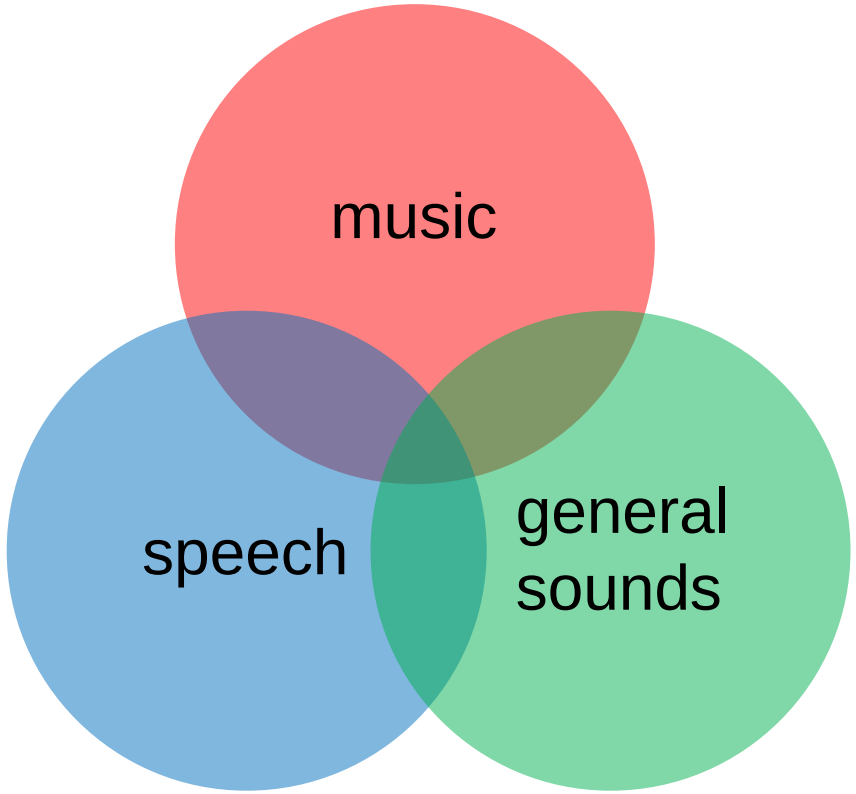
Audio Information Research (AIR) Lab Overview

Zhiyao Duan
Assistant Professor
Department of Electrical & Computer Engineering (Primary)
Department of Computer Science (Secondary)
University of Rochester
http://www.ece.rochester.edu/projects/air/index.html

Presentation at 2019 North East Music Information Special Interest Group (NEMISIG)
February 9, 2019

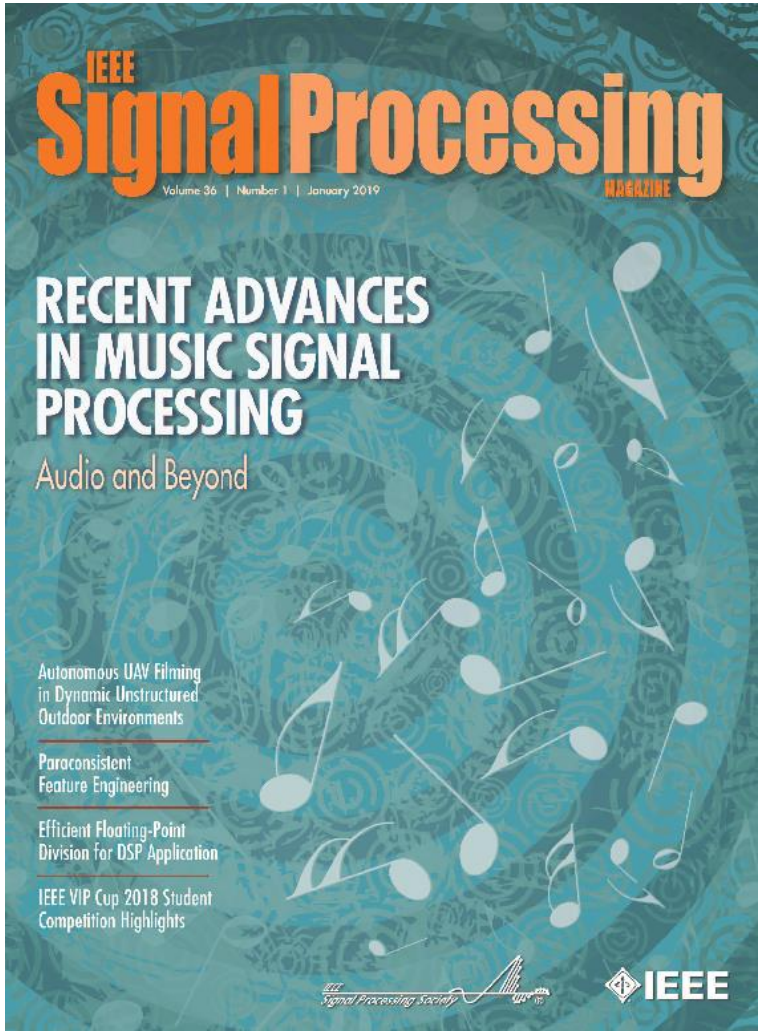
Theme: Machine Understanding of Sound Leveraging Other Modalities

