CVPR 2021 Tutorial on
Audio-Visual Scene Understanding

Audio Scene Understanding

Zhiyao Duan
University of Rochester

June 19, 2021

My Background

- Associate professor in ECE and CS at U Rochester
- Directs Audio Information Research (AIR) lab

MUSIC INFORMATION RETRIEVAL
- Music transcription, alignment, source separation, generation, interactive performance

SPEECH PROCESSING
- Speech separation, enhancement, verification, emotion analysis, diarization, text-to-speech, voice transfer

ENVIRONMENTAL SOUND UNDERSTANDING
- Sound search by vocal imitation, sound event detection, source localization

AUDIO-VISUAL PROCESSING
- Talking face generation, music performance analysis and generation, source separation

Motivations and Goals

- Audio is a critical modality in audio-visual scenes (e.g., videos), but has received considerably less attention than the visual modality
- Computer Audition (or machine listening) is a much smaller field than CV
- Bring some new thoughts and perspectives to the CV community
- Receive new ideas from you for solving audio-visual and audio problems

Audio Scene Understanding

In human perception, this is called Auditory Scene Analysis.

The cocktail party problem

(image from http://www.justellus.com/)

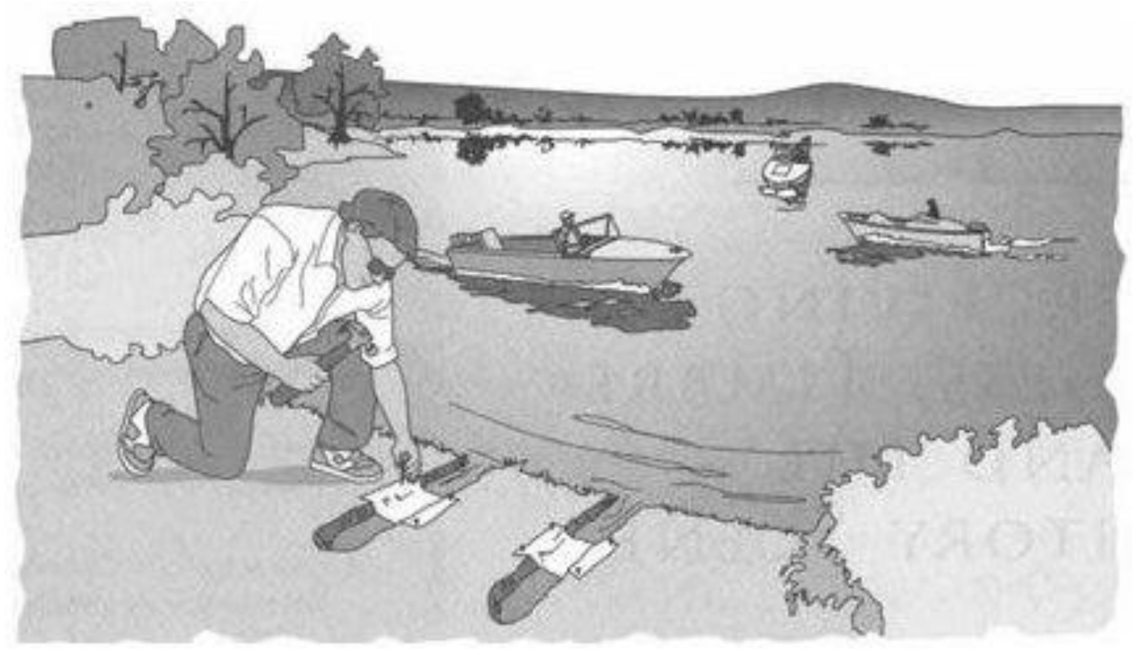
Important Tasks

- What are the sound sources? ---- sound event detection / speaker recognition
- What are they talking about? ---- speech recognition
- What musical notes are played? ---- music transcription

- Where are the sound sources? ---- sound source localization

- What does each source sound like? ---- sound source separation
- Make a particular voice clearer ---- speech enhancement
- Remove the room effect ---- de-reverberation
- ……

It’s not easy!



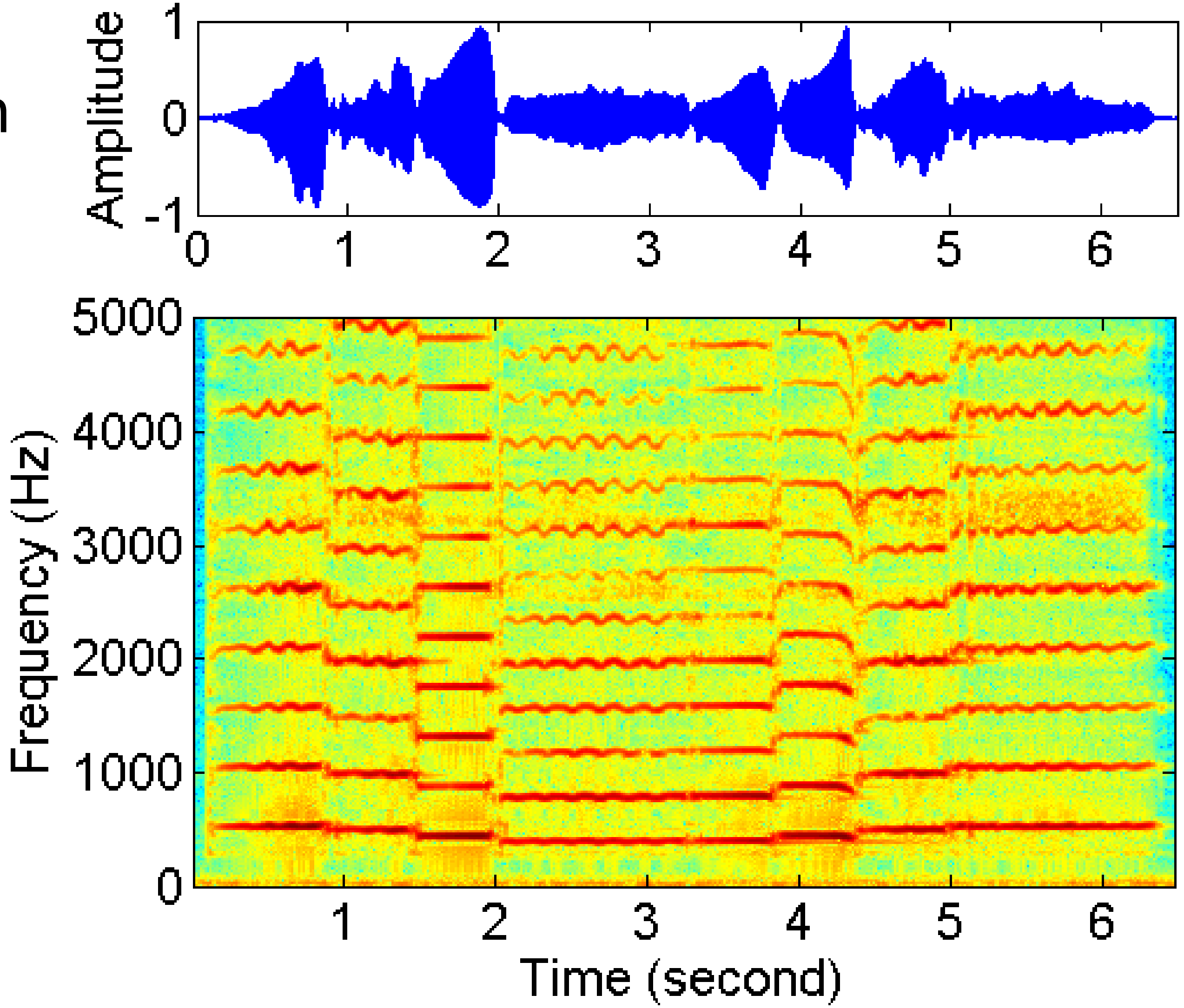
Example from Albert S. Bregman, Auditory Scene Analysis: The Perceptual Organization of Sound. The MIT Press, 1990.

Fundamental Challenges

- Sound sources often overlap (in both time and frequency).
- Various kinds of sound sources
  - Harmonic (e.g., vowel) vs. percussive (e.g., consonants)
  - Short (e.g., mouse clicking) vs. long (car engine)
  - Natural (e.g., environmental sounds) vs. artificial (e.g., speech, music)
- Rich semantic structures (also an advantage!)
  - (Long-term) temporal dependencies in speech and music
  - Harmonic relations among simultaneous sources in music
- Reverberation: ubiquitous and smears sounds significantly
- Difficult to annotate

