Uncertainty Estimation for Trustworthy AI

Part 2: Real World Applications

Vipul Arora

MADHAV lab
Machine Analysis of Data for Human Audition and Vision

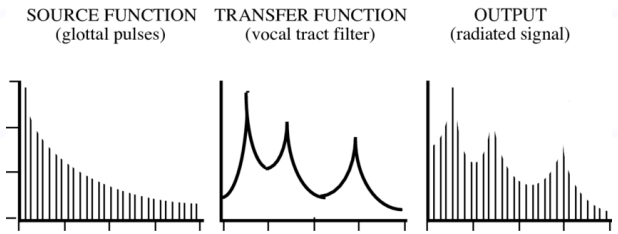
भारतीय प्रौद्योगिकी संस्थान कानपुर
INDIAN INSTITUTE OF TECHNOLOGY KANPUR

Speech

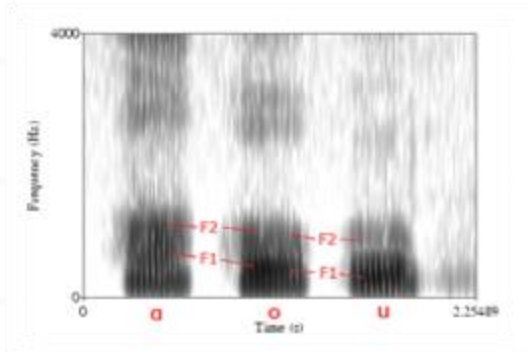
[Image from Amazon Science]

Evolution of Speech Technologies

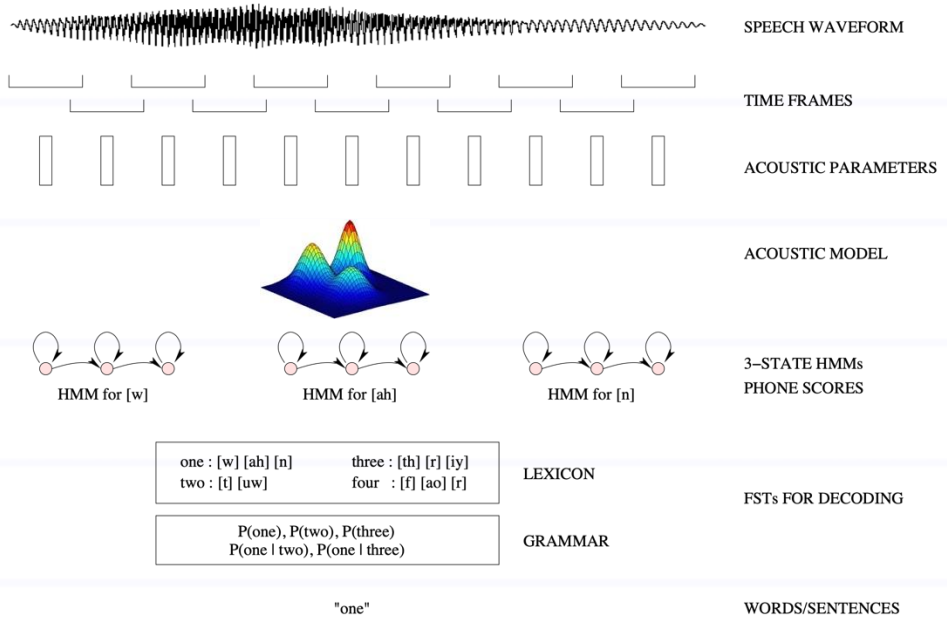
- Linguistic rules and signal processing
- GMM-HMM
- DNN-HMM



[Image from Jacques Koreman, 1995]



[Image from Ken Stevens, 2000]



Machine Learning: end-to-end

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

Hello, I am giving
this talk

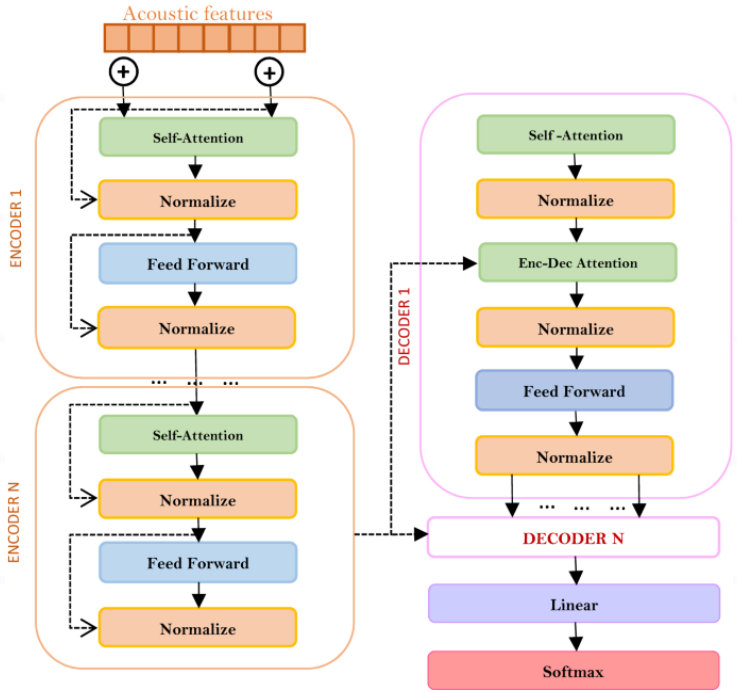


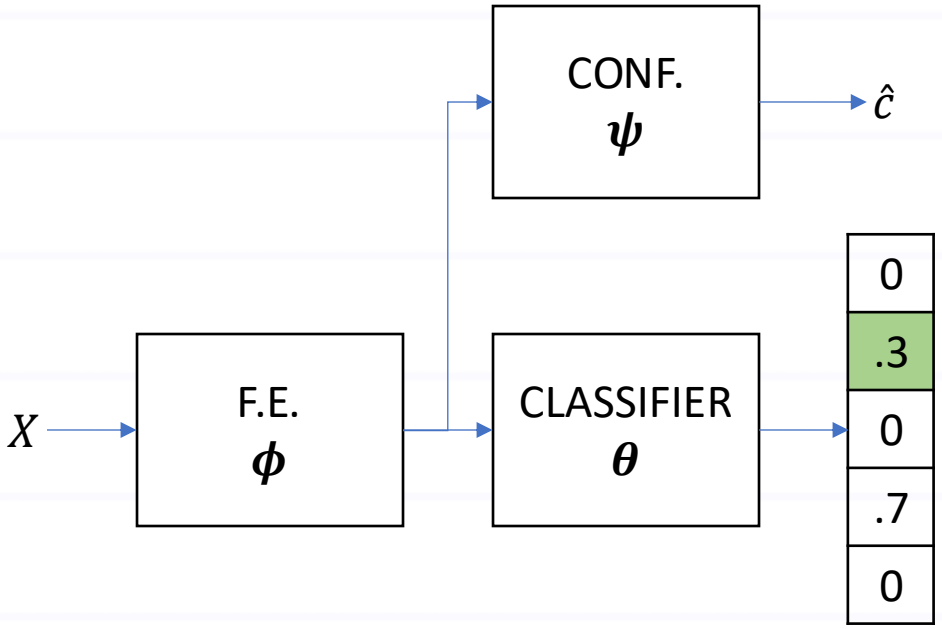
Image from https://doi.org/10.1038/s41598-022-12260-y