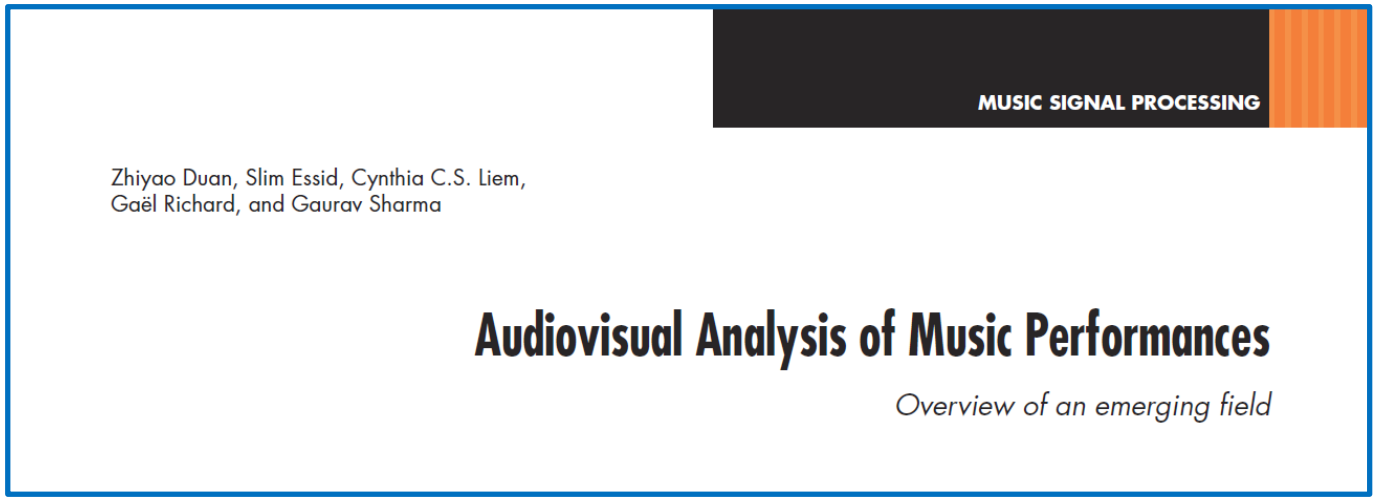


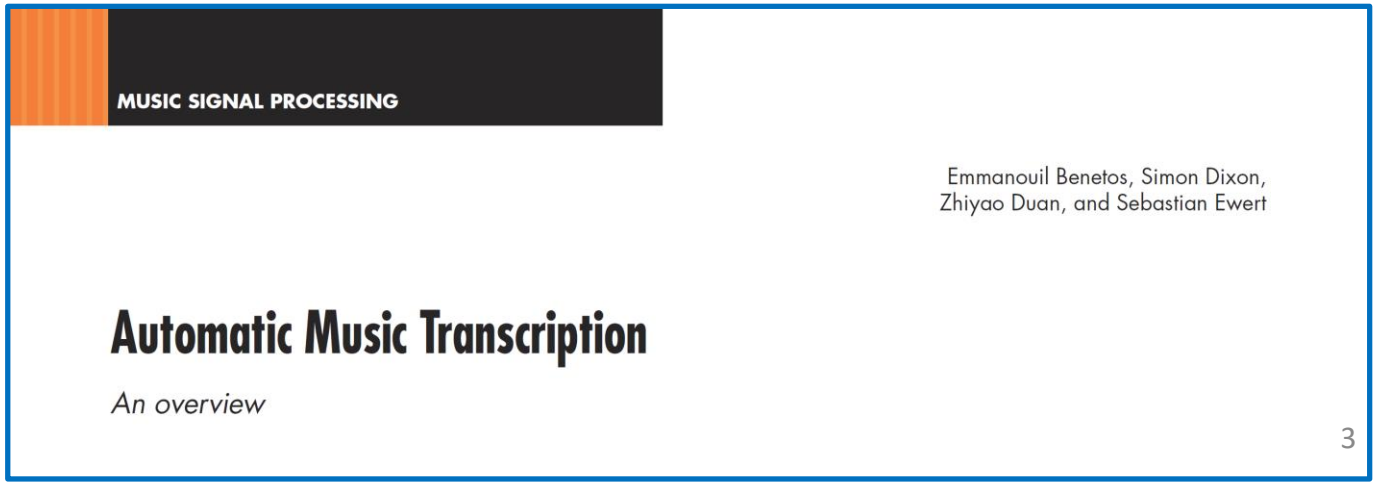
Active Research Grants in 2018

- NSF-IIS: Algorithms for Query by Example of Audio Databases
- NSF-BIGDATA: Audio-Visual Scene Understanding
- UR-ARVR: Real-Time Synthesis of a Virtual Talking Face from Acoustic Speech
- UR-ARVR: Spatial Audio VR Recordings of Music Concerts

Music Overview Articles

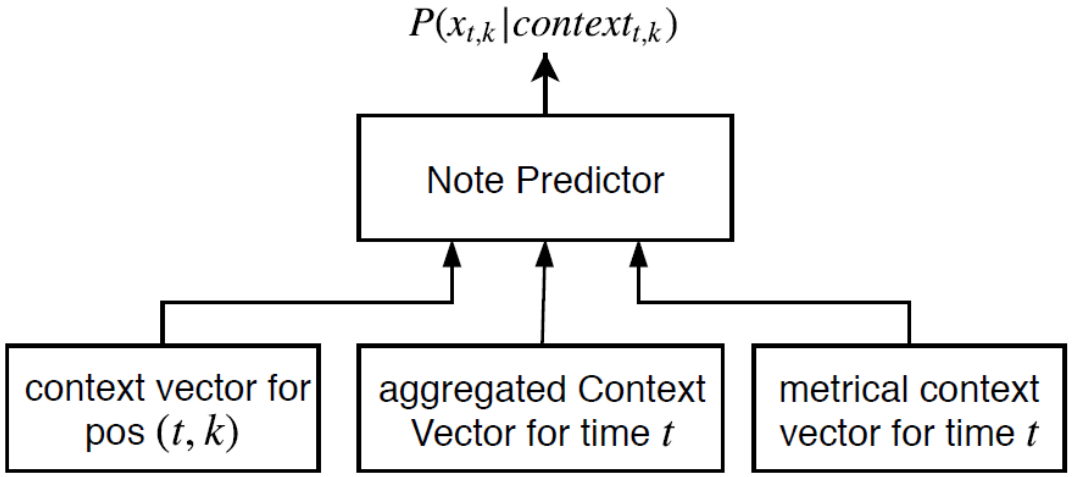


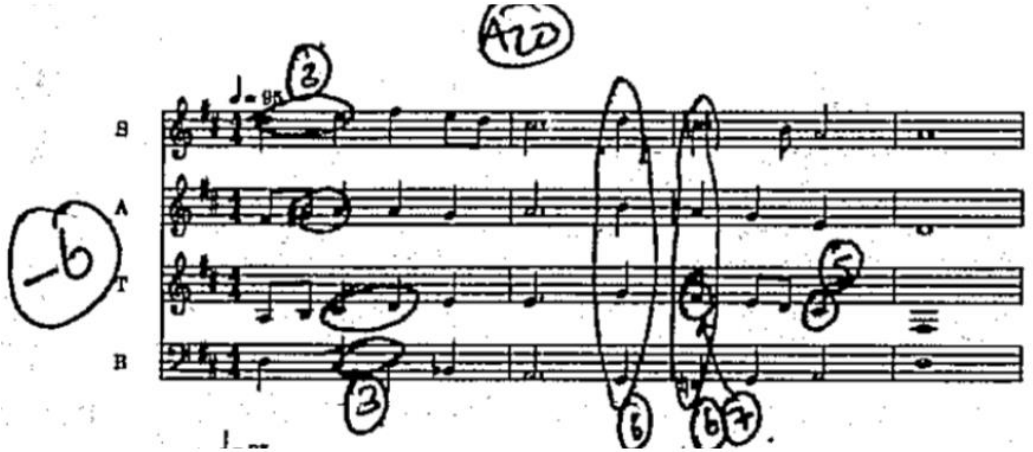




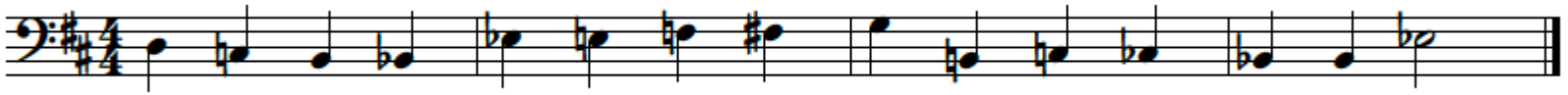
Music Generation
(Yan et al. ISMIR'18)