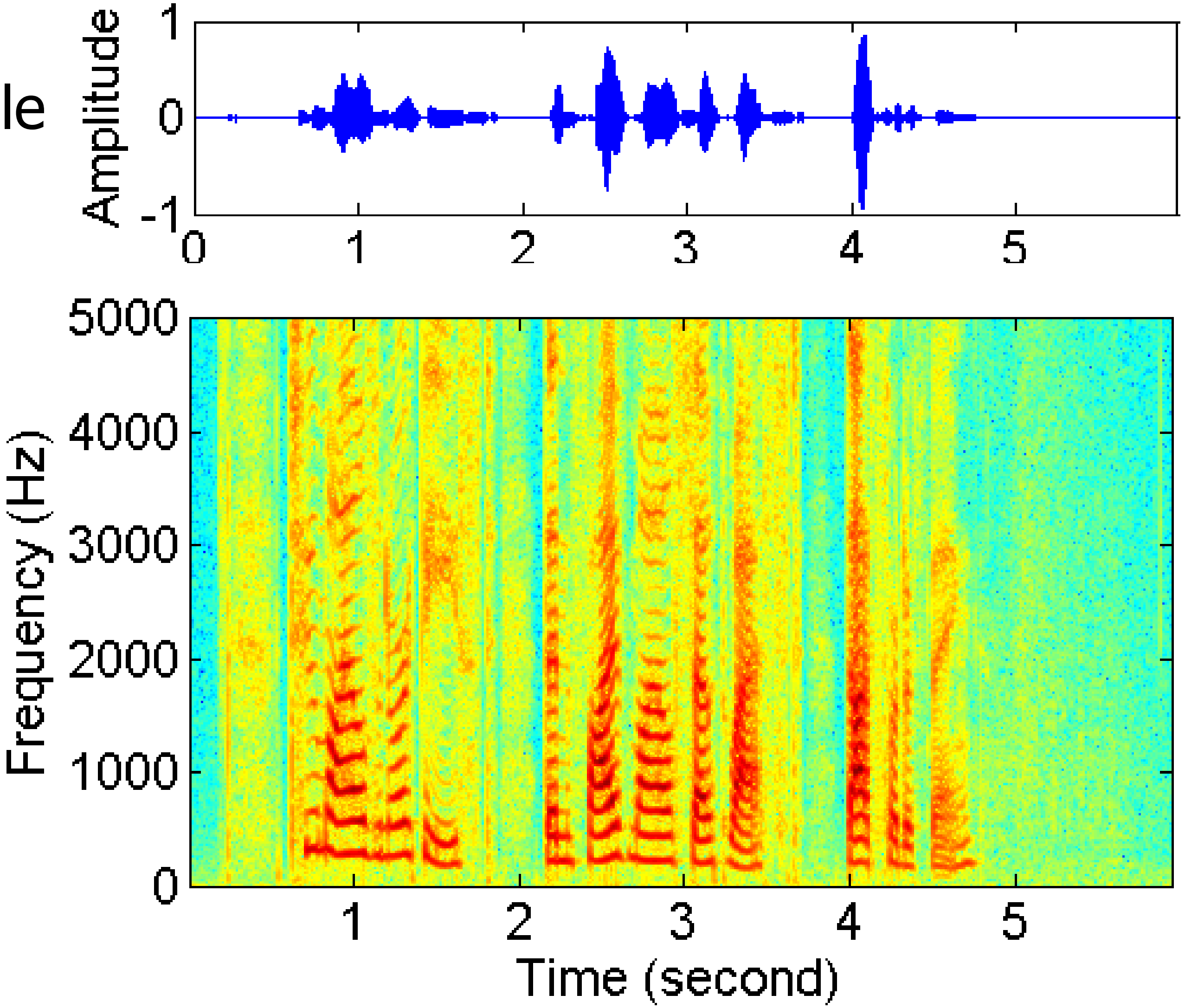
Spectrogram

violin



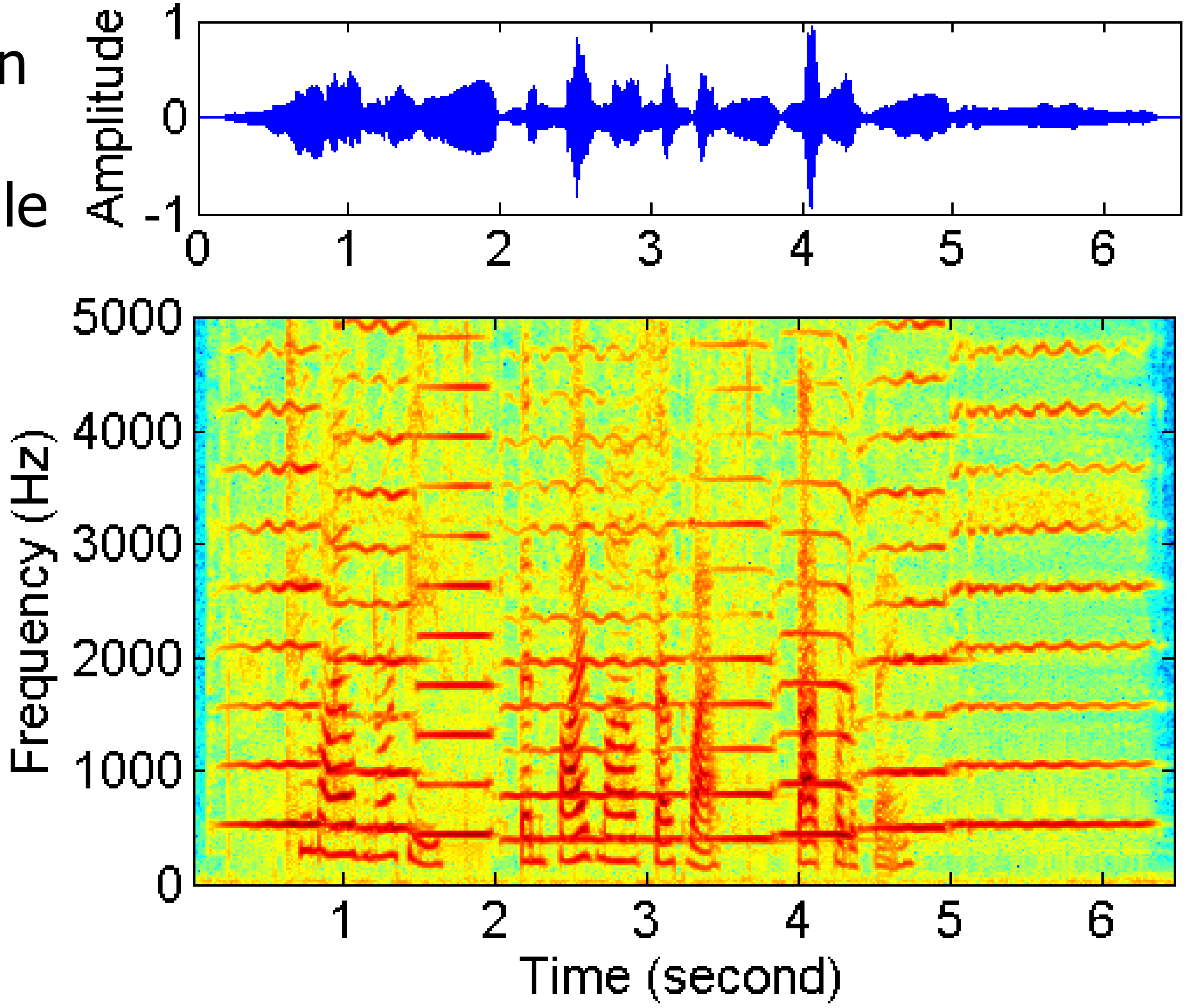
Spectrogram

female



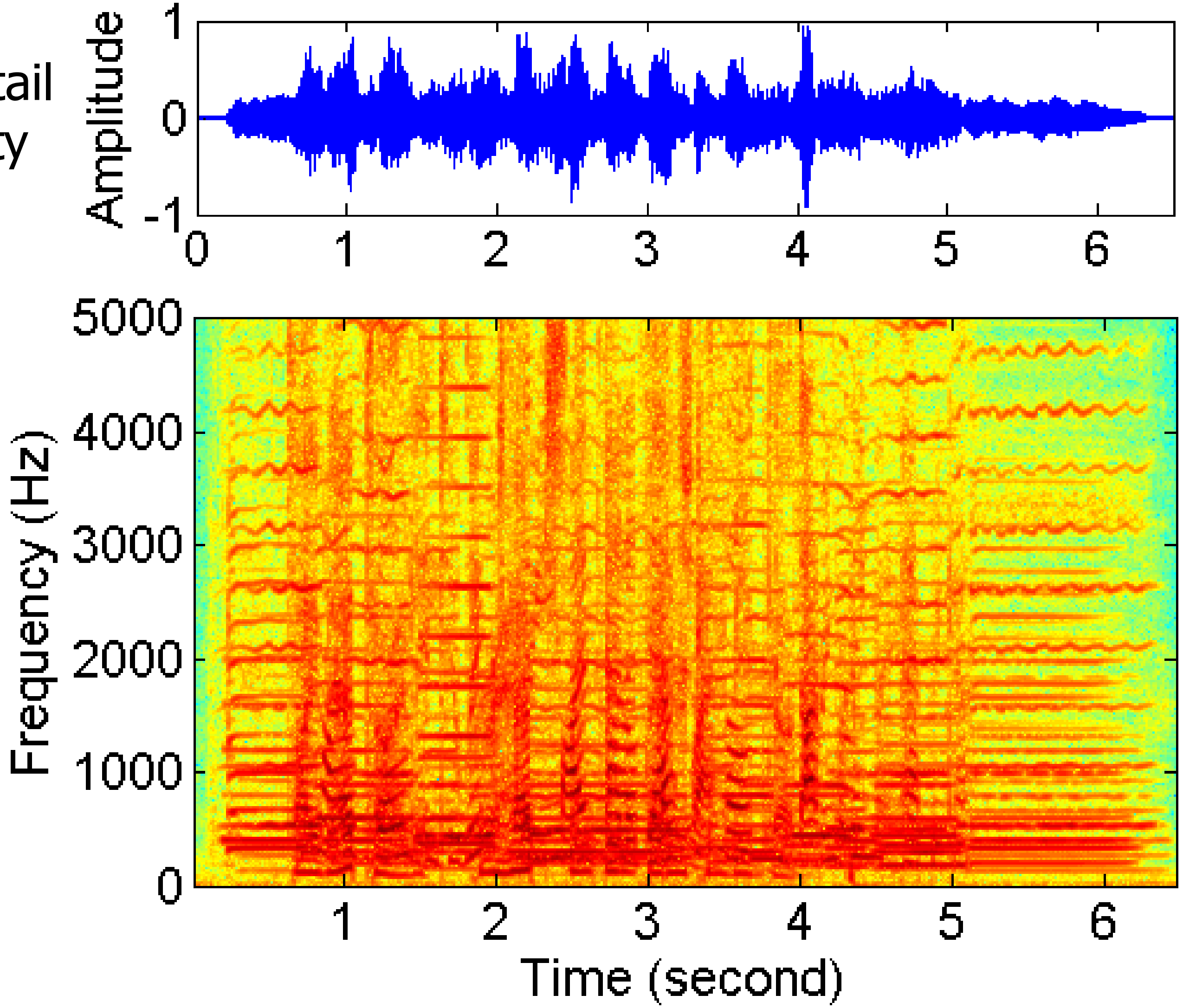
If they sound together

violin
+
female

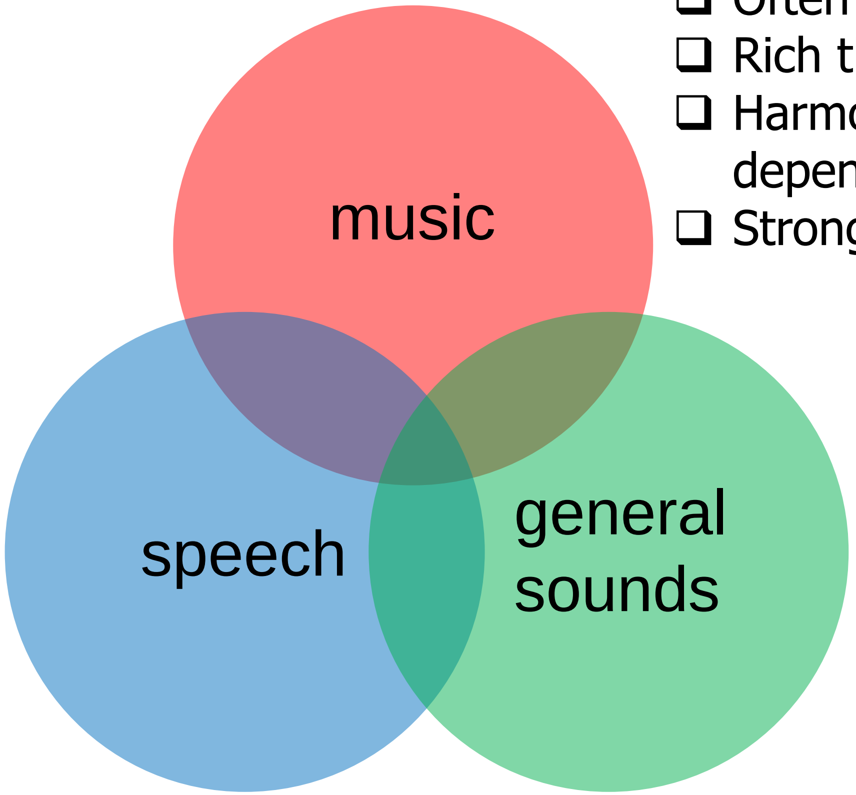


How about this?

cocktail party



Three General Kinds of Sound



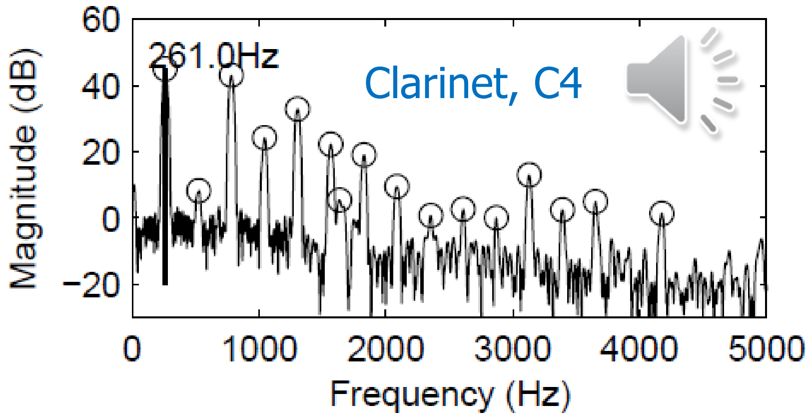
music
- ❑ Often polyphonic
- ❑ Rich timbre
- ❑ Harmonic and temporal dependencies
- ❑ Strong reverberation

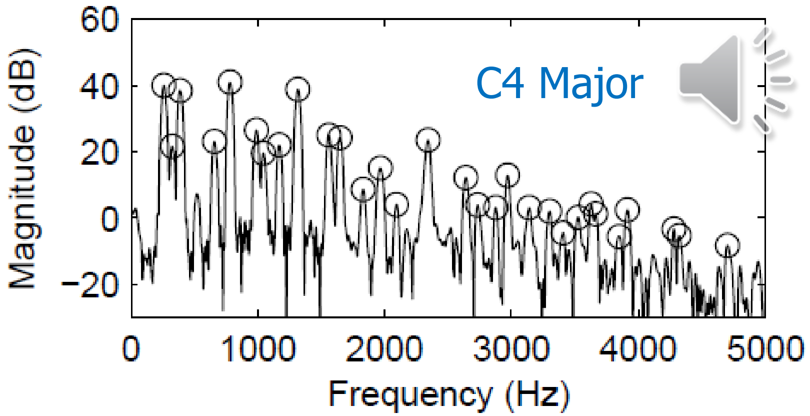
speech
- ❑ Often monophonic but maybe overlapped by noise or other sources
- ❑ Rich semantic meaning
- ❑ Often with reverberation

general sounds
- ❑ Sporadic but can be polyphonic
- ❑ Long-tail distribution
- ❑ Wide range of duration
- ❑ Less structured

Polyphonic Music

- Overlapping harmonics
  - Fundamental frequencies of simultaneous notes are often of small integer ratios, causing many harmonics of different notes to overlap with each other
    - E.g., C4:C5 = 1:2, C4:G4 = 2:3, C4:F4 = 3:4, C4:E4 = 4:5