Speech Technologies

- Speech Recognition
- Emotion Recognition
- Clinical Diagnosis
- …

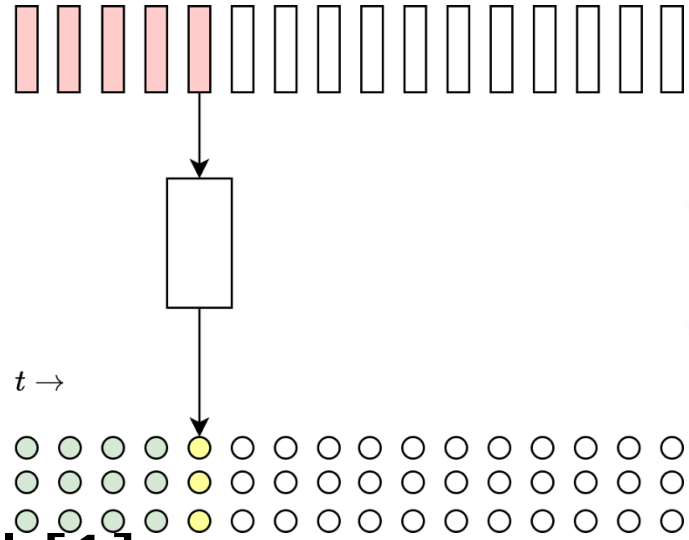
Uncertainty Estimation for Speech Technologies

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

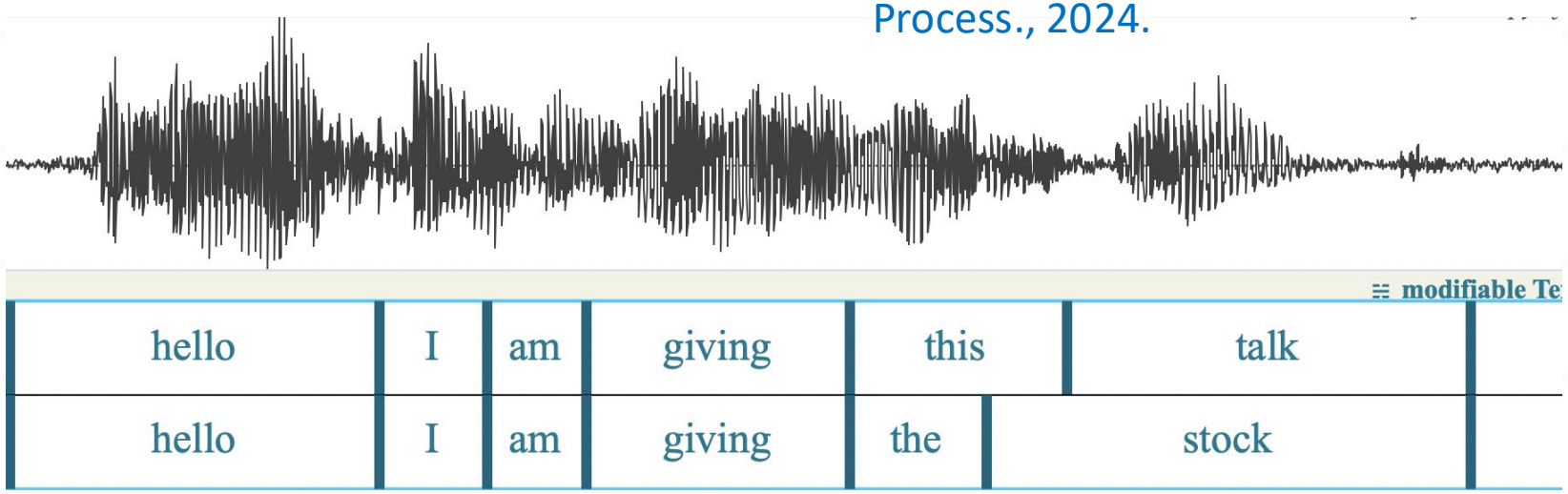


- Speech is a **time series**
- Word-level confidence estimation model [1]
  - `Hello, I am giving the stock`
  - `  1, 1, 1,    1,  0,   0`
  - Minimize binary cross entropy w.r.t. 0/1 target

1. Q. Li et al., “Confidence Estimation for … Speech Recognition,” ICASSP 2021

TeLeS

N. Ravi, Thishyan 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

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

TeLeS

- Treat time series as bag-of-words
- |{this}∩{the}| = 2
- |{this}∪{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

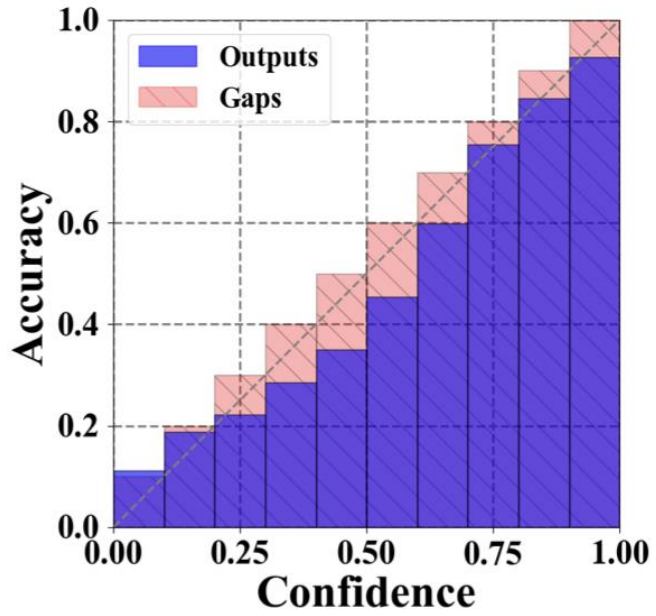


TABLE V: TeLeS-WLC and SOTA - Hindi PB Testset

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

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	**32.96/11.09**
1/7	31.71/10.59	33.09/11.00	36.93/12.42	**28.78/9.51**

Impact

- Active Learning
- 10x faster annotation workflows


project5

Active Learning for different languages

Hindi

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	**32.96/11.09**
1/7	31.71/10.59	33.09/11.00	36.93/12.42	**28.78/9.51**

Tamil

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	**47.38/9.83**
1/7	45.62/9.28	45.11/9.13	45.49/9.19	**44.35/8.87**

