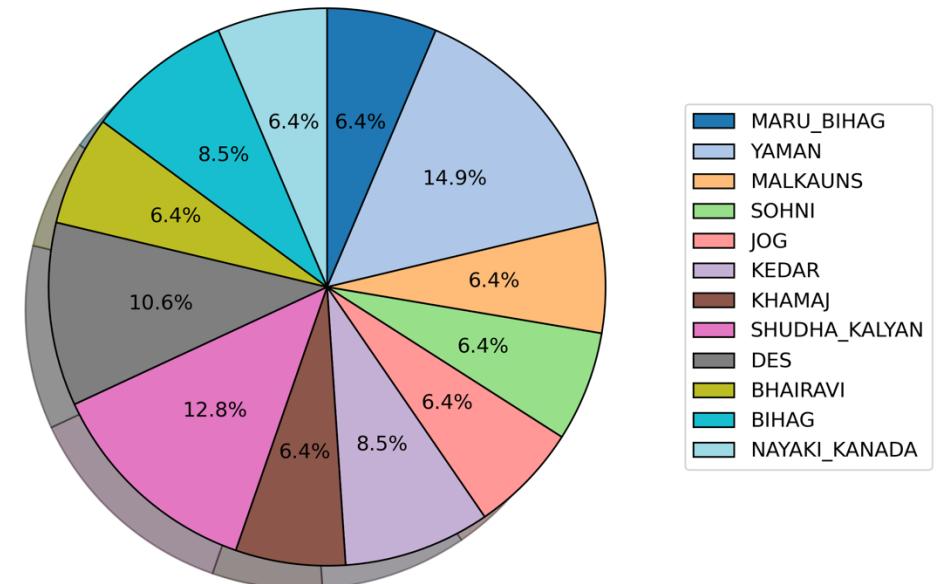


# Andolan or Kanswar



# Raga

- Melodic framework to incite specific emotions



# Need

---

- Music recommendation
- Music search
- Music teaching (incorrect evaluations are misleading)

# Challenges

---

- Limited/no labeled data
- Annotation is difficult and experts are rare

# Ornamentation Detection

- F1 score = 58%

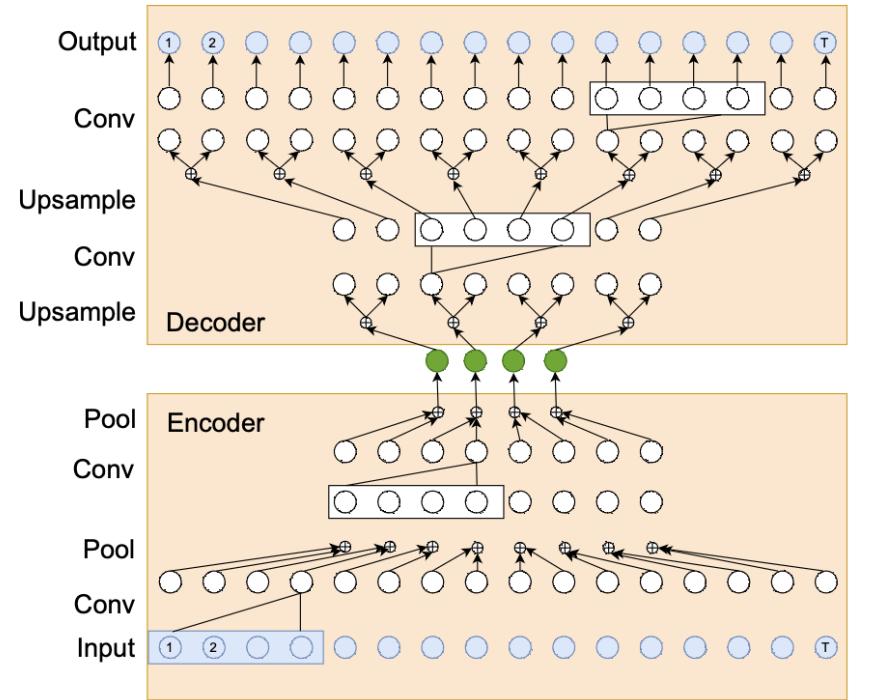


FIG. 6: Encoder-Decoder Temporal Convolutional Network

# Raga Classification

- Classify 12 Ragas
- Estimate tonic
- Normalize chromagram
- Achieves 89% F1-score