- A part-invariant model for music generation and harmonization
- Objective and subjective comparison with human harmonization

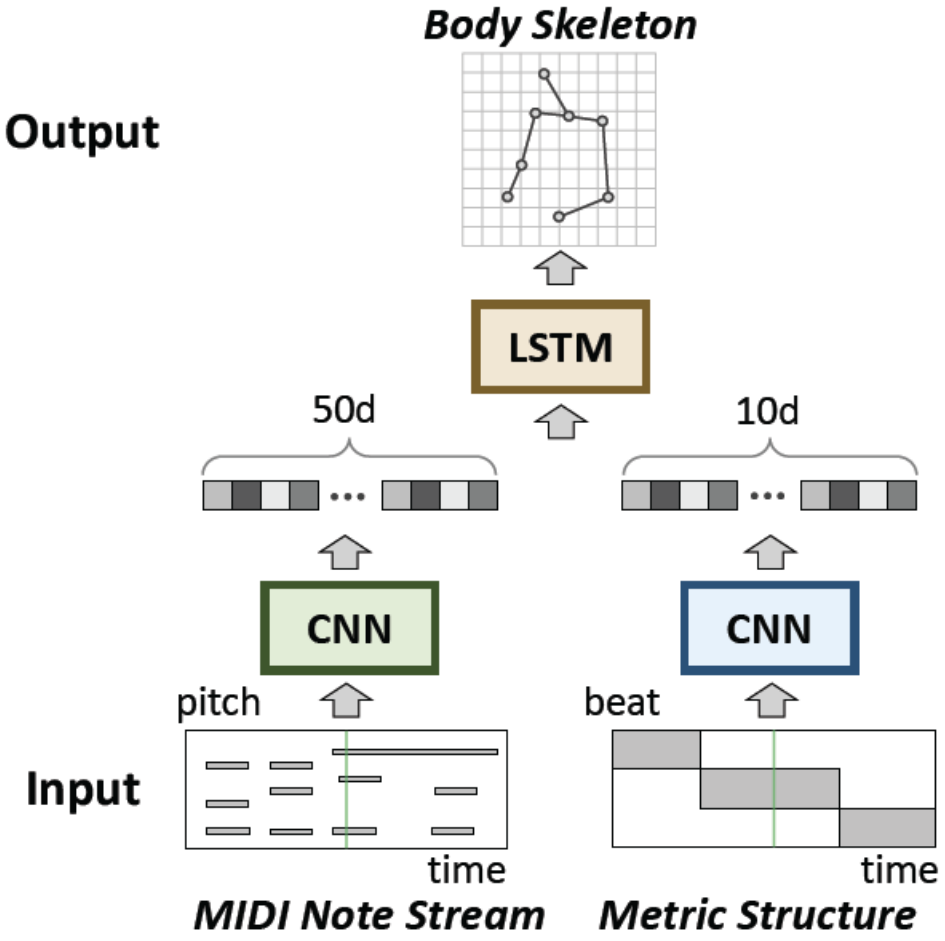




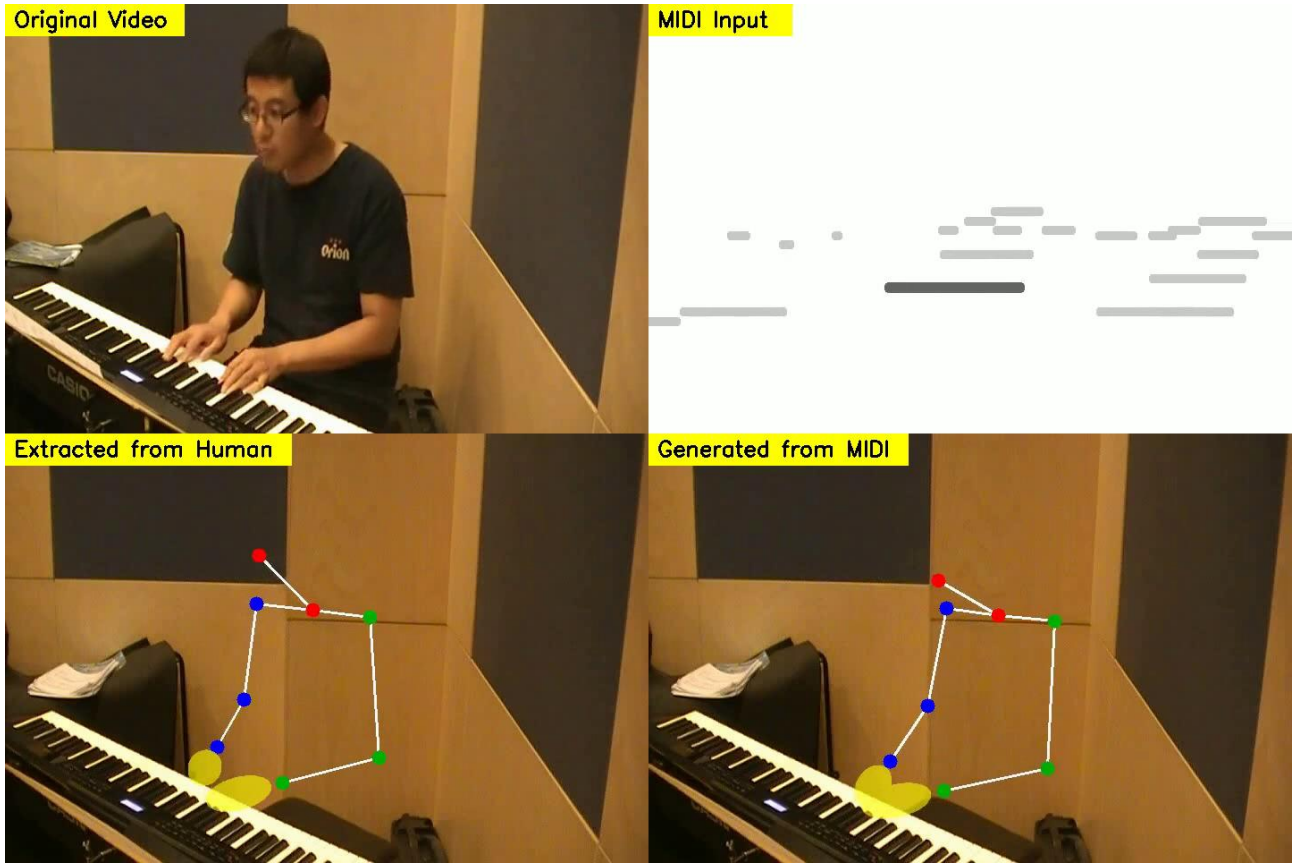
Baseline harmonization examples



Skeleton Plays Piano
(Li et al., ISMIR'18)