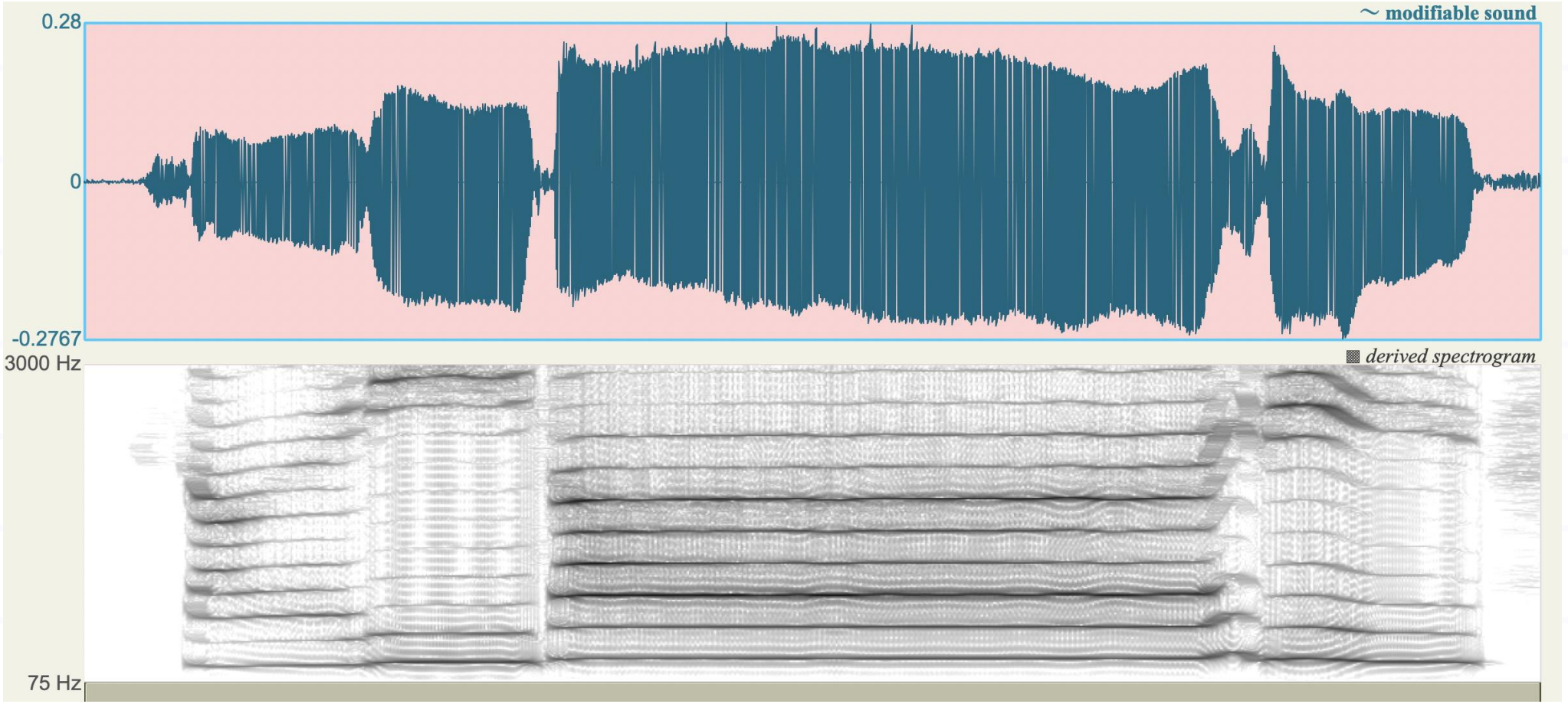
Kannada

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	**47.98/10.71**
1/7	43.75/9.57	60.87/13.67	74.47/17.69	**41.91/9.15**

Uncertainty Estimation for Music Analysis

Melody Estimation

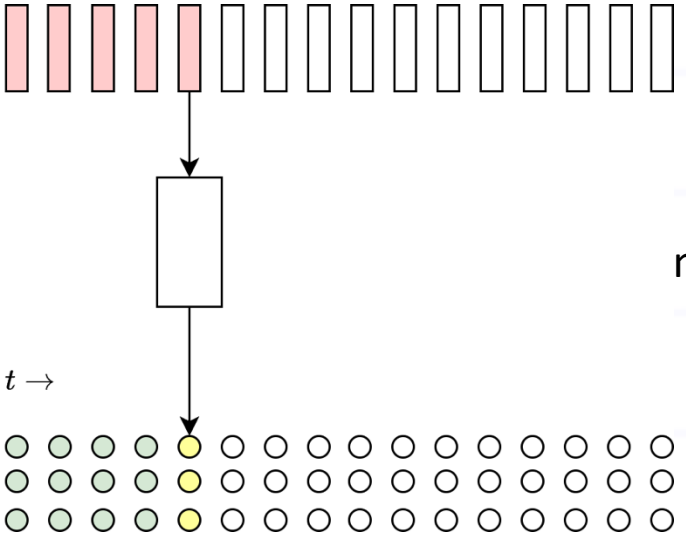
Singing


0.28
0
-0.2767
~ modifiable sound
3000 Hz
derived spectrogram
75 Hz

Polyphonic Music

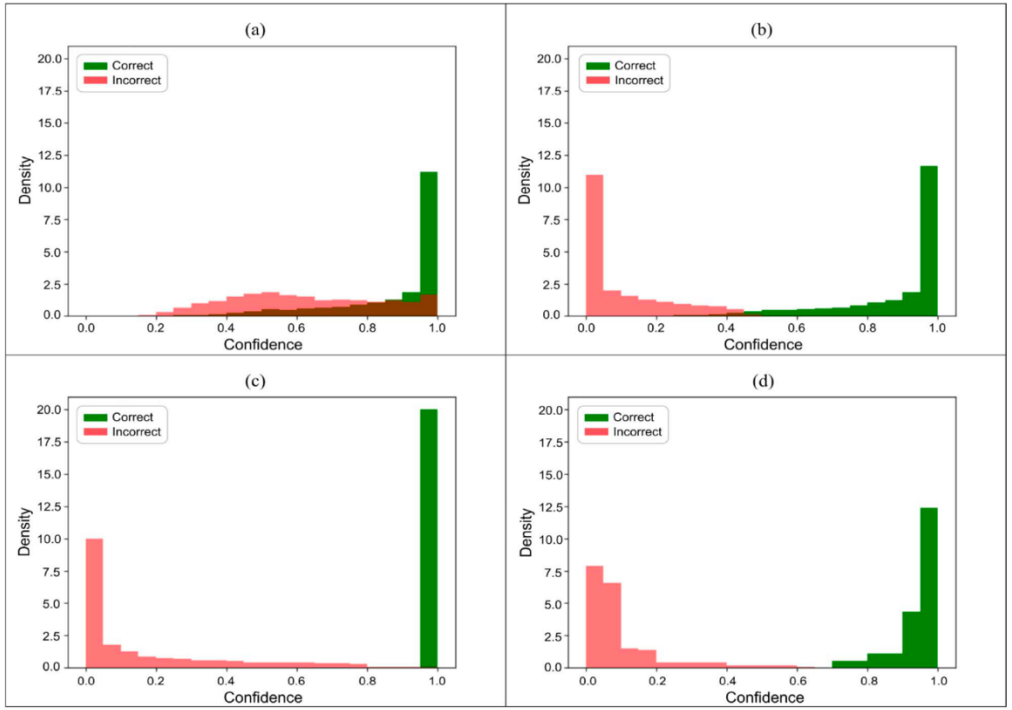
Melody Estimation

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.



max class prob.

Normalized true class prob.