    - For C4-E4-G4 major chord, harmonic overlap ratios are: C4 (46.7%), E4 (33.3%), G4 (60%)

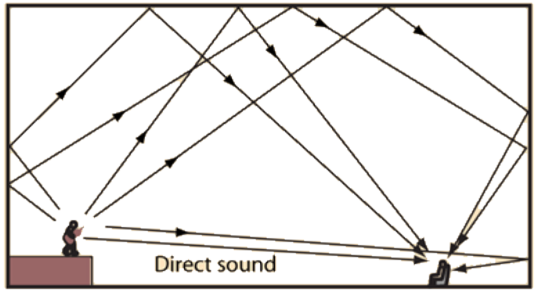




- Temporal structures
  - Repetitions and variations at different time scales: section, phrase, measure, beat
  - Transformations of motifs: transposition, inversion, retrograde (reverse), etc.

Reverberation

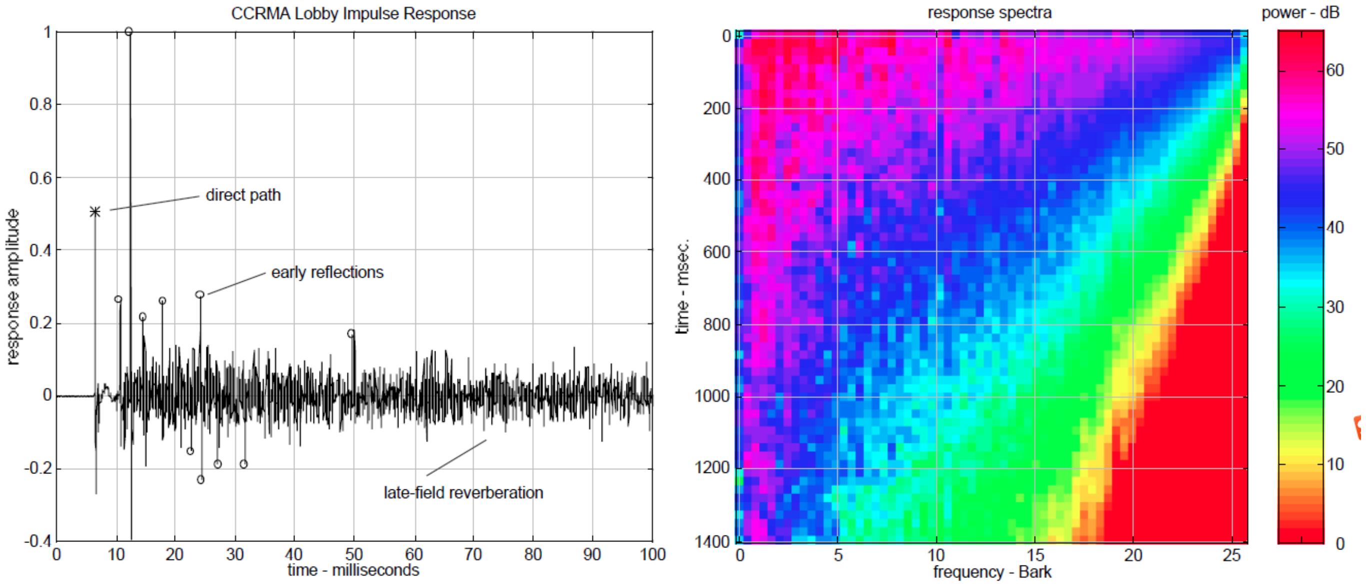
- Room Impulse Response (RIR)
  - Reverberation time RT60: time takes for sound to decay by 60 dB
    - Office ~0.5s, home ~0.7s, classroom ~1s, concert hall ~2s, cathedral ~3.5s
- 1 second is 44,100 samples at 44.1 KHz sampling rate
- Similar to motion blur for images, but with a much large "blurring kernel"



(image from http://hyperphysics.phy-astr.gsu.edu/hbase/Acoustic/reverb.html)

Kernel size
95*95

(images from http://www.cse.cuhk.edu.hk/~leojia/projects/robust_deblur/)



(images from https://ccrma.stanford.edu/~adnanm/SCI220/Music318ir.pdf)

Difficulties in Annotation

- Approach 1: annotate a real recording directly
  - Time consuming to listen through
  - Difficult to attend to simultaneous sound sources