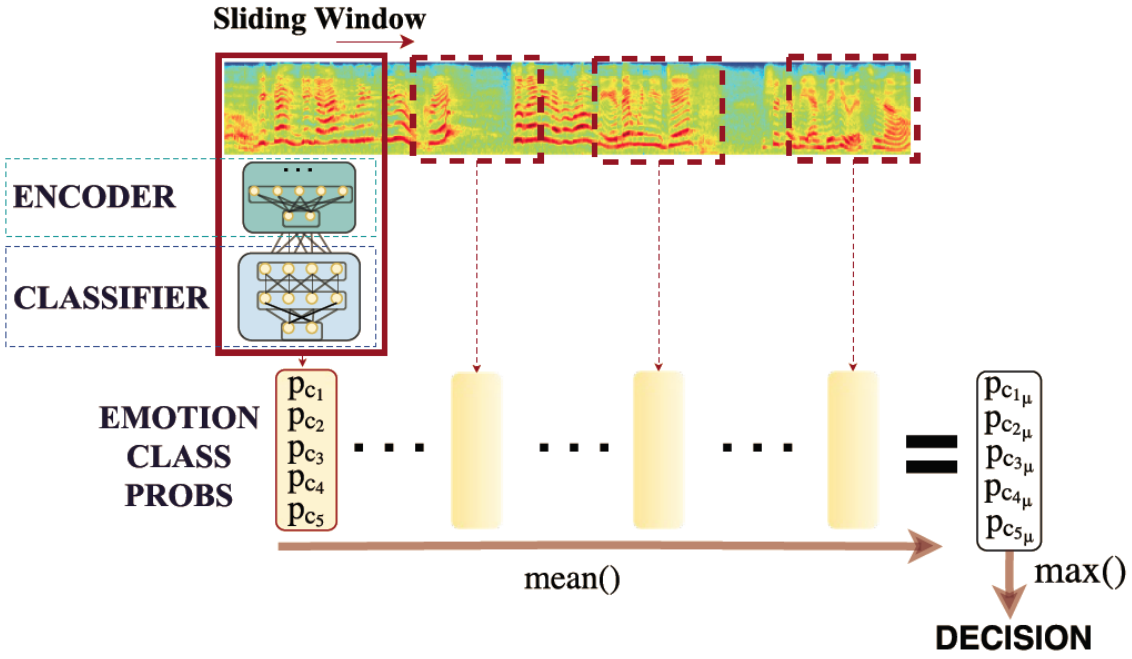
Example



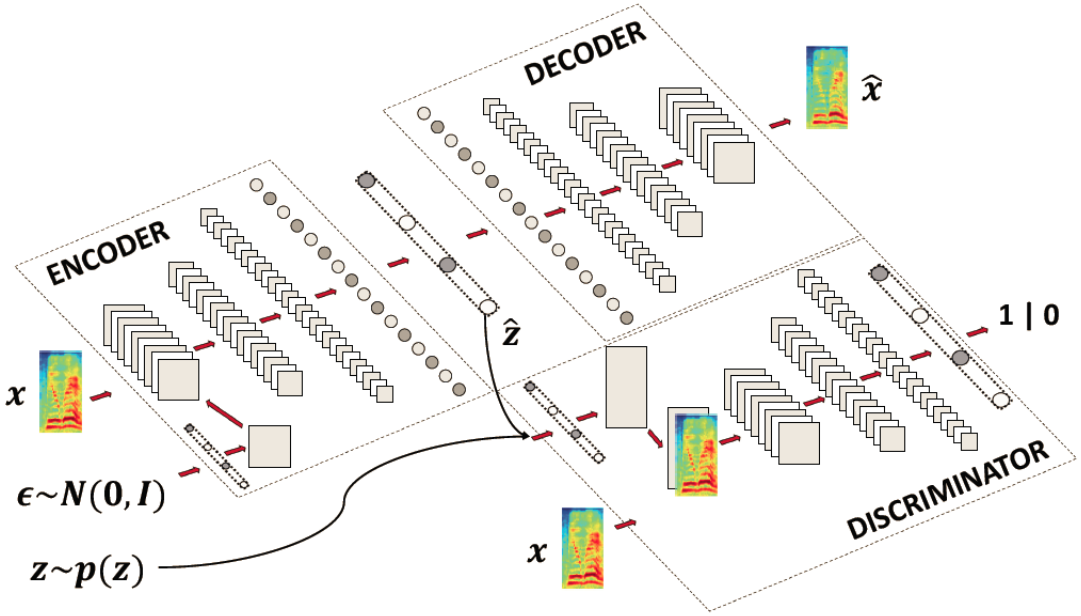
Speech Emotion Recognition
(Eskimez et al., ICASSP'18)

- Use unsupervised feature learning (e.g., autoencoders) on non-emotional speech to significantly improve emotion recognition accuracy