## Raga Identification

Youtube Video Link \*

Model \*

Ragas: Bhairavi, Bihag, Des, Jog, Kedar, Khamaj, Malkauns, Maru\_Bihag, Nayaki\_kanada, Shuddha, Sohni, Yaman

Start Time (seconds)

End Time (seconds)

**Run Script**

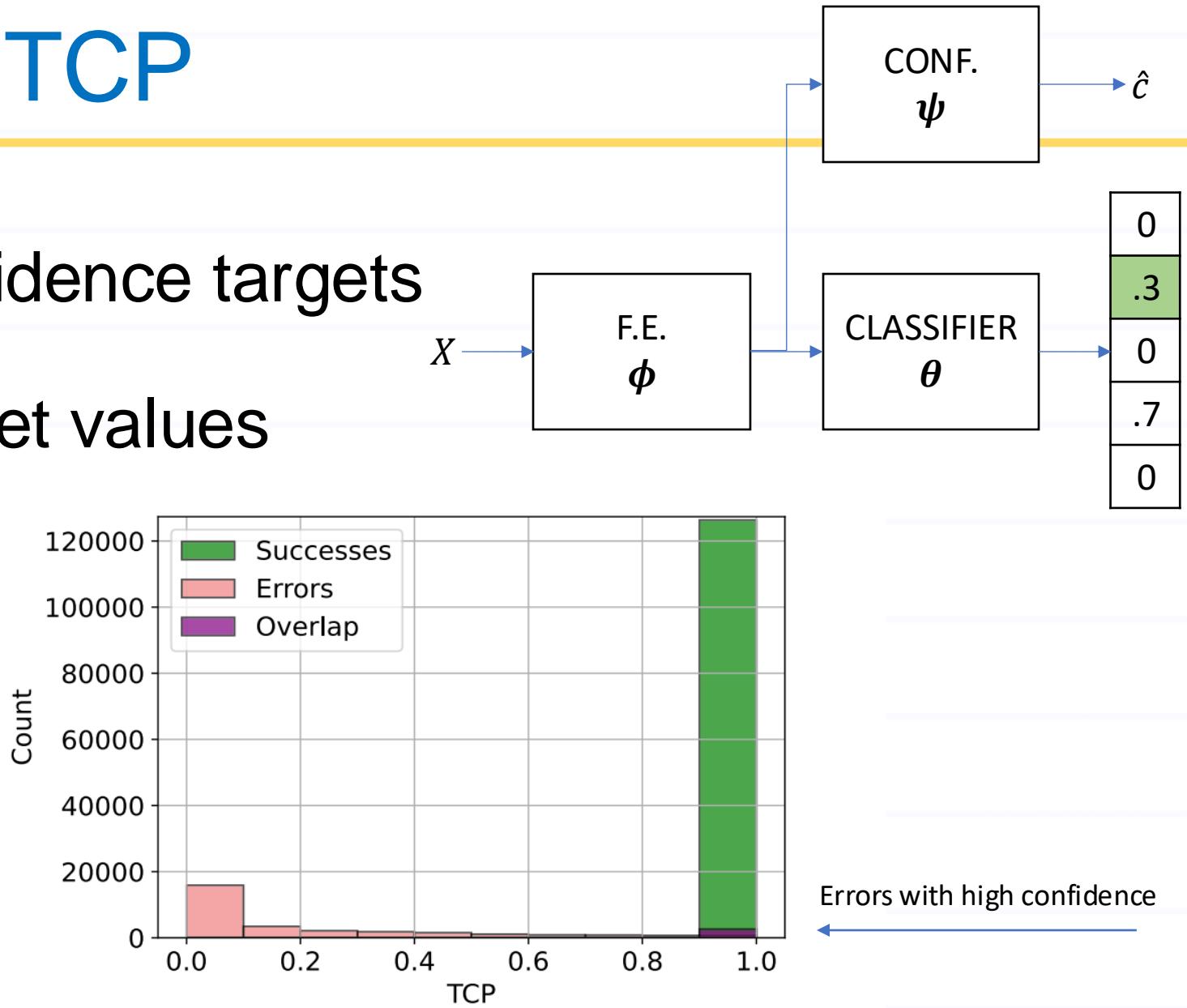
Show log stream

Downloading...  
Removing Speech and making chunks...  
Finding Tonics now...  
Raag = Bhairavi  
Confidence = 90.0%