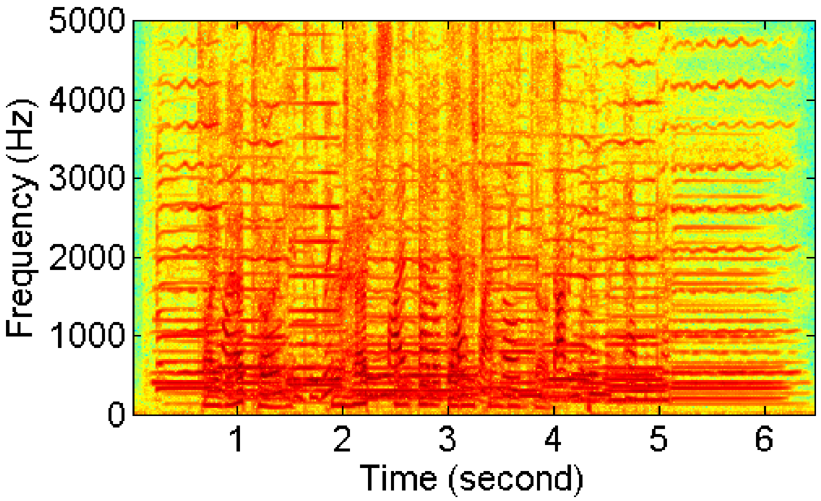
- Approach 2: record each source in isolation and then mix them
  - Difficult to ensure synchronization and coordination
  - Still needs to annotate each source

- Approach 3: mix sound events (musical note samples) based on a transcript (musical score)
  - Requires a concatenative synthesis engine
  - Costly to obtain authentic sound samples
  - Less realistic room acoustics

Vision vs. Audition

- Visual scenes mainly describe objects that reflect light
  - Shape, color, brightness, texture, motion, etc.
- Audio scenes mainly describe sources that emit sound
  - Time, frequency, loudness, location, temporal evolution, etc.

- Visual objects occlude; auditory objects overlap
  - Analyzing audio scenes is like computer vision where
    - Objects are half-transparent
    - Objects change transparency over time
    - Objects disappear and reappear unexpectedly
    - (if with reverb) objects are all strongly motion blurred



Auditory Scene Analysis

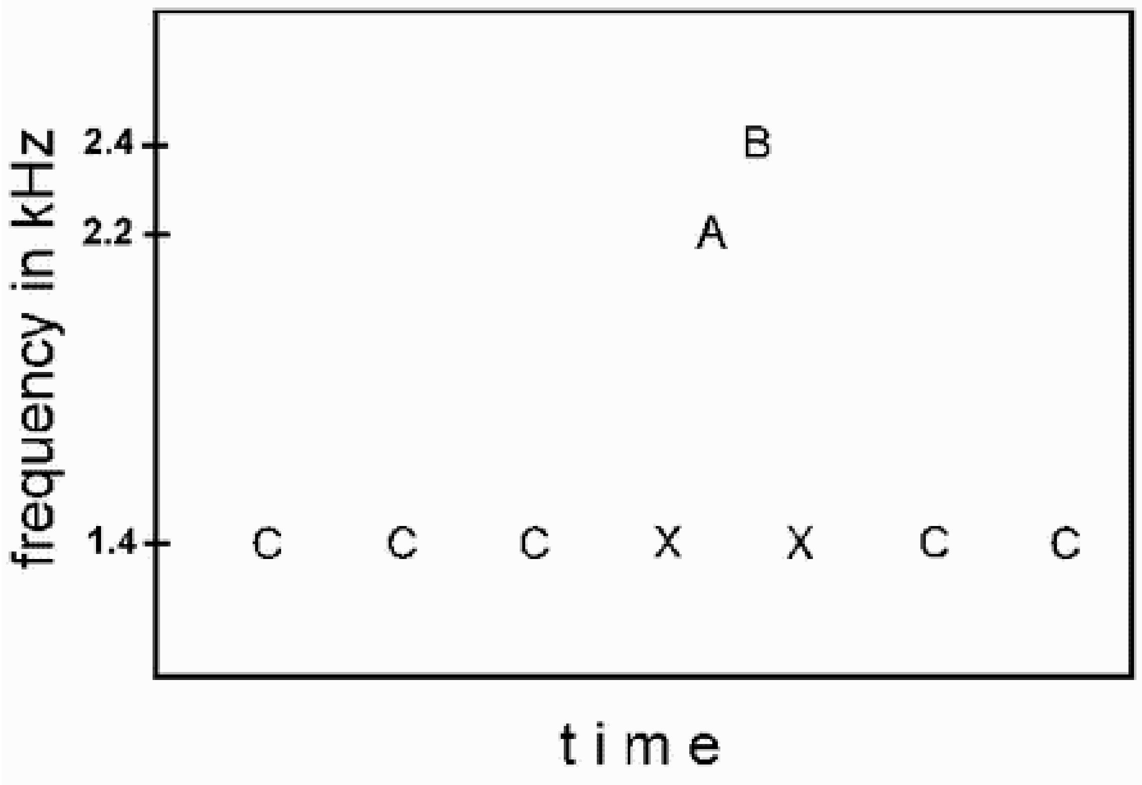
- Studies how human auditory systems analyze auditory scenes through psychoacoustic experiments [1]

- The analysis-synthesis process
  - Decompose scenes into small auditory segments
  - Group segments into auditory streams

- Sequential grouping
  - proximity and similarity in time, frequency, loudness, timbre, spatial location; related rhythm
- Simultaneous grouping
  - harmonicity; common fate in onset/offset, frequency, amplitude, and spatial location

- [1] Albert S. Bregman, Auditory Scene Analysis: The Perceptual Organization of Sound. The MIT Press, 1990.

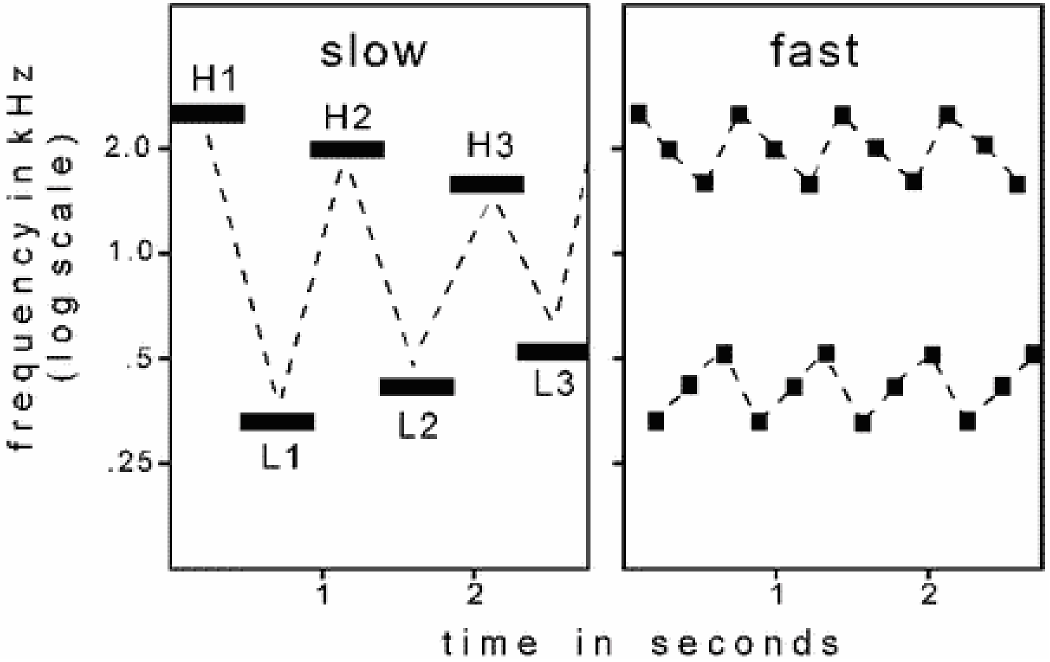
Exclusive Allocation



- The allocation of the X tones are different when the C tones are played or not, and it affects our perception of the A and B tones.

Example from Albert S. Bregman, Auditory Scene Analysis: The Perceptual Organization of Sound. The MIT Press, 1990.

Stream Segregation



- High and low tones are segregated when played fast
- Can you tell the order of the six tones?

Example from Albert S. Bregman, Auditory Scene Analysis: The Perceptual Organization of Sound. The MIT Press, 1990.

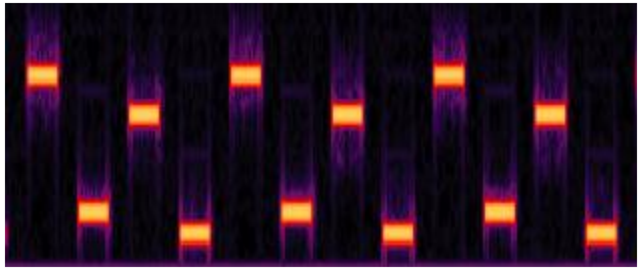
Stream Segregation in Music

Toccata and Fugue in d minor, J.S. Bach

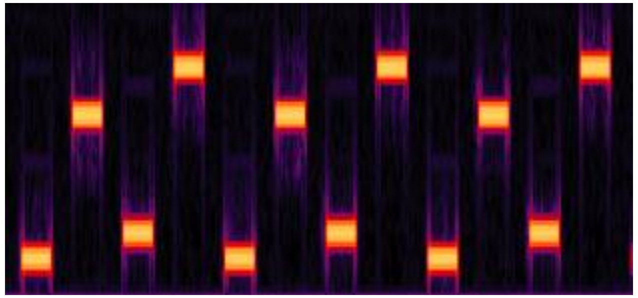
Arrangement for violin solo, performed by Sergei Krylov
(video from https://www.youtube.com/watch?v=R_tu63ypB6I)

Primitive vs. Learned

H1-L1-H2-L2



L2-H2-L1-H1



- Infants cannot discriminate the two stimuli, which indicates that they perform stream segregation of the high and low tones.

Example from Albert S. Bregman, Auditory Scene Analysis: The Perceptual Organization of Sound. The MIT Press, 1990.