General Framework

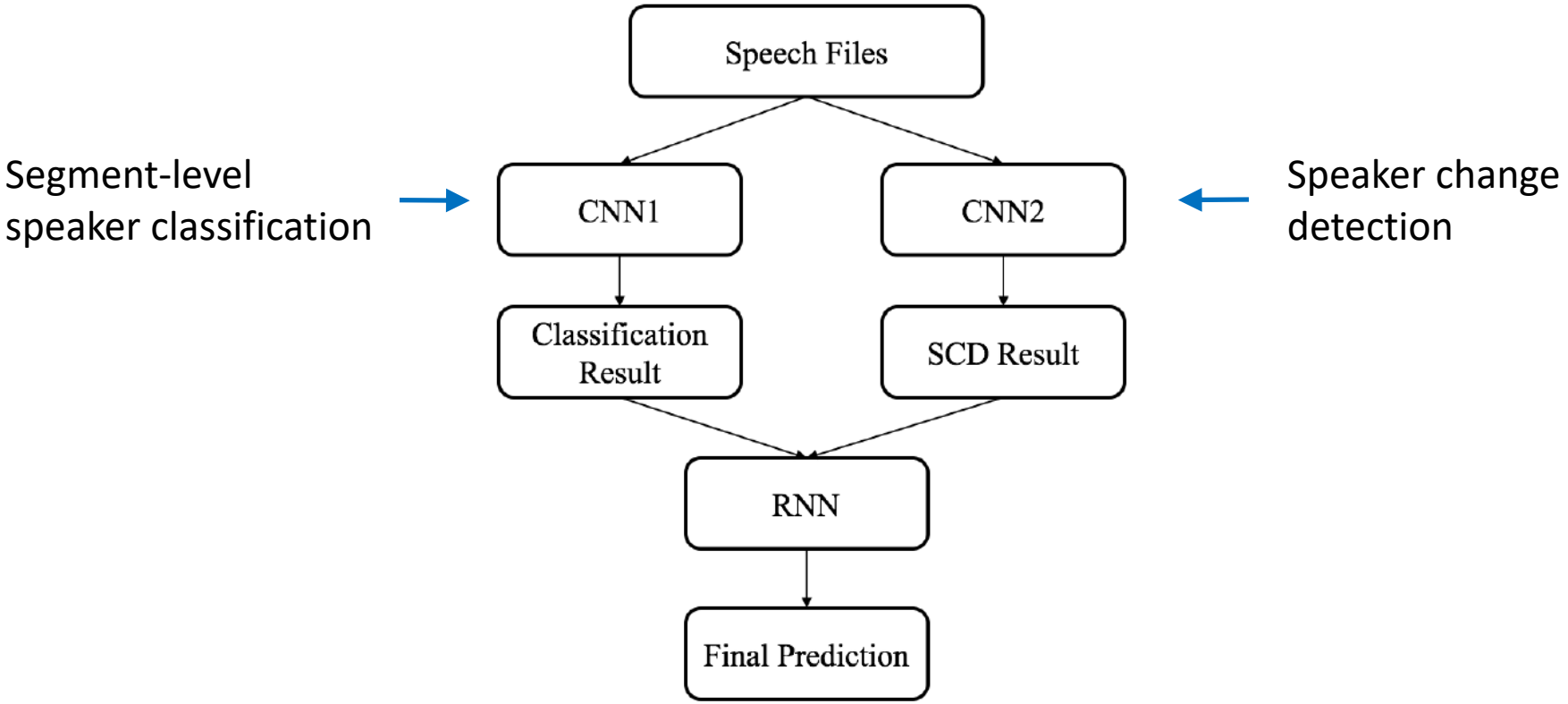


Adversarial Variational Bayes (AVB)

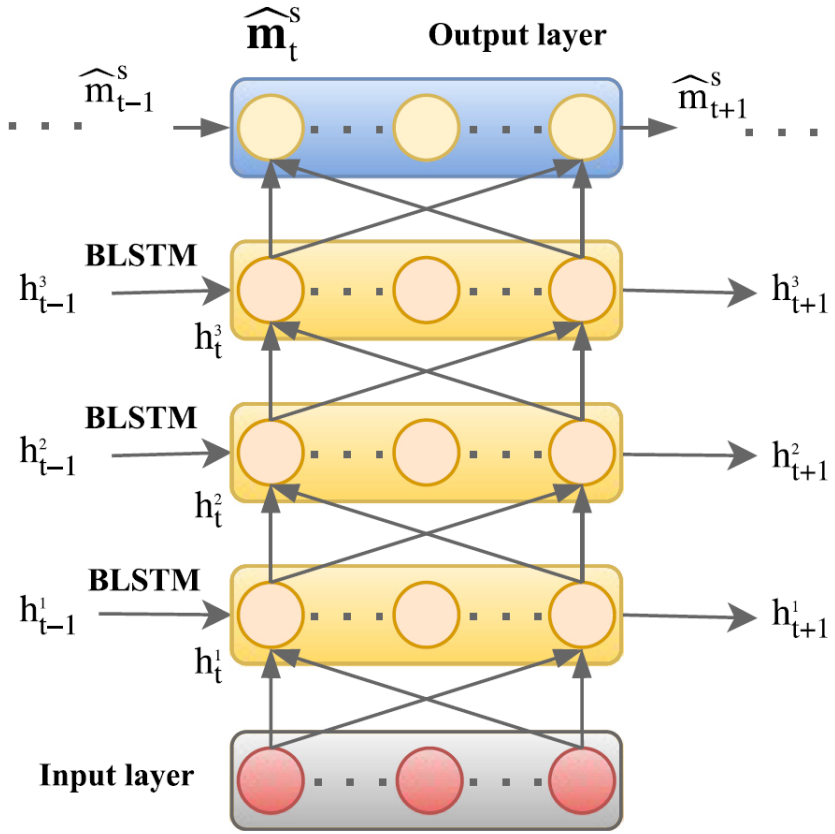


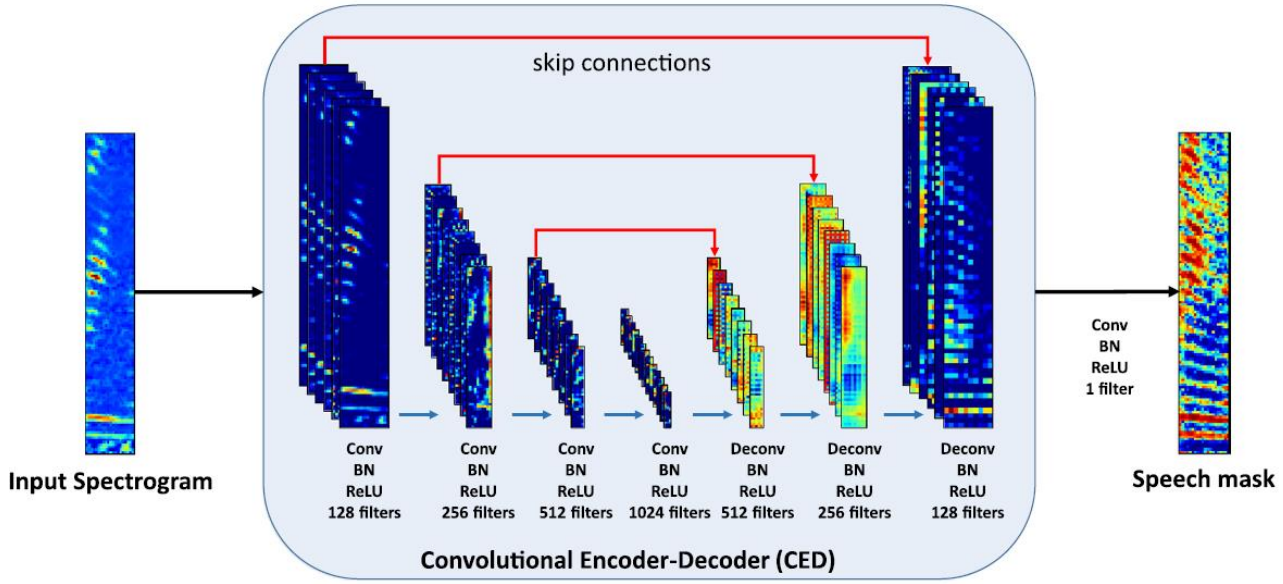
Joint Speaker Recognition and Diarization (Zhou et al., ICASSP'18)

- Recognition: identify a speaker against others, mainly using timbre information
- Diarization: cluster speech segments by identity, using timbre and temporal continuity



Speech Enhancement for Speaker Verification (Eskimez et al., SpeechComm.'18)