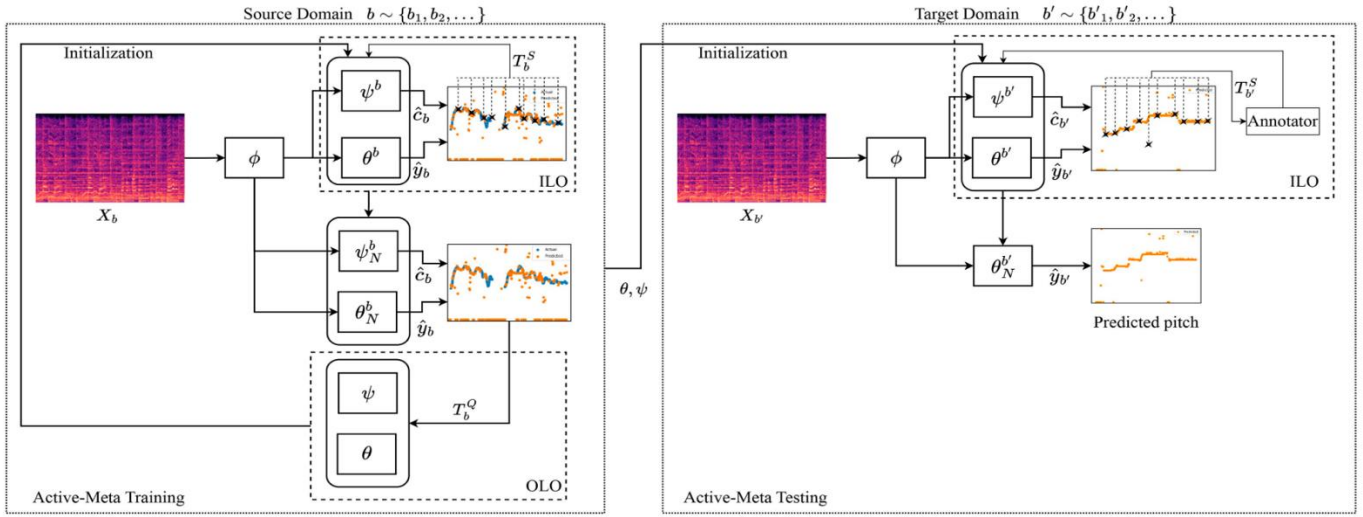


true class prob.

CEM trained with Normalized true class prob.

Active Learning with CEM

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



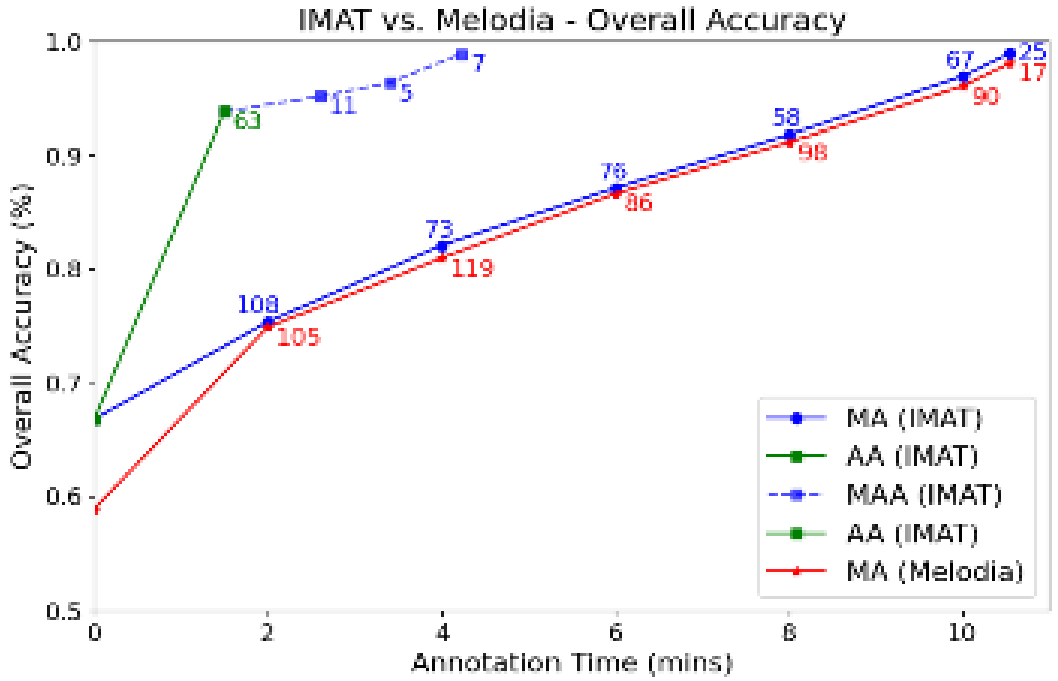
Dealing with heavy class imbalance

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

Comparison with other adaptation methods

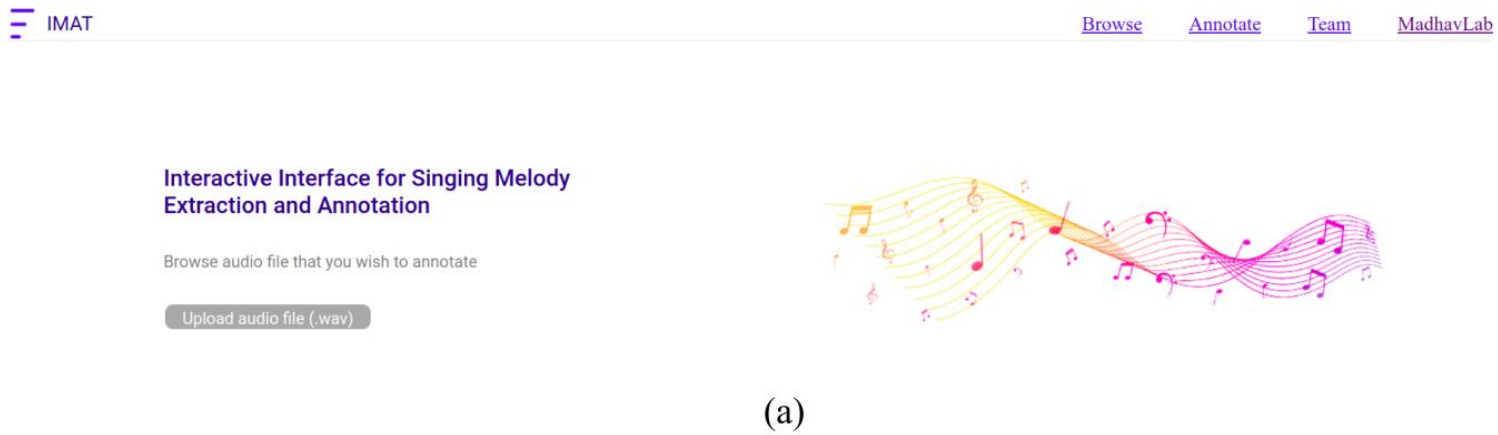
Experiments				ADC2004			MIREX05			HAR		
Method	**MW**	**AA**	**RA**	**RPA**	**RCA**	**OA**	**RPA**	**RCA**	**OA**	**RPA**	**RCA**	**OA**
FT-RA	-	✗	✓	80.34	81.45	80.98	81.16	81.98	82.10	76.45	77.10	76.88
FT-AA	-	✓	✗	81.55	81.59	81.66	81.80	81.78	81.85	76.95	77.34	76.17
MAML-RA	✗	✗	✓	81.10	82.56	81.41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