Primitive vs. Learned

- Listening to a stimulus repeatedly can improve performance in stream segregation

- Easier to follow a friend’s voice than a stranger’s in a noisy environment
  - Prior knowledge of timbre helps

- Music training helps music scene understanding
  - Prior knowledge of music theory, composition rules, music style, etc. helps

Super Ability in Music Scene Understanding

- "In Rome, he (14 years old) heard Gregorio Allegri's *Miserere* once in performance in the Sistine Chapel. He wrote it out entirely from memory, only returning to correct minor errors..."

  -- Gutman, Robert (2000). *Mozart: A Cultural Biography*

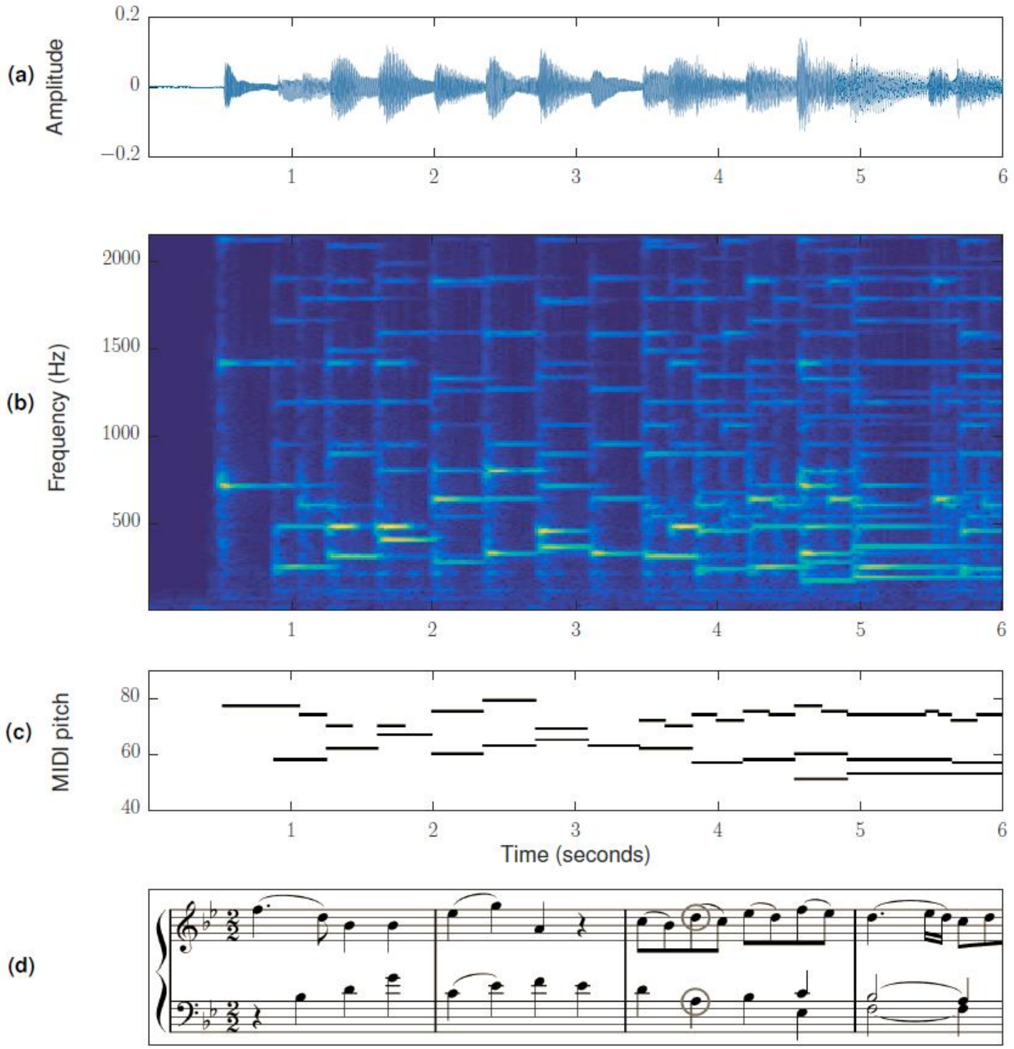
Wolfgang Amadeus Mozart

Selected Important Tasks

- Automatic Music Transcription
- Sound Event Detection
- Audio Source Separation

Automatic Music Transcription

- Converting music audio into a symbolic representation (e.g., MIDI or music notation)
- Consider by many the "Holy Grail" in Music Information Retrieval (MIR)
- Applications: performance analysis, education, search, etc.
- Challenges
  - Polyphonic
  - Rich timbre
  - Music language model
  - Lack of annotated data

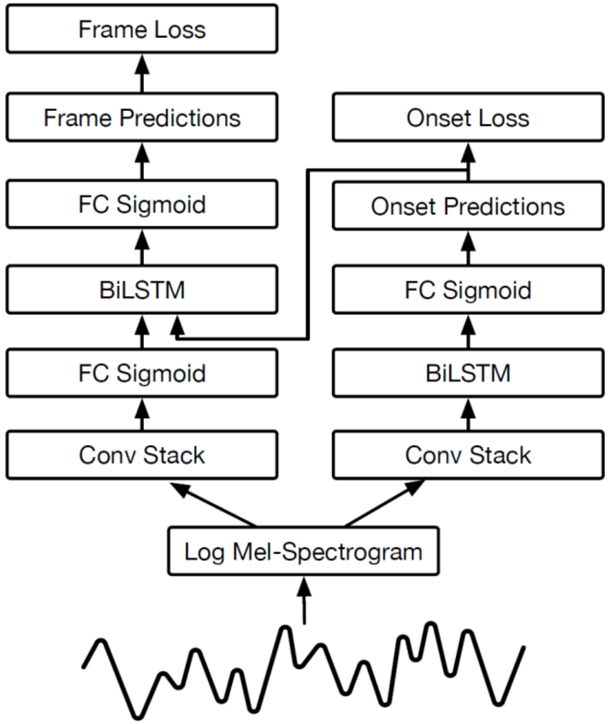


Emmanouil Benetos*, Simon Dixon*, Zhiyao Duan*, and Sebastian Ewert*, **Automatic music transcription: an overview**, *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 20-30, 2019. (*alphabetic order)

Piano Transcription

- Disklavier piano: acoustic piano that records MIDI and can reproduce audio from MIDI
  - In this way, audio recordings and MIDI transcriptions are obtained easily
- Datasets: MAPS [1], MAESTRO [2]

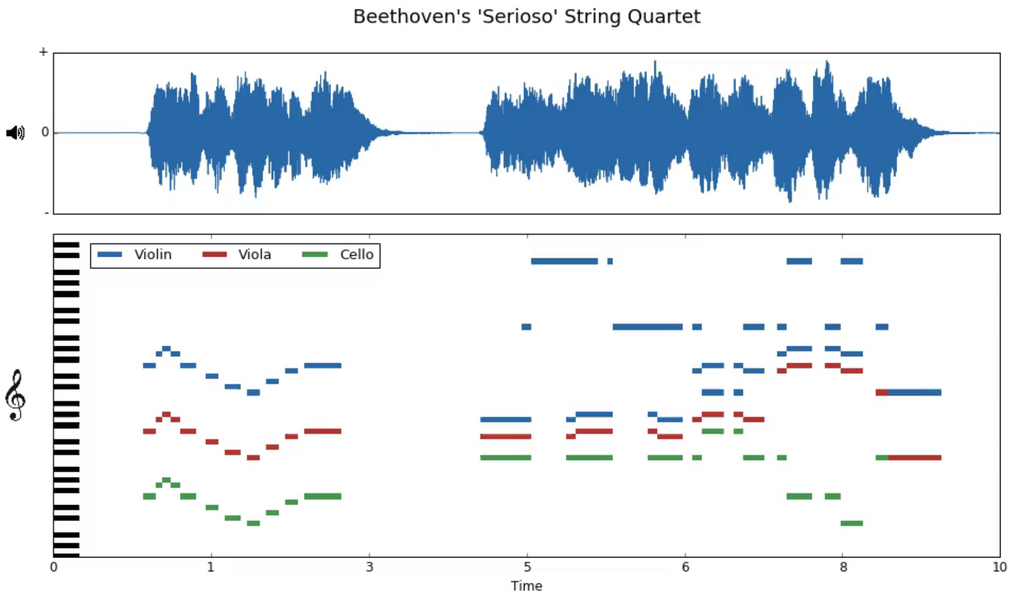
- Onsets & Frames [3]