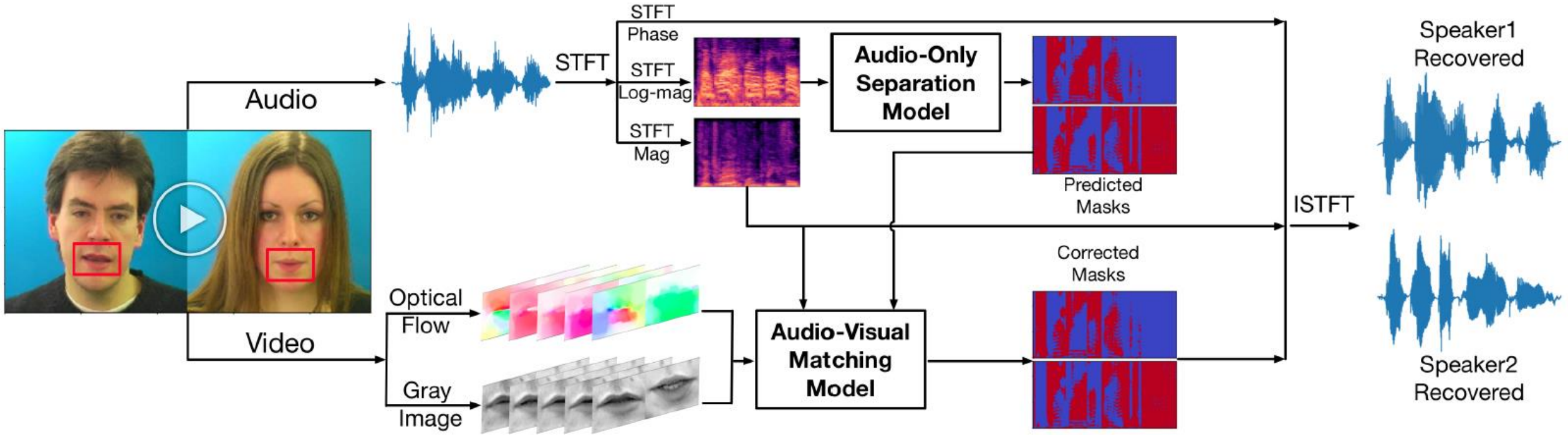
Noisy

Enhanced by CED

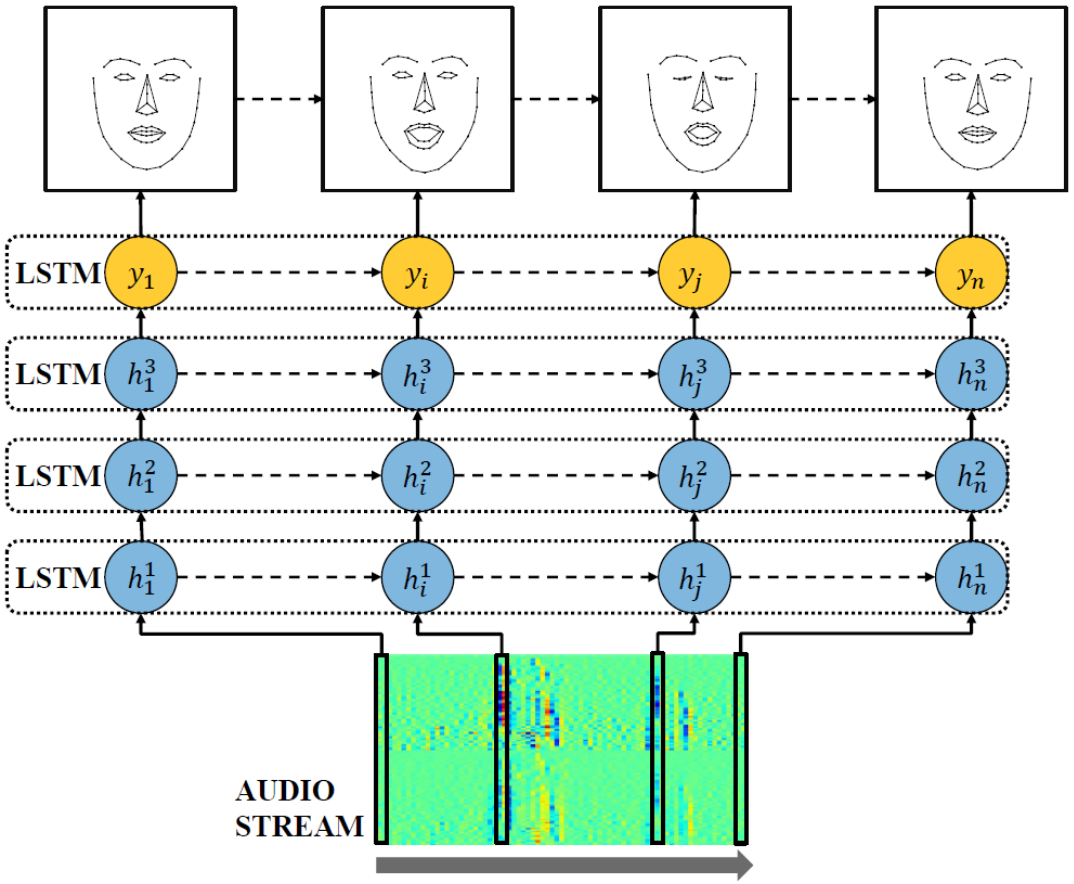
Enhanced by BLSTM

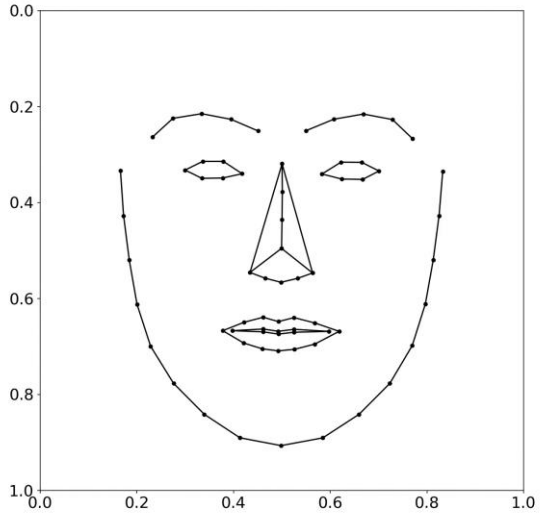
Audio-Visual Speech Separation
(Lu et al., SPL'18)

- Use audio-visual matching to solve source permutation problem

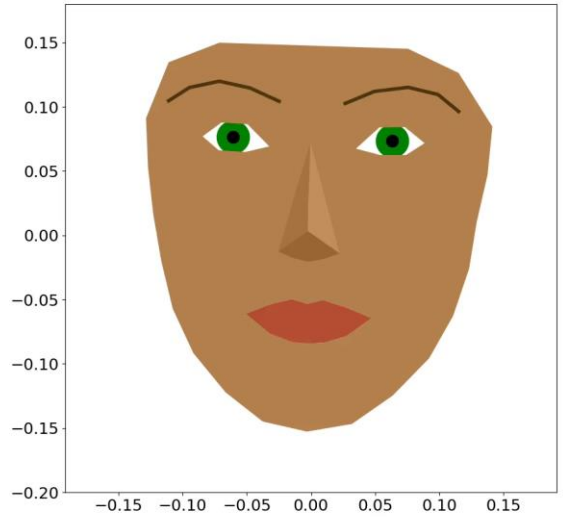


Talking Face Landmark Generation
(Eskimez et al., LVA/ICA'18)