Experiments

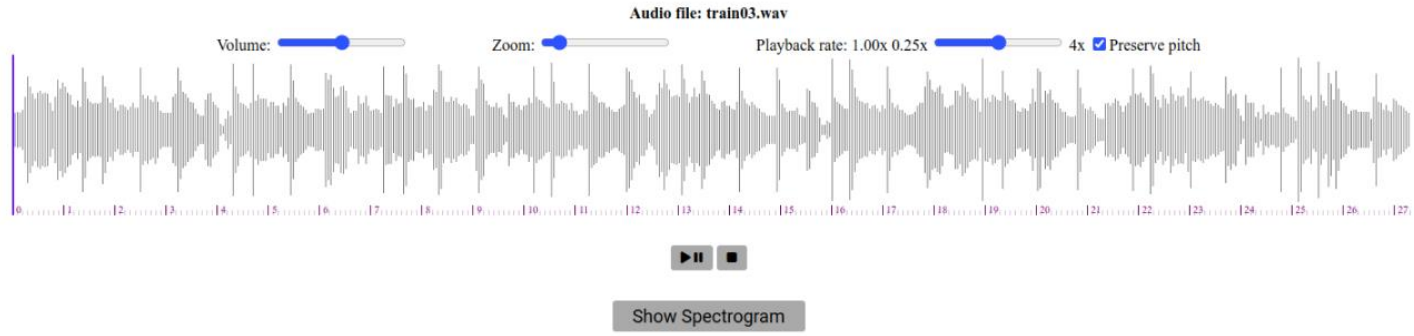


Code Repository and Tutorial

- 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://youtu.be/1SSQOE8ClHE?si=ktSEyapmJvTaWsZ5



(a)



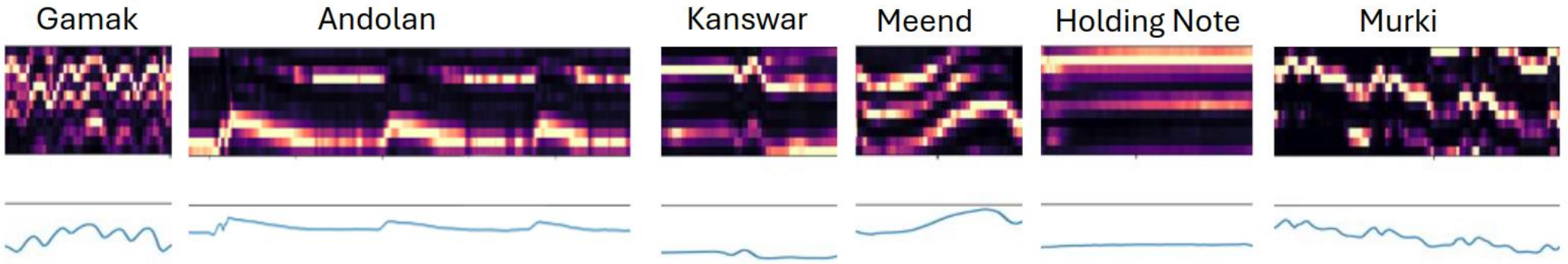
(b)

Uncertainty Estimation for Music Analysis

Ornaments and Raga

Ornamentation Detection

submitted to TASLP



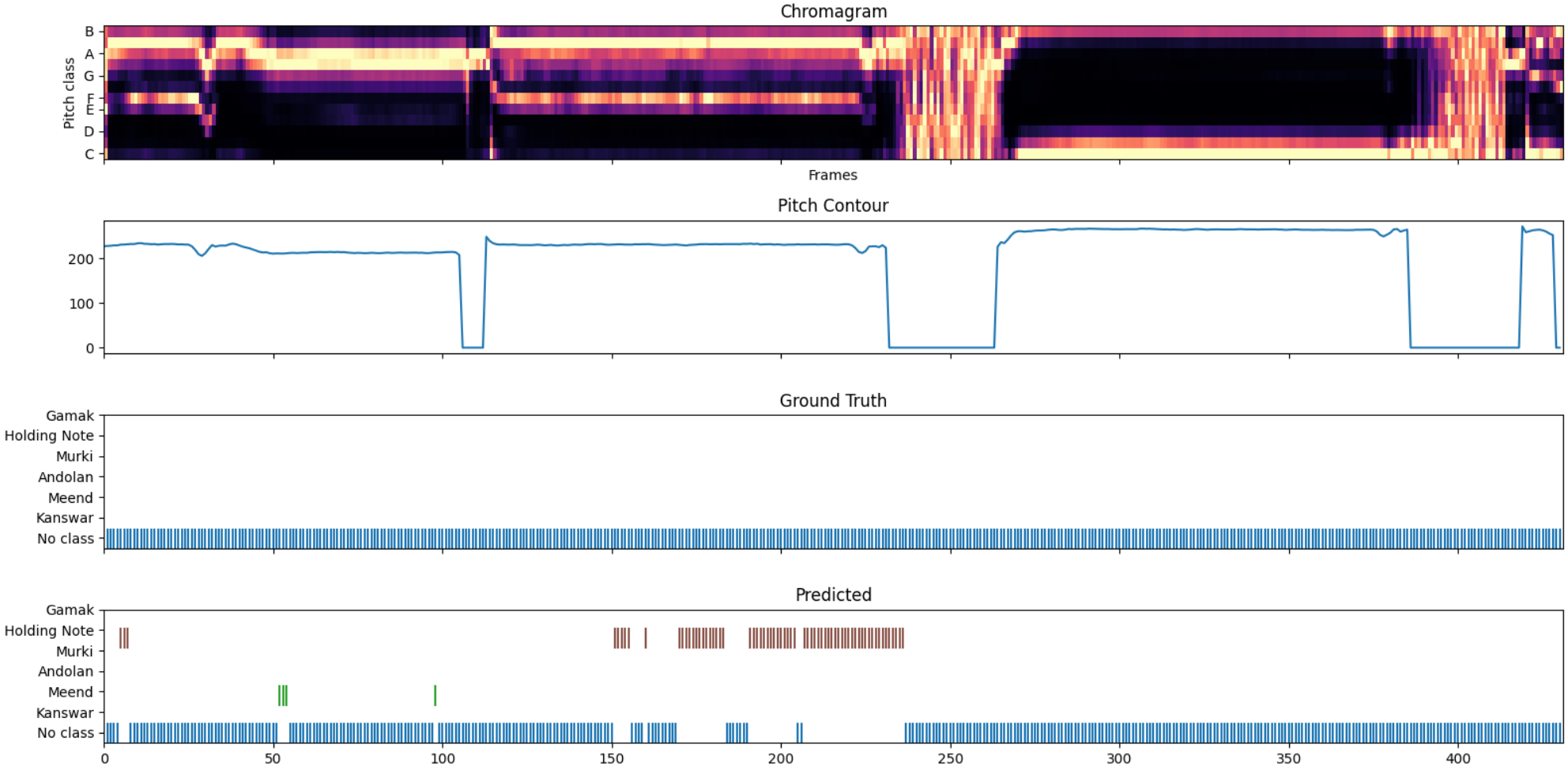
Recognizing Ornamentations in Singing Voices in Indian Art Music

Sumit Kumar and Vipul Arora
Department of Electrical Engineering, Indian Institute of Technology, Kanpur, India[a]