# Limitations of TCP

- Ambiguity in confidence targets
- Imbalance in target values

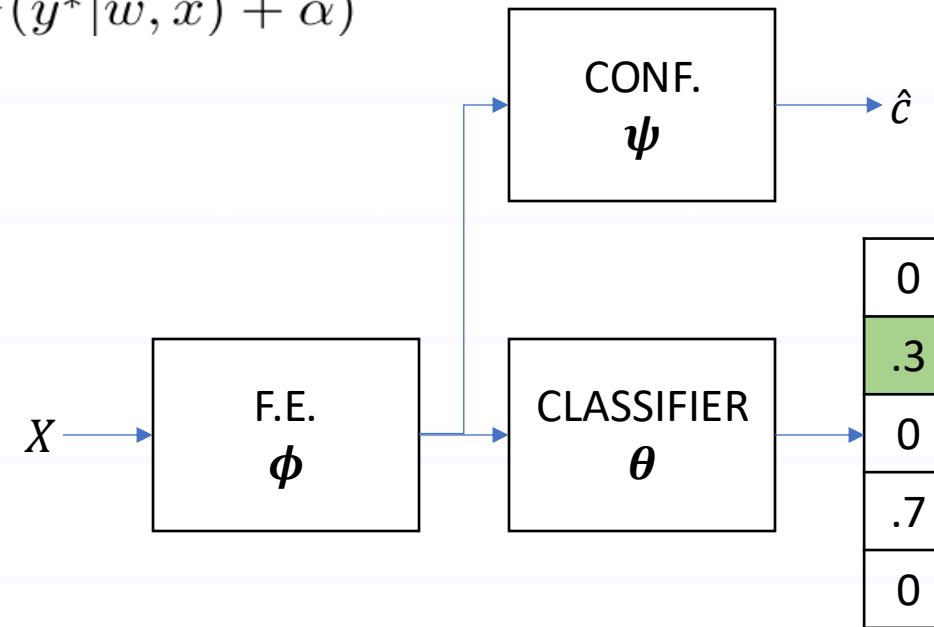
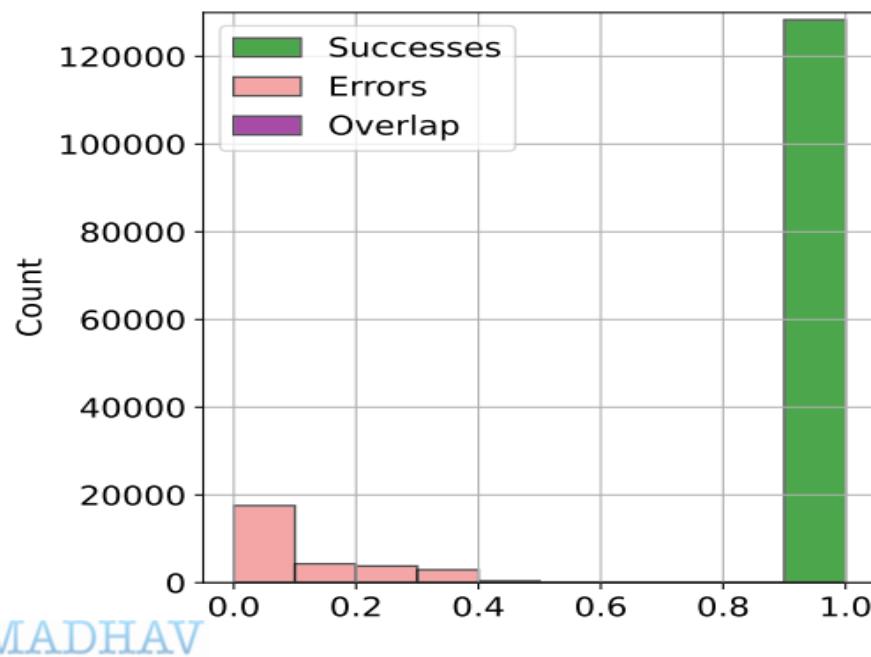
$$\text{TCP}^n(\mathbf{x}, y) = \frac{P(Y = y^* | \mathbf{w}, \mathbf{x})}{P(Y = \hat{y} | \mathbf{w}, \mathbf{x})}.$$