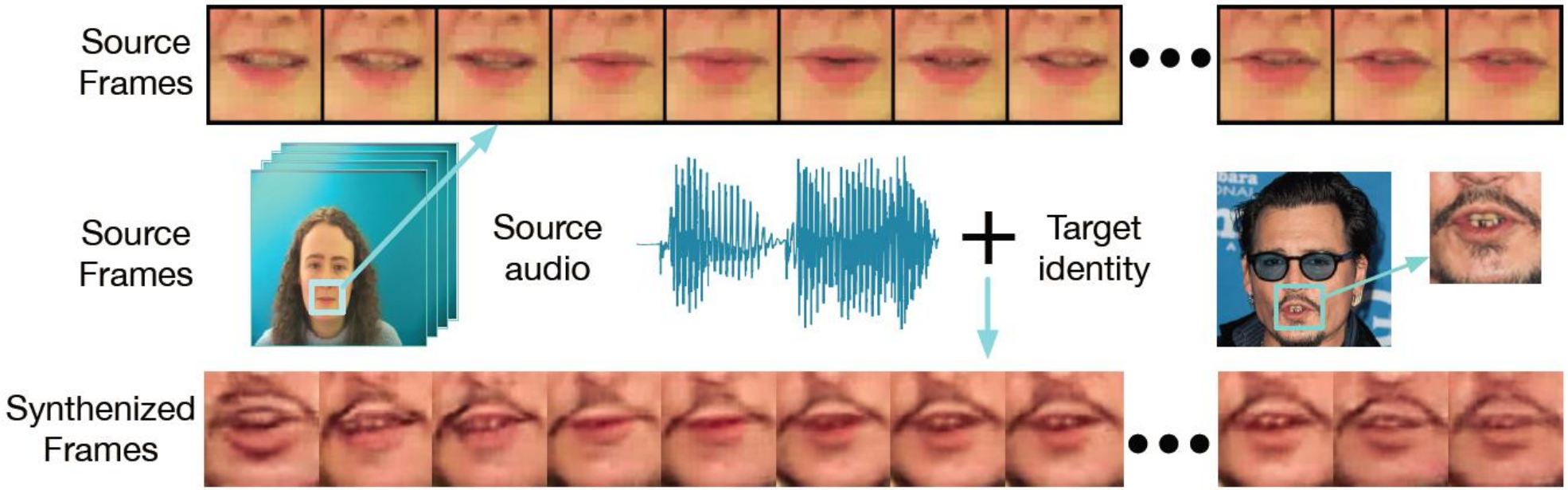
LSTM predicting 2D landmarks



1-D CNN predicting PCA coefficients of 3D landmarks

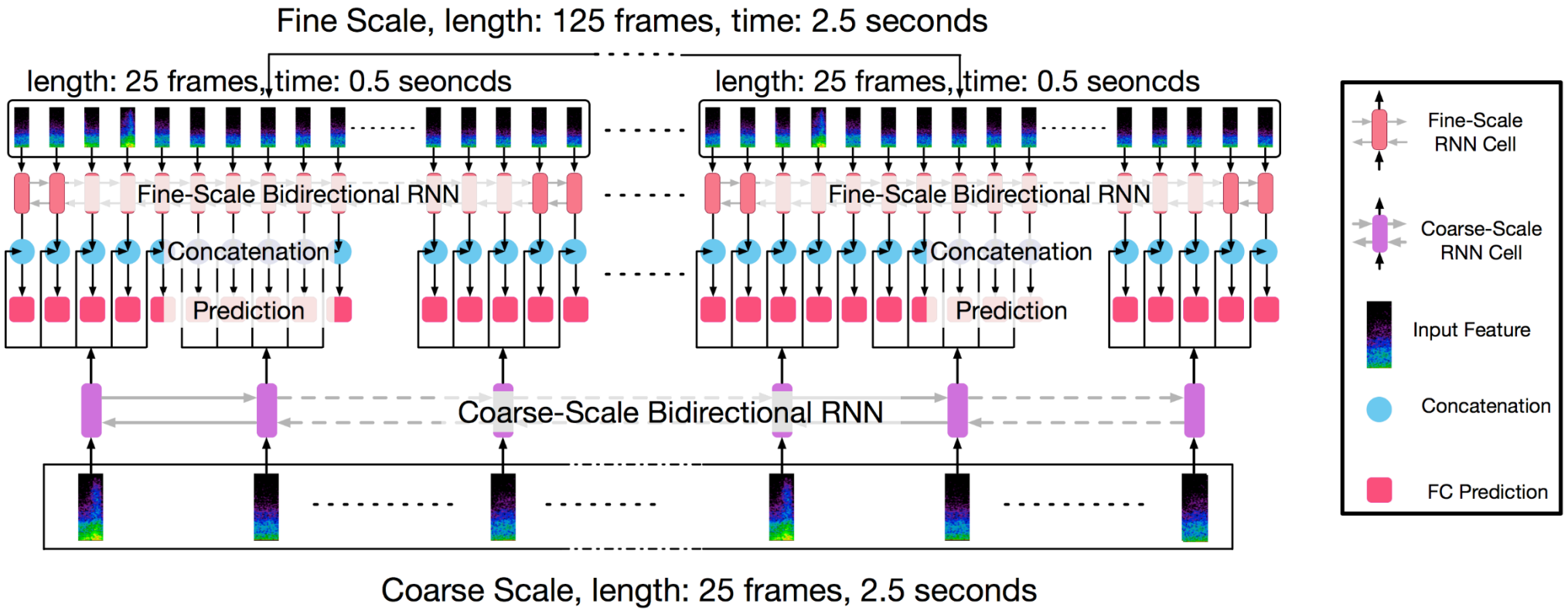
Lip Movement Generation
(Chen et al., ECCV'18)

- Input: speech audio from A + one lip image from B
- Output: lip movement of B speaking what A said
- Training objective: image reconstruction loss + perceptual loss + audio-visual correlation loss + adversarial loss

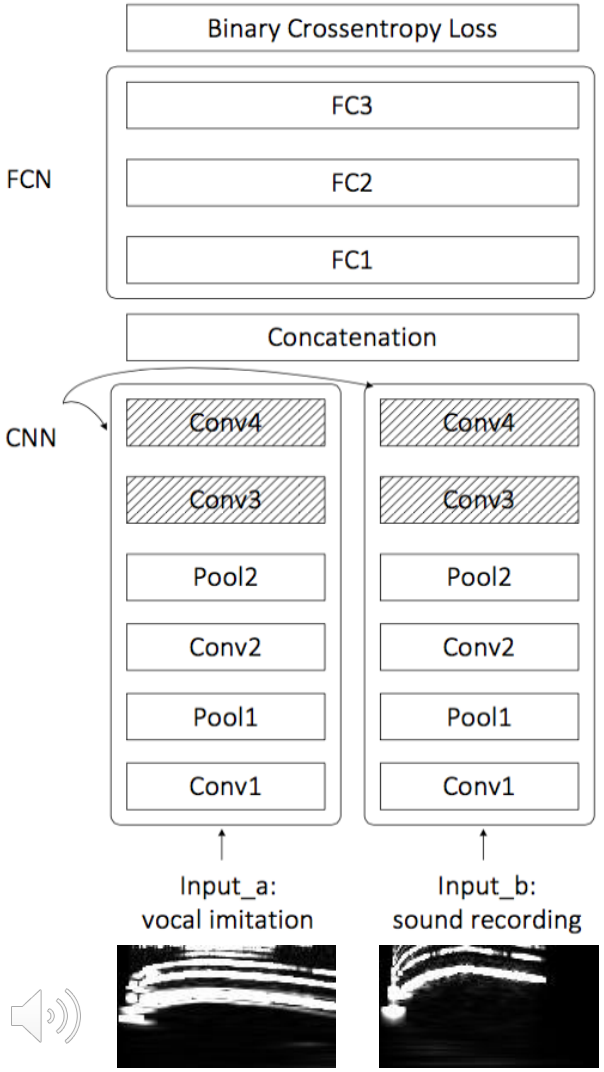


Multi-Scale RNN for Sound Event Detection (Lu et al., ICASSP'18)