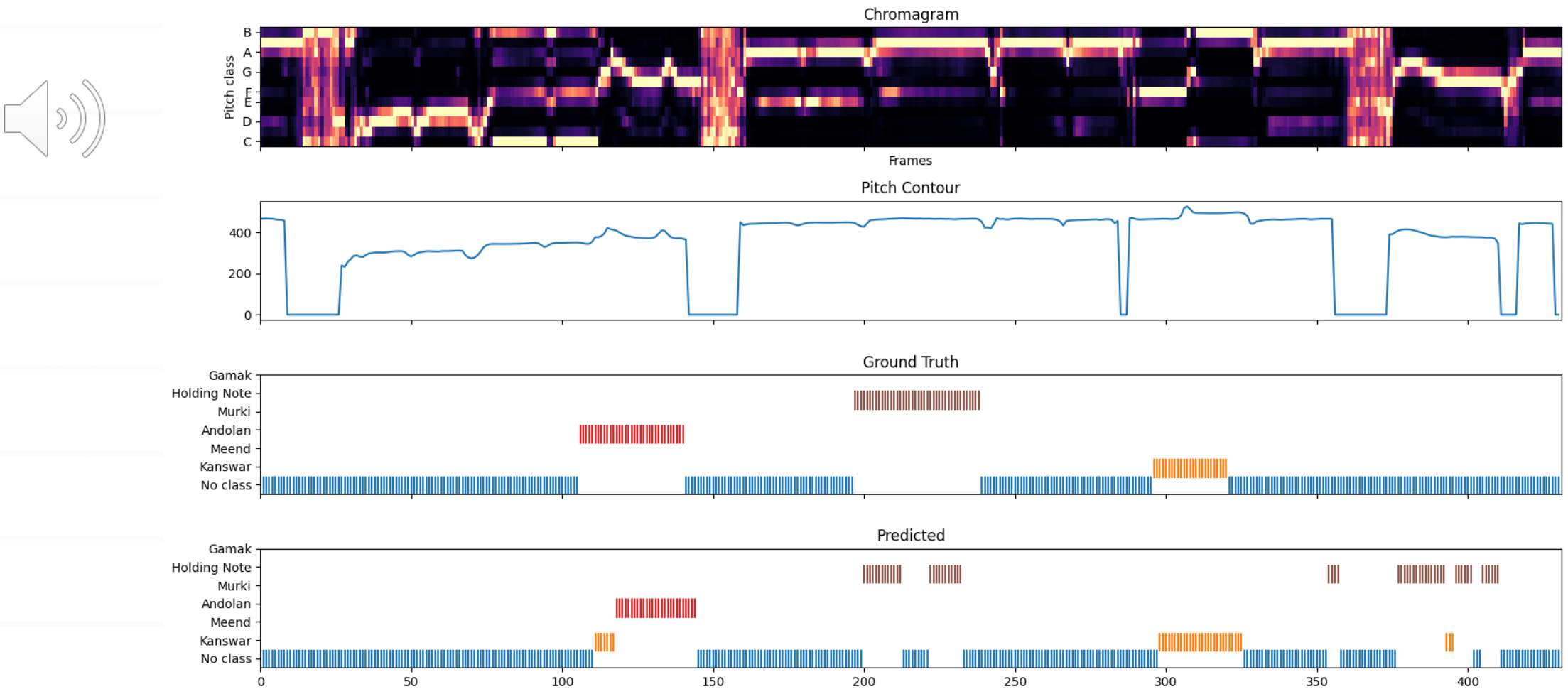
(Dated: 15 November 2024)