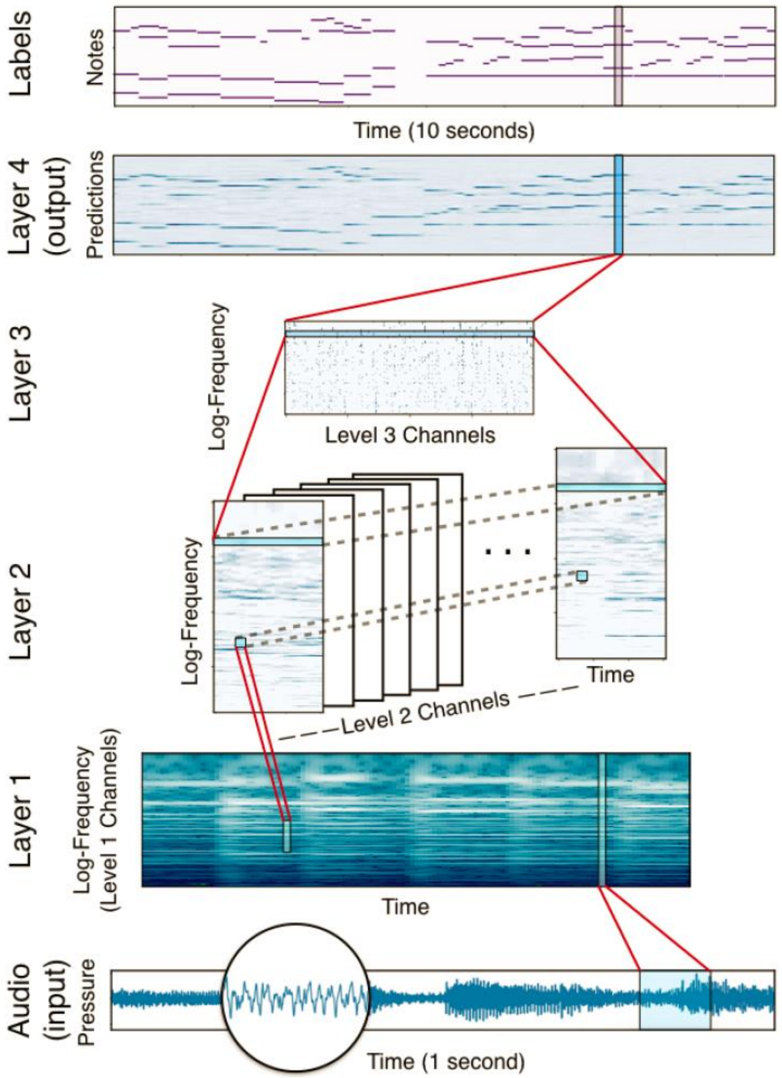


- [1] V. Emiya, R. Badeau, and B. David. **Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle**. IEEE/ACM TASLP, 2010.
- [2] C. Hawthorne, A. Stasyuk, A. Roberts, I. Simon, C.-Z. A. Huang, S. Dieleman, E. Elsen, J. Engel, & D. Eck. **Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset**. ICLR, 2019.
- [3] C. Hawthorne, E. Elsen, J. Song, A. Roberts, I. Simon, C. Raffel, J. Engel, S. Oore, & D. Eck. **Onsets and frames: Dual-objective piano transcription**. arXiv preprint arXiv:1710.11153. 2017.

Multi-Instrument Transcription

- MusicNet [1]
  - 330 classical pieces with MIDI alignments using Dynamic Time Warping (DTW)





[1] J. Thickstun, Z. Harchaoui, and S. Kakade, **Learning features of music from scratch**, ICLR, 2017.
[2] J. Thickstun, Z. Harchaoui, D.P. Foster, S.M. Kakade, **Invariances and data augmentation for supervised music transcription,** ICASSP, 2018.

Music is not just about sound

- University of Rochester Multimodal Music Performance Dataset (URMP)
  - 44 ensemble performances with 13 kinds of instruments
  - Isolated recordings and annotations

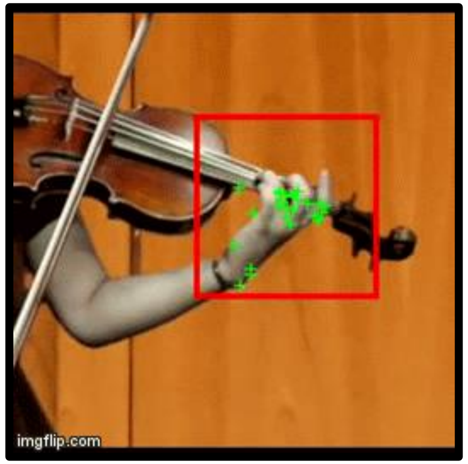
Bochen Li*, Xinzhao Liu*, Karthik Dinesh, Zhiyao Duan, and Gaurav Sharma, **Creating a multitrack classical music performance dataset for multi-modal music analysis: challenges, insights, and applications**, *IEEE Transactions on Multimedia*, vol. 21, no. 2, pp. 522-535, 2019. (*equal contribution)

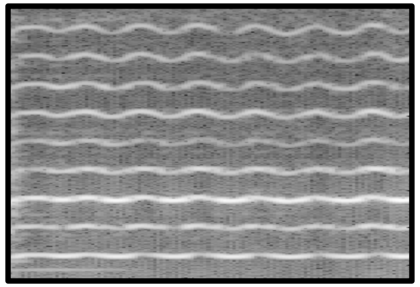
Audio-Visual Music Analysis

- Key is to build audio-visual correspondence
- Static
  - Fixed image ←→ Audio frame, e.g., [1]
  - E.g., Posture of a flutist ←→ Play/Nonplay activity
  - E.g., Piano fingering ←→ Music transcription
- Dynamic, instrument specific
  - Dynamic movement ←→ Audio feature fluctuation
  - E.g., Guitarist's strumming hand ←→ Rhythmic pattern
  - E.g., Violinist rolling left hand ←→ Vibrato [2]

Dynamic, general
  - Co-factorization of audio/visual fluctuations [3]
  - Learning audiovisual motion embeddings [4,5]

(images from https://www.123rf.com/photo_39591413_young-flute-player-performing-indoors-against-white-background.html)

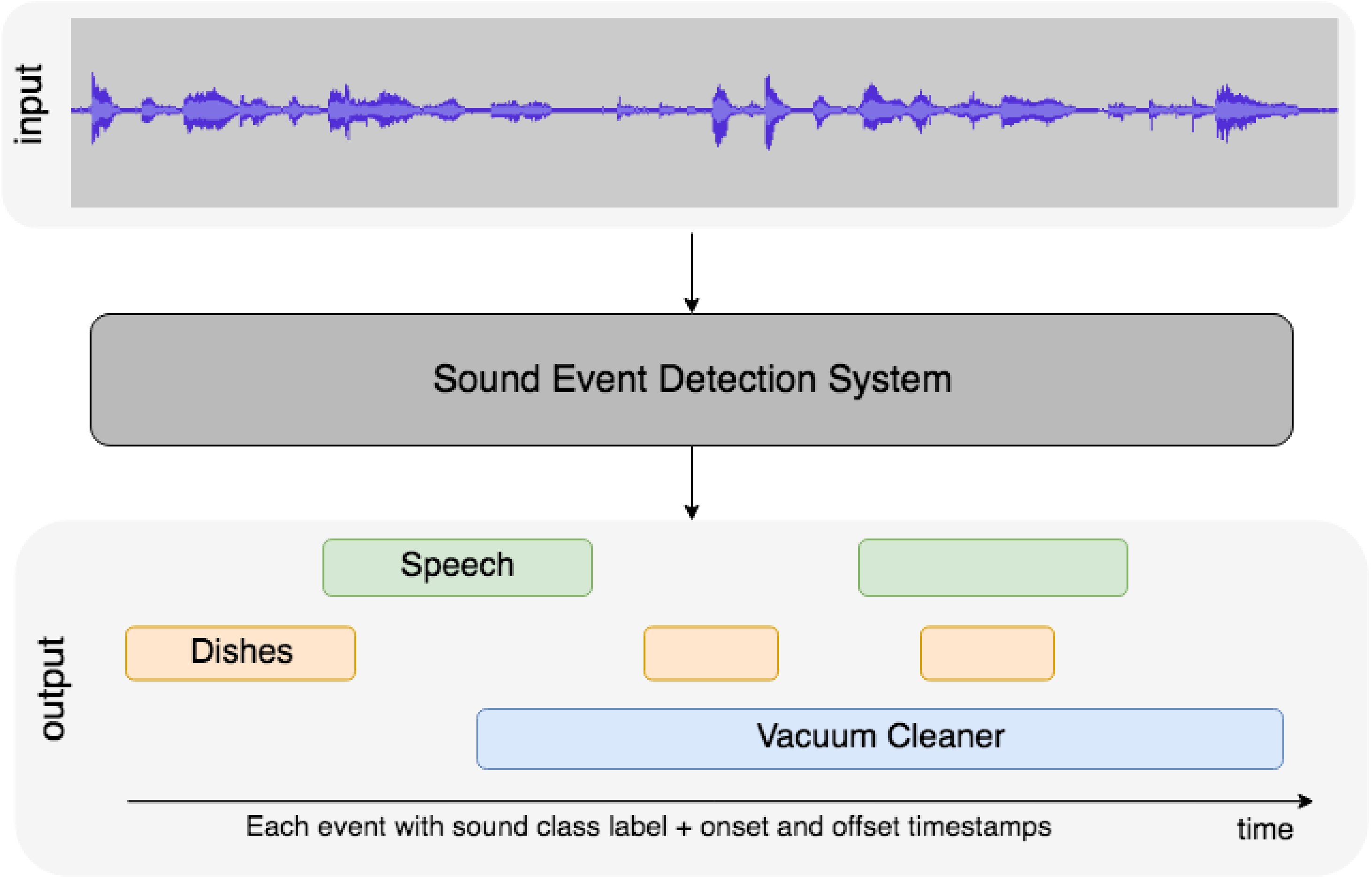




[1] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, and A. Torralba, **The Sound of Pixels**, ECCV, 2018.
[2] B. Li, K. Dinesh, G. Sharma, and Z. Duan, **Video-based vibrato detection and analysis for polyphonic string music**, *ISMIR*, 2017.
[3] S. Parekh, S. Essid, A. Ozerov, N.Q. Duong, P. Pérez, & G. Richard. **Motion informed audio source separation**. ICASSP 2017.
[4] H. Zhao, C. Gan, W.-C. Ma, A. Torralba. **The Sound of Motions**, ICCV, 2019.
[5] C. Gan, D. Huang, H. Zhao, J. B. Tenenbaum, A. Torralba, **Music gesture for visual sound separation**, CVPR 2020.