- Captures long-term and short-term temporal evolution of sound events
- Separate training → collective fine tuning



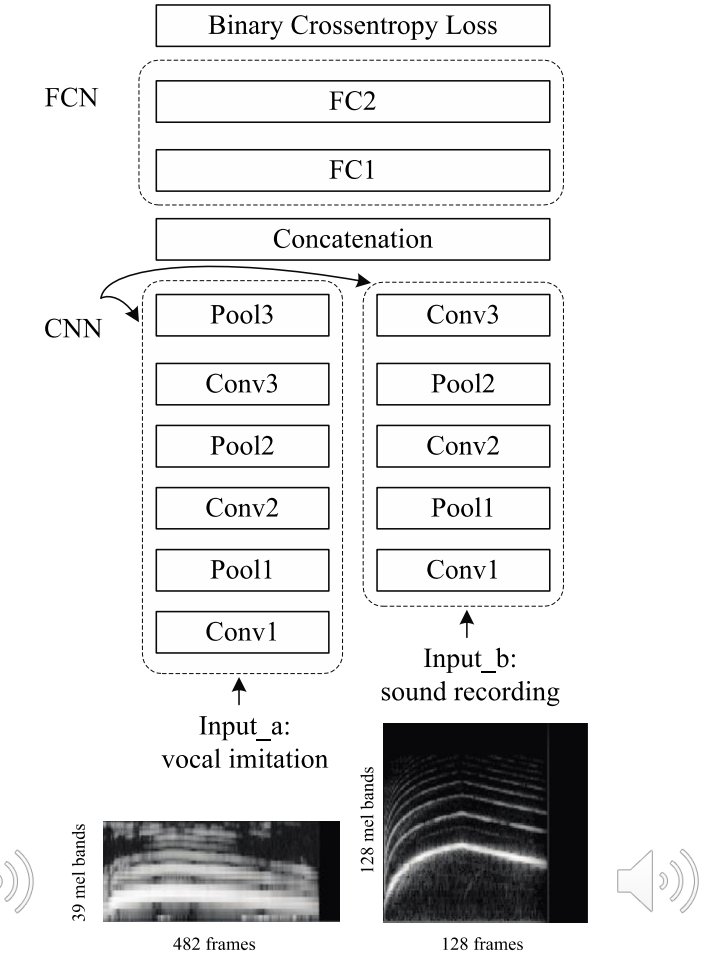
Sound Search by Vocal Imitation (Zhang et al., TASLP'19)



SYMM-IMINET: same architecture and partially shared weights for imitation and recording towers

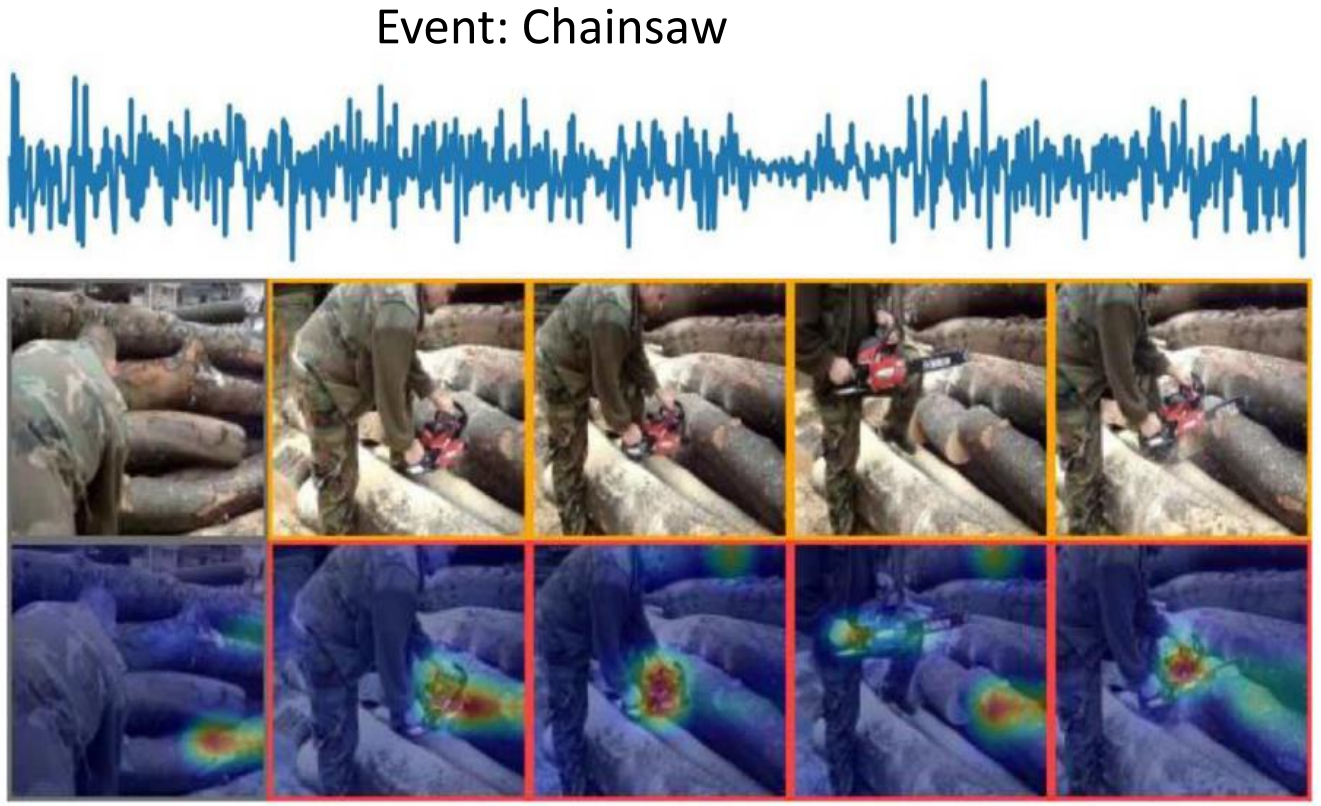
TL-IMINET: different architectures for imitation and recording towers

- Imitation tower: pre-trained for language recognition
- Recording tower: pre-trained for environmental sound classification



Audio-Visual Event Localization (Tian et al., ECCV'18)

- Audio-visual event: an event that is both audible and visible
- Training with weakly labeled videos
- Cross-modal event localization




Thank You!