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

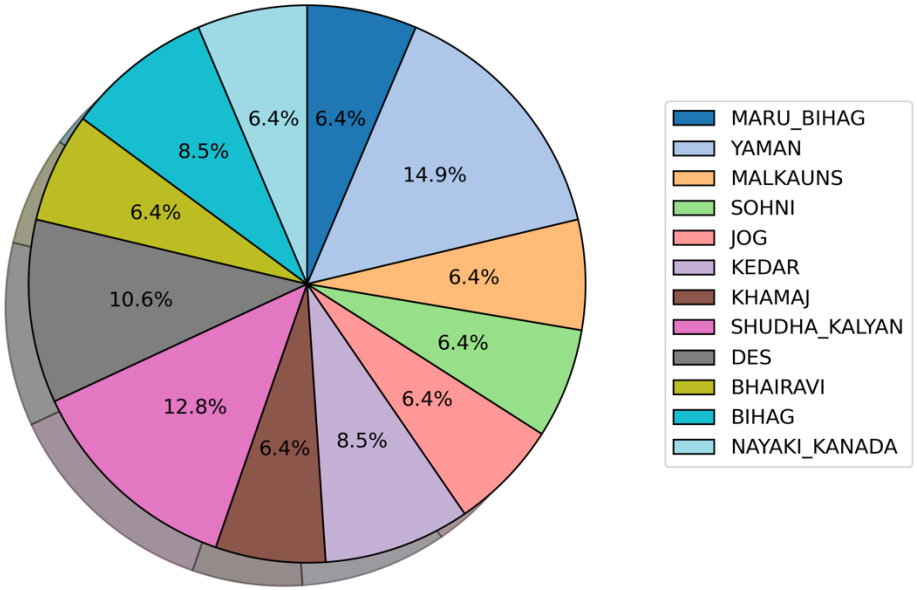


Andolan or Kanswar



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%

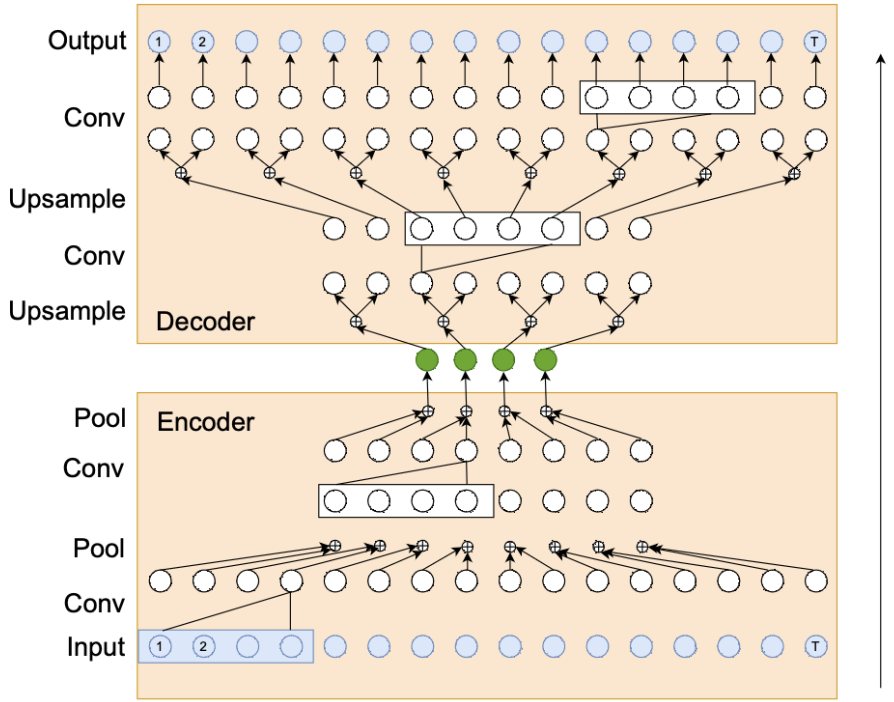
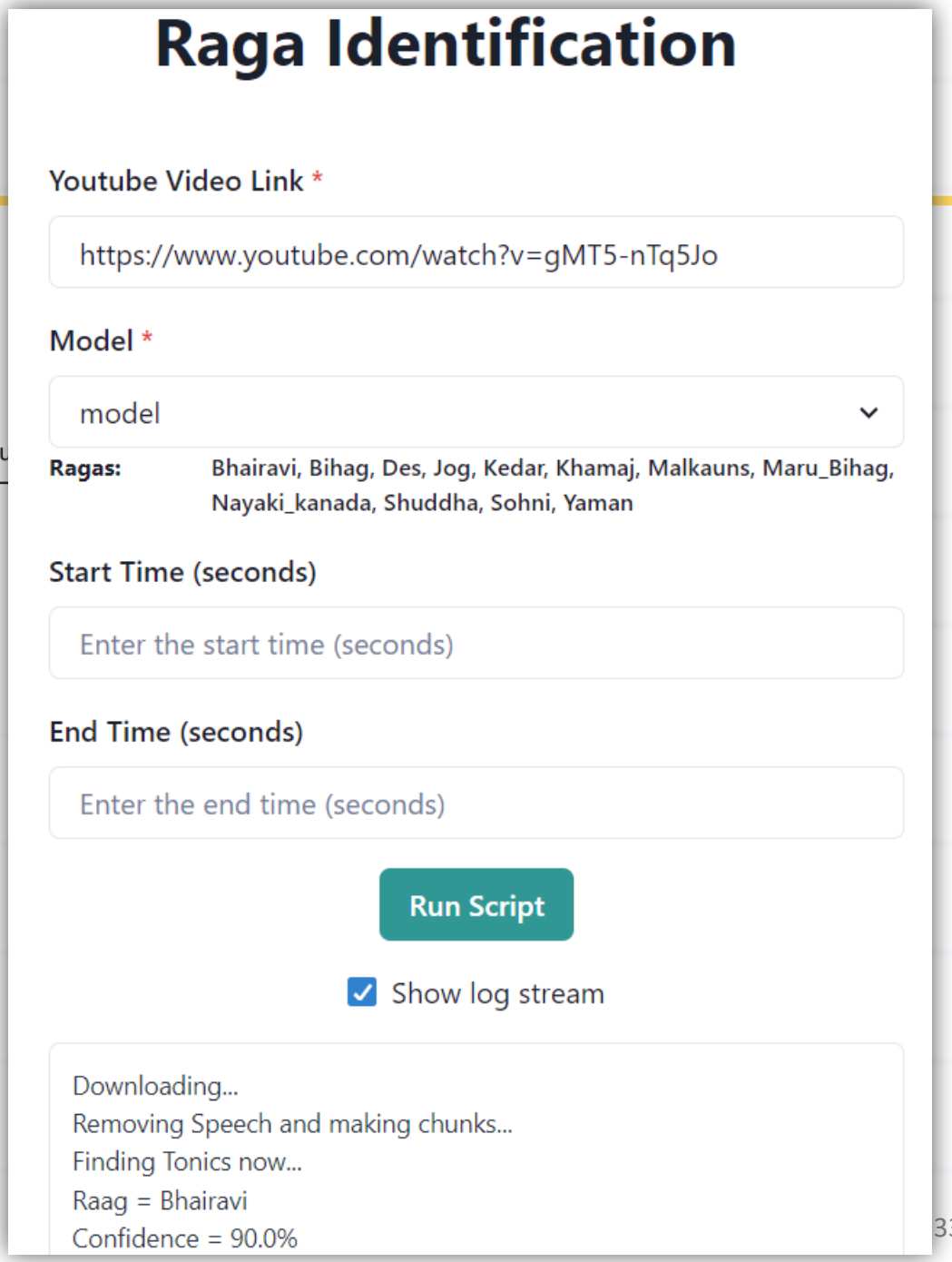