# Confidence Estimation Model

- Modified True Class Probability

$$\text{TCP}^{n*}(\mathbf{x}, \hat{y}, y^*) = \frac{P_Y(y^*|w, x)}{P_Y(\hat{y}|w, x) + \mathbb{I}[y^* \neq \hat{y}](P_Y(y^*|w, x) + \alpha)}$$



# Results

Sumit, Parampreet and Vipul,  
ICASSP 2025 workshop

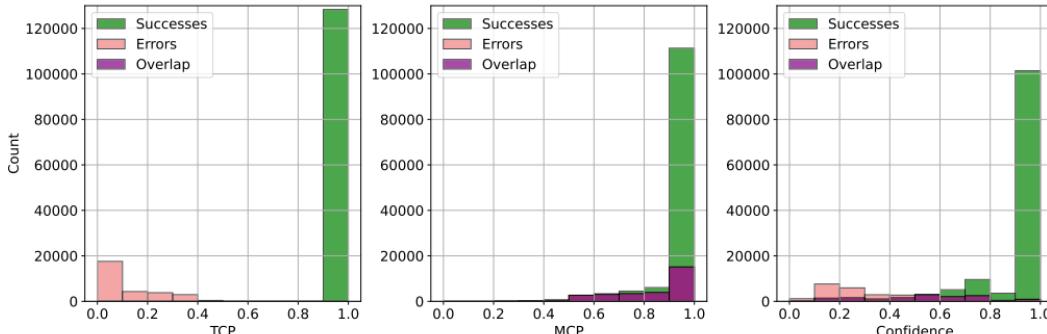


Fig. 3. (a) Proposed TCP<sup>n\*</sup> (b) MCP (c) Predicted confidence for Ornamentation Detection Task

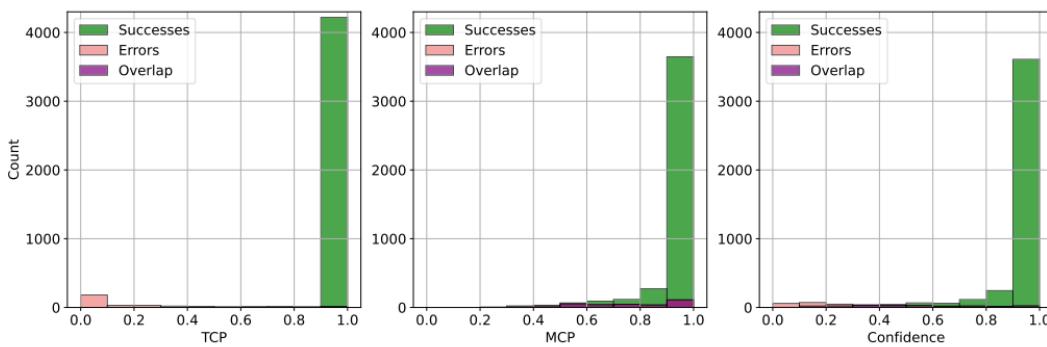


Fig. 4. (a) Proposed TCP<sup>n\*</sup> (b) MCP (c) Predicted confidence for Rāga Identification Task

Task	Models	Metrics			
		Precision	Recall	F-1	ECE
<i>Rāga</i>	Baseline	0.89	0.90	0.89	-
	Confidnet	0.89	0.90	0.90	0.04
	Proposed Model	<b>0.92</b>	<b>0.93</b>	<b>0.92</b>	<b>0.03</b>
Ornamentation	Baseline	0.53	0.72	0.58	-
	Confidnet	0.56	0.72	0.60	0.08
	Proposed Model	<b>0.62</b>	<b>0.74</b>	<b>0.66</b>	<b>0.05</b>

# Further Reading

---

- Ravi et al., TeLeS: Temporal Lexeme Similarity Score to Estimate Confidence in End-to-End ASR, IEEE TASLP 2024
- Saxena and Arora, Interactive Singing Melody Extraction Based on Active Adaptation, IEEE TASLP 2024  
([www.github.com/madhavlab/2023\\_interactiveMelEx](https://www.github.com/madhavlab/2023_interactiveMelEx))
- Kumar et al., Confidence-Enhanced Models for Indian Art Music Analysis, WIMAGA workshop, ICASSP 2025