Sound Event Detection

- IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE) – Task 4
- Datasets
  - Synthetic mixtures (strong labels)
  - Real recordings (weak labels)



(image from http://dcase.community/challenge2021/task-sound-event-detection-and-separation-in-domestic-environments)

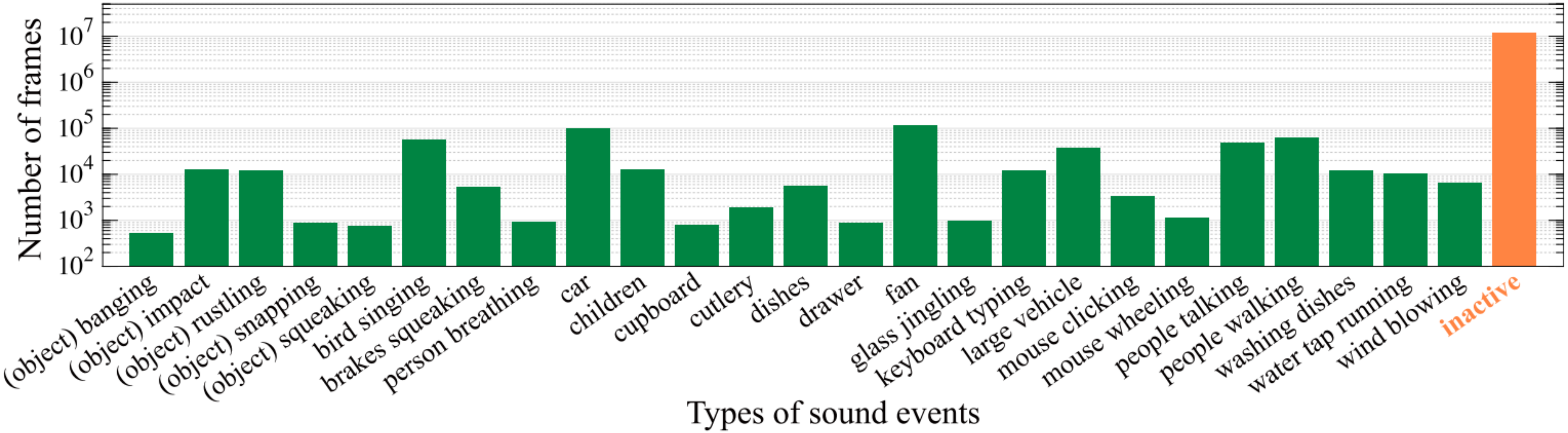
Sound Event Detection

- Best Scoring System [1] in DCASE2020
  - Conformer model (CNN + Transformer) [2]
  - Semi-supervised learning with Mean-Teacher technique [3]
  - Data augmentation with time shifting and mixup [4]
  - Median filtering and score fusion

- [1] K. Miyazaki, T. Komatsu, T. Hayashi, S. Watanabe, T. Toda, K. Takeda, **Convolution-augmented transformer for semi-supervised sound event detection**, DCASE2020 Challenge, 2020.
- [2] A. Gulati, J. Qin, C.-C. Chiu, et al., **Conformer: convolution-augmented transformer for speech recognition**, arXiv preprint arXiv:2005.08100, 2020.
- [3] A. Tarvainen and H. Valpola, **Mean teachers are better role models: Weight-averaged consistency targets improve semisupervised deep learning results**, NIPS, 2017.
- [4] H. Zhang, M. Cisse, Y. N. Dauphin, and D. LopezPaz, **Mixup: Beyond empirical risk minimization**, arXiv preprint arXiv:1710.09412, 2017.

Interesting Directions

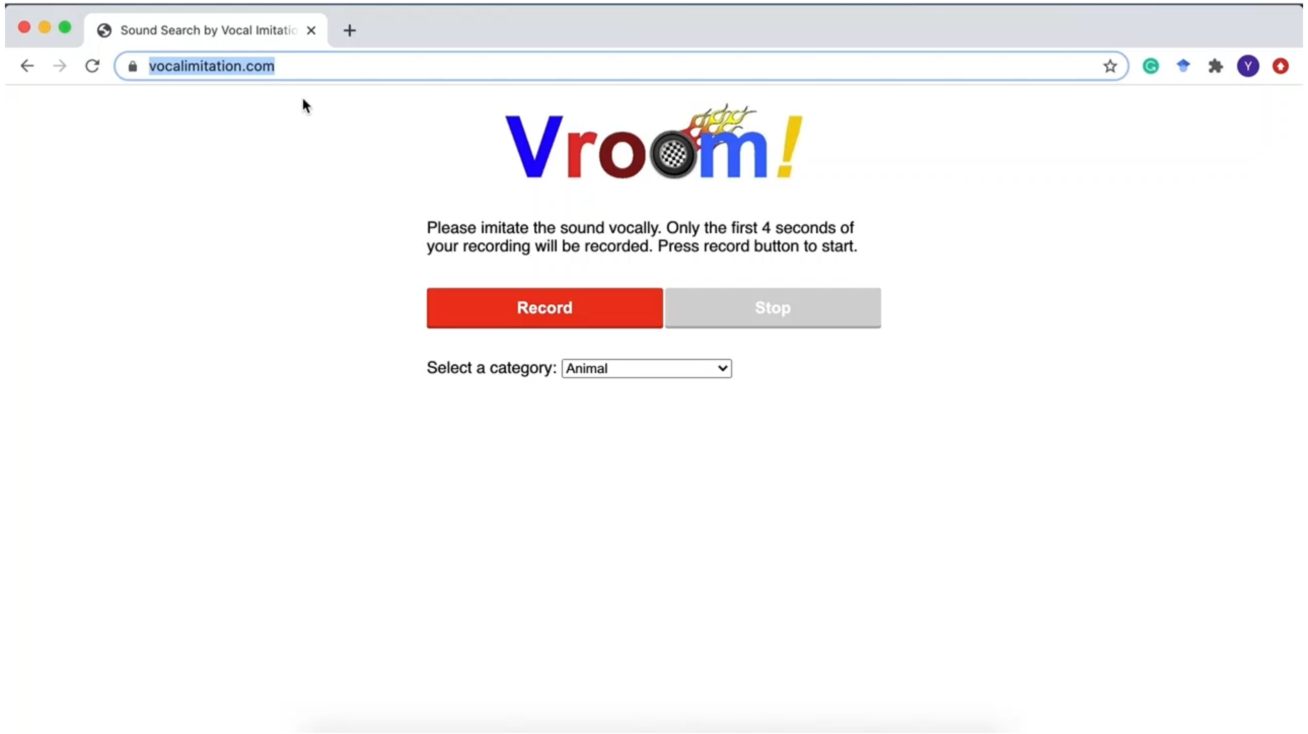
- Addressing data imbalance issue [1]
  - Modify binary cross entropy loss to: *simple reweighting loss, inverse frequency loss, asymmetric focal loss, focal batch Tversky loss*

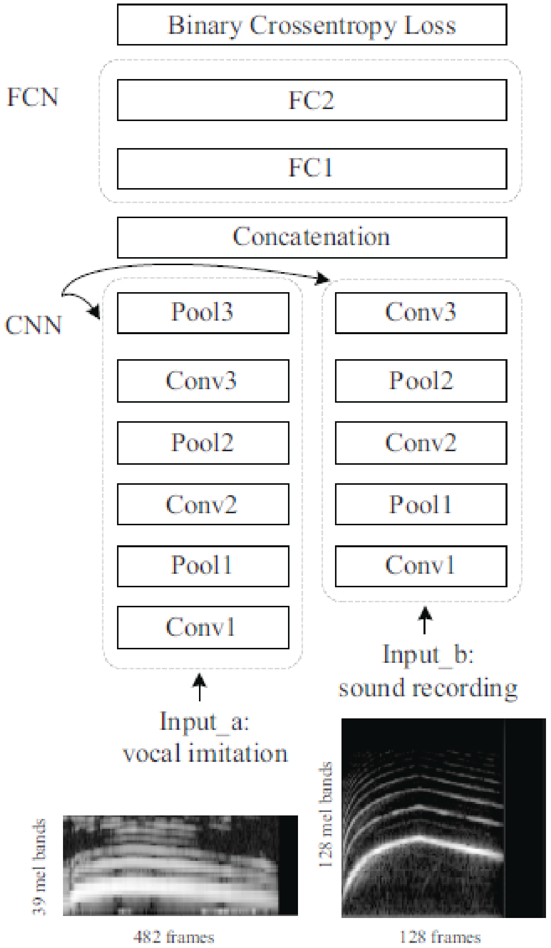


- [1] K. Imoto, S. Mishima, Y. Arai, & R. Kondo, **Impact of sound duration and inactive frames on sound event detection performance**, ICASSP, 2021.

Interesting Directions

- Few-shot learning to open-set scenarios [1]
- Sound retrieval (by vocal imitation [2, 3])

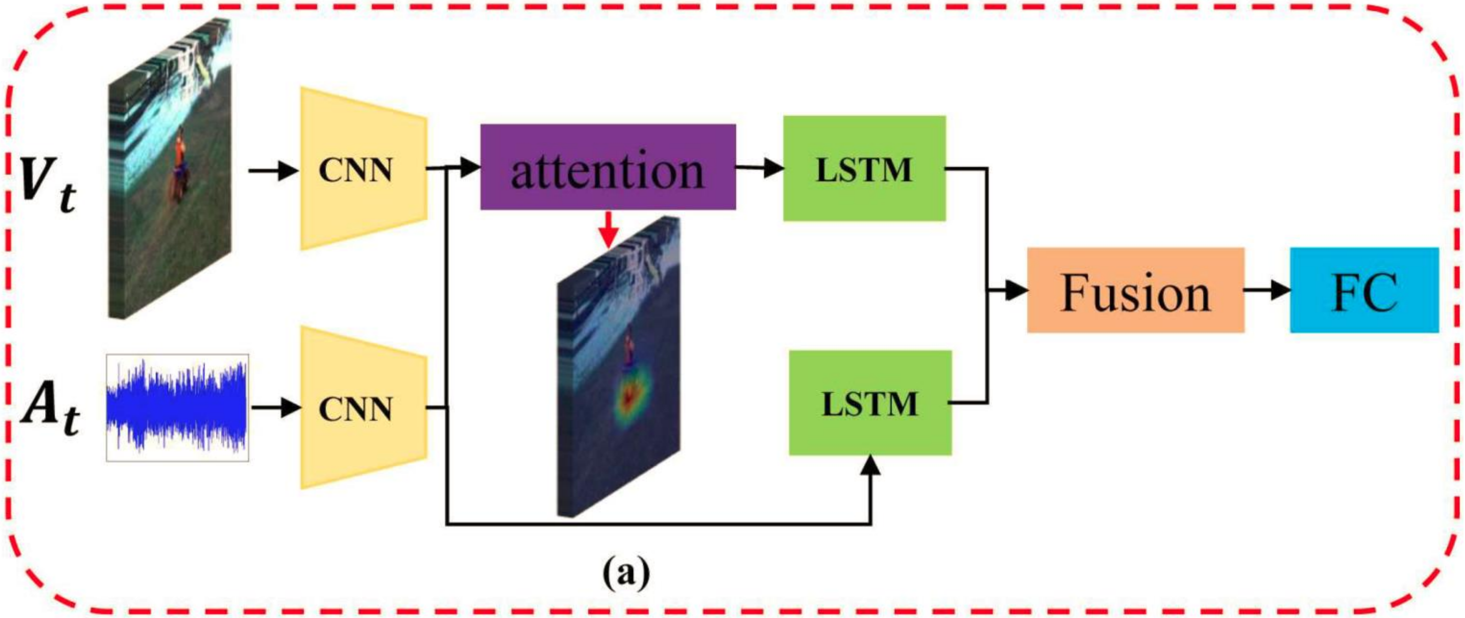




- [1] Y. Wang, J. Salamon, N. J. Bryan, & J. P. Bello. **Few-shot sound event detection**, ICASSP, 2021.
- [2] Y. Zhang, B. Pardo, & Z. Duan, **Siamese style convolutional neural networks for sound search by vocal imitation**, IEEE/ACM TASLP 2019.
- [3] Y. Zhang, J. Hu, Y. Zhang, B. Pardo, & Z. Duan, **Vroom!: A search engine for sounds by vocal imitation queries**, CHIIR, 2020.

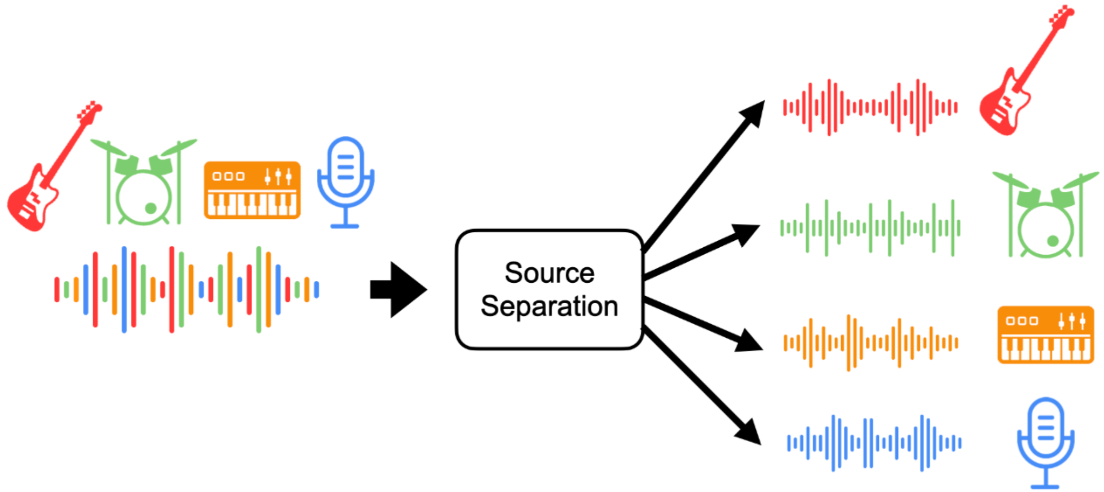
With Visual Information

- Audio-Visual Event Detection
  - Audio-visual association helps to fuse information from both modalities



Y. Tian, J. Shi, B. Li, Z. Duan, & C. Xu, **Audio-visual event localization in unconstrained videos**, ECCV, 2018.

Audio Source Separation

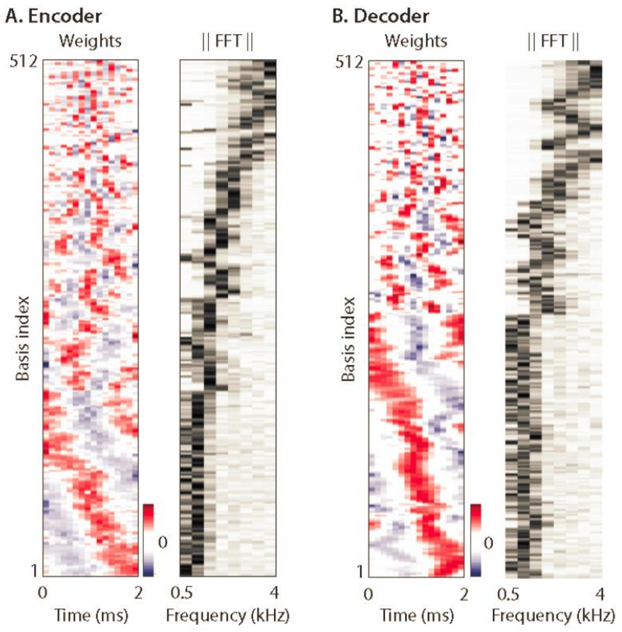


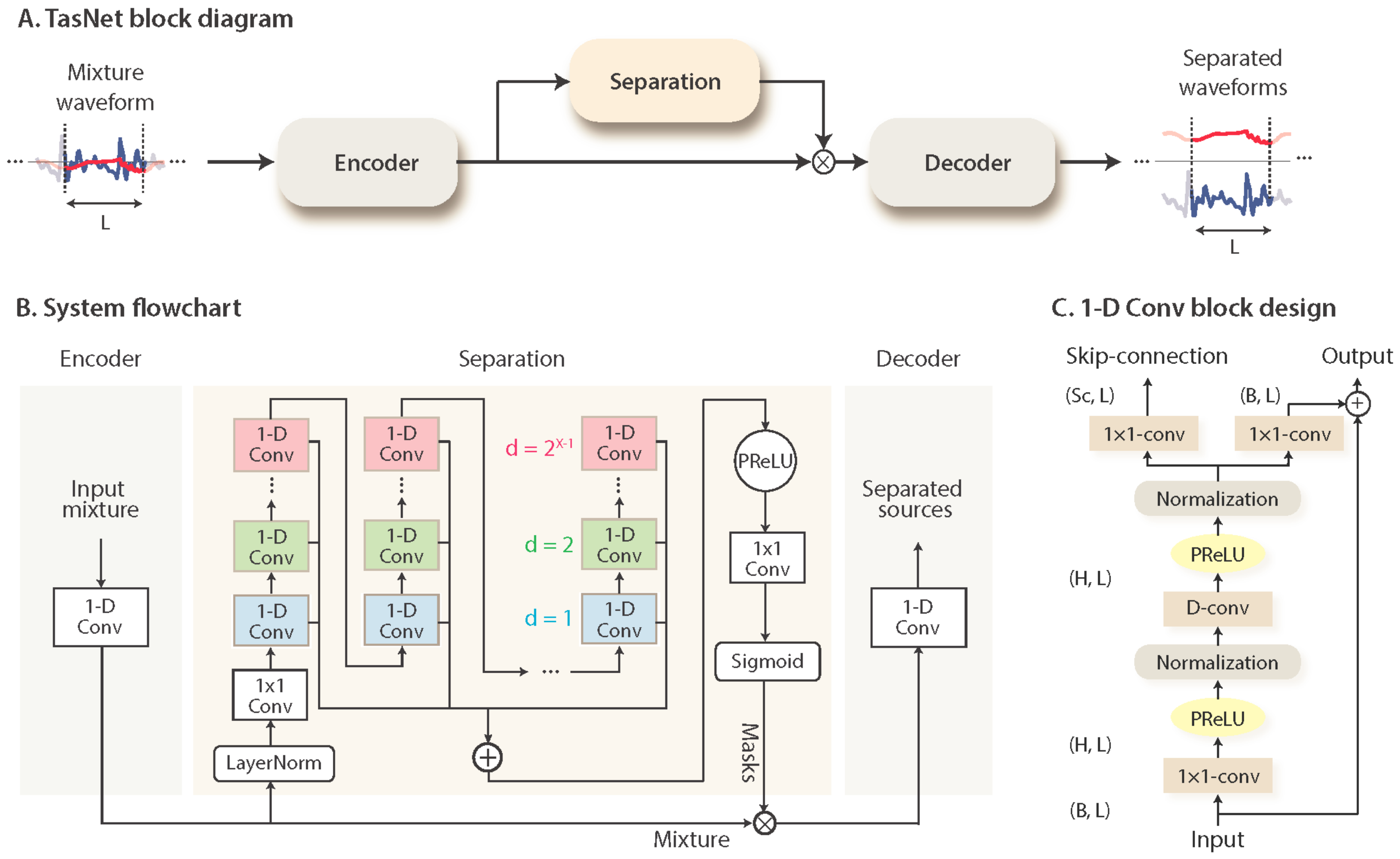
(image from https://source-separation.github.io/tutorial/landing.html)

- Speech separation, speech enhancement
  - Training supervised methods on random mixtures of speech (and noise)
- Music: singing voice separation, multi-instrument separation
  - Interesting finding: it is helpful to use a large amount of random mixtures of instrumental sources in training!

State of The Art

- Conv-TasNet [1]: time-domain audio separation network
- The separation module was later replaced by Dual-Path RNN (DPRNN) [2]





- [1] Y. Luo, N. Mesgarani, **Conv-TasNet: surpassing ideal time-frequency magnitude masking for speech separation**, IEEE/ACM TASLP, 2019.
- [2] Y. Luo, Z. Chen, T. Yoshioka, **Dual-path RNN: efficient long sequence modeling for time-domain single-channel speech separation**, ICASSP, 2020.

Unseen Number of Sources

- Methods with supervised training cannot generalize to unseen numbers of sources (e.g., train on 2-speaker mixtures but test on 4-speaker mixtures)
- Key idea to generalization of SANet [1]: anchor each source to a fixed position in an embedding space through speaker loss and compactness loss.