FIG. 6: Encoder-Decoder Temporal Convolutional Network

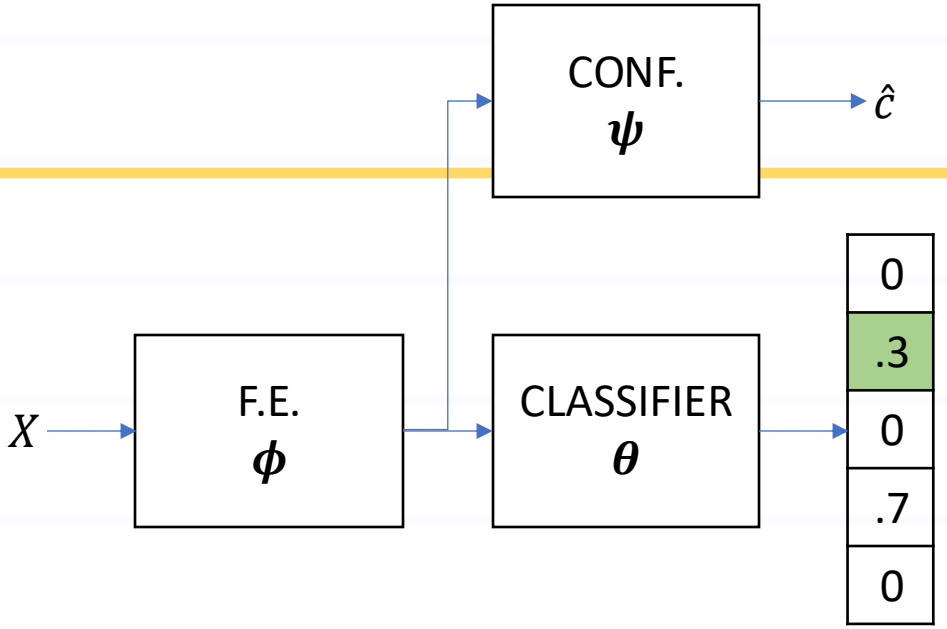
Raga Classification

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

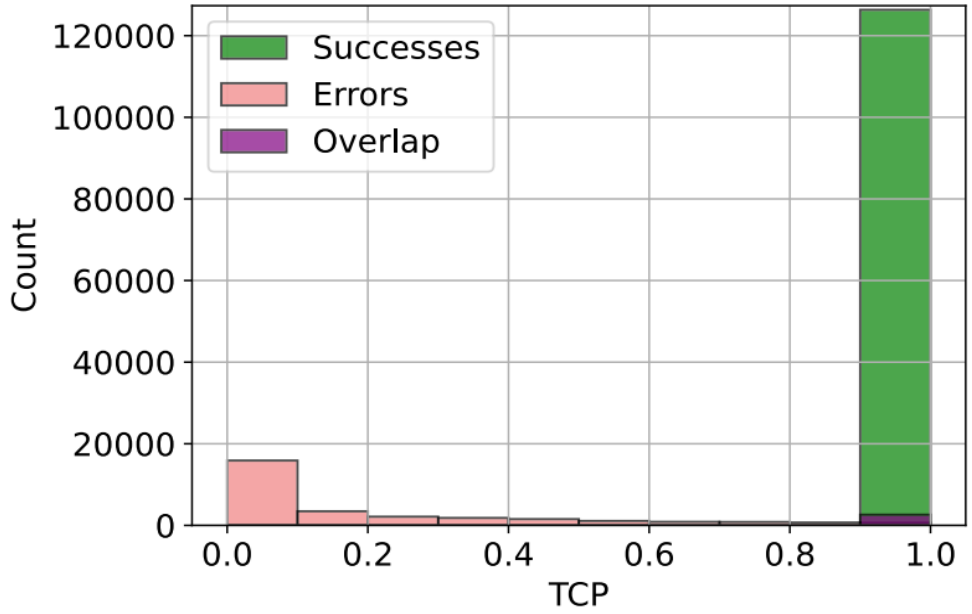


Limitations of TCP

- Ambiguity in confidence targets
- Imbalance in target values



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

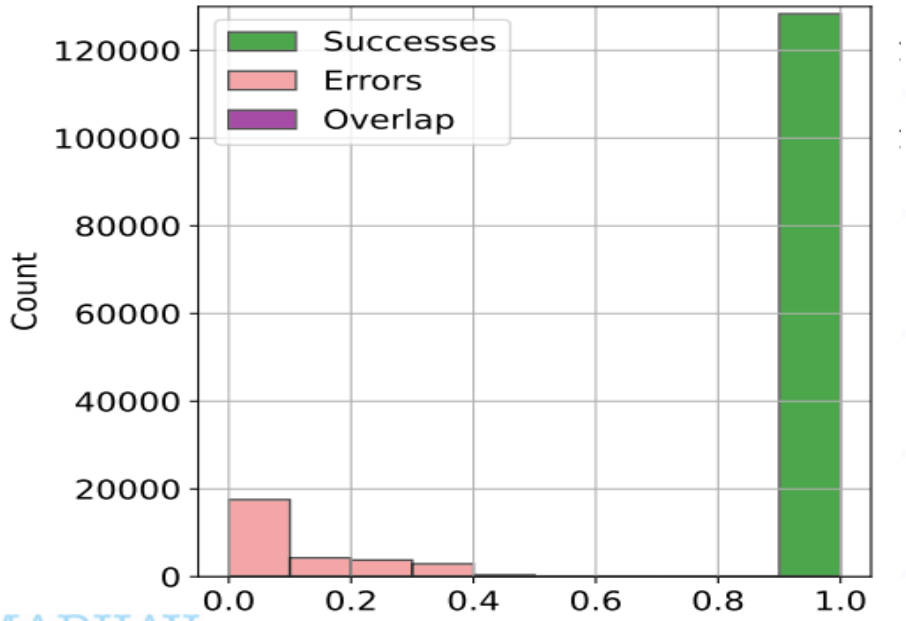


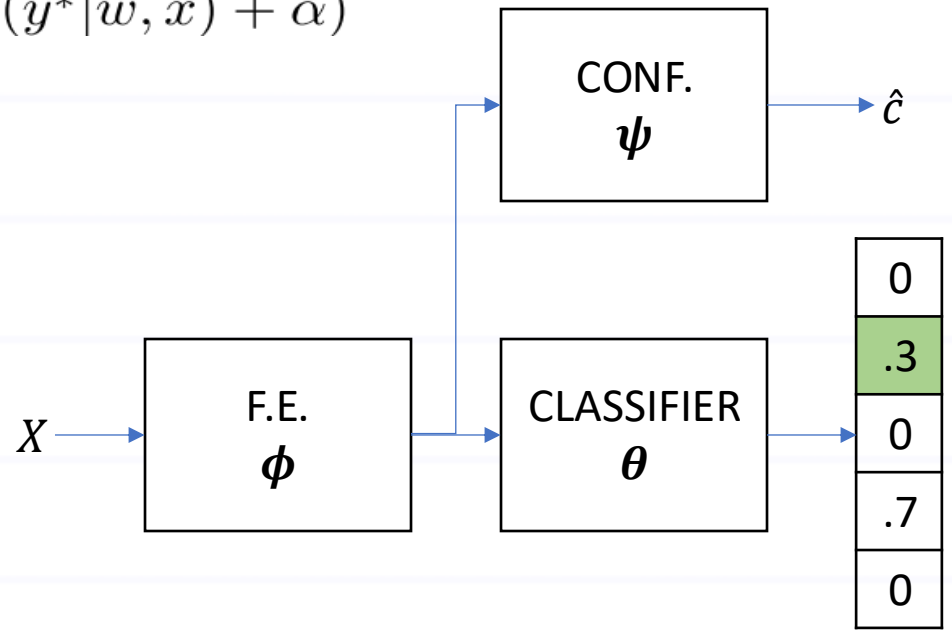
Errors with high confidence

Confidence Estimation Model

- Modified True Class Probability

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





Results

Sumit, Parampreet and Vipul,
ICASSP 2025 workshop

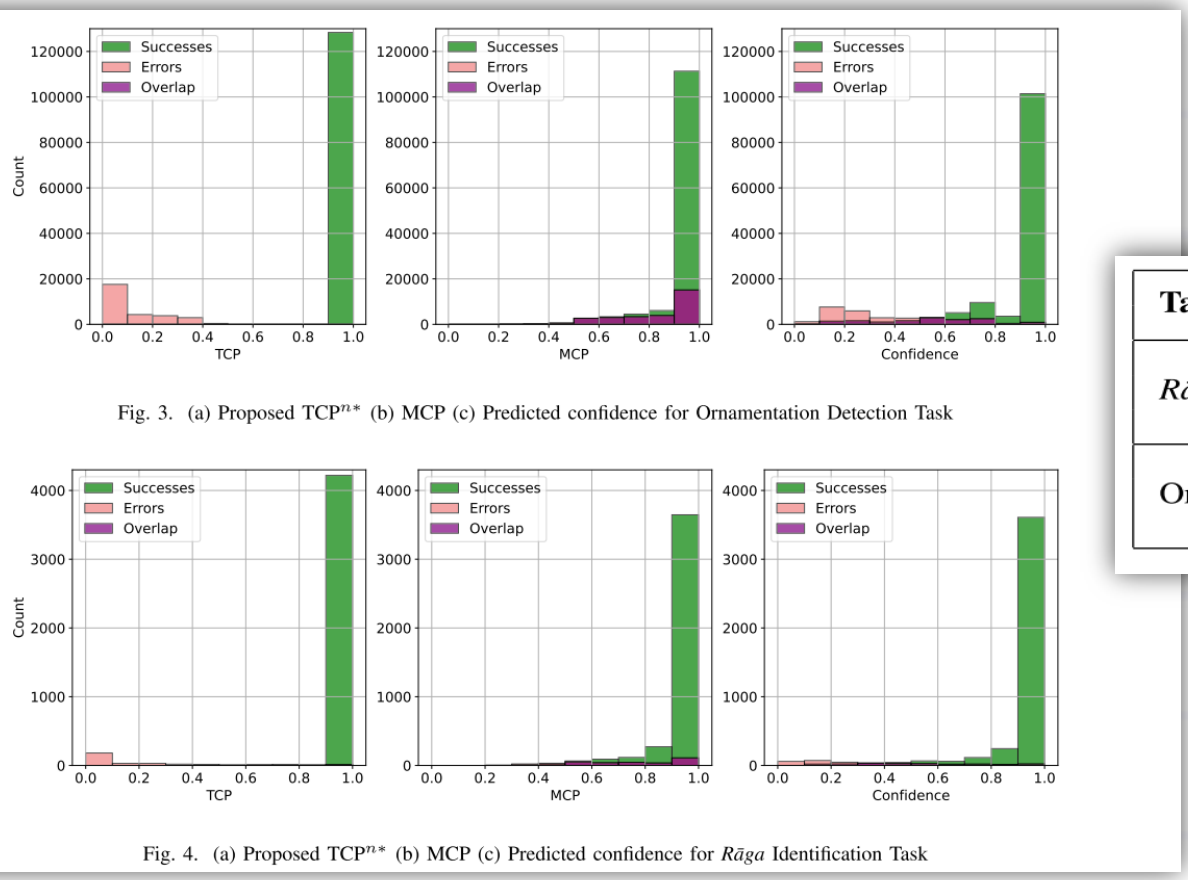


Fig. 3. (a) Proposed TCP^{n*} (b) MCP (c) Predicted confidence for Ornamentation Detection Task

Fig. 4. (a) Proposed TCP^{n*} (b) MCP (c) Predicted confidence for *Rāga* Identification Task

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

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)
- Kumar et al., Confidence-Enhanced Models for Indian Art Music Analysis, WIMAGA workshop, ICASSP 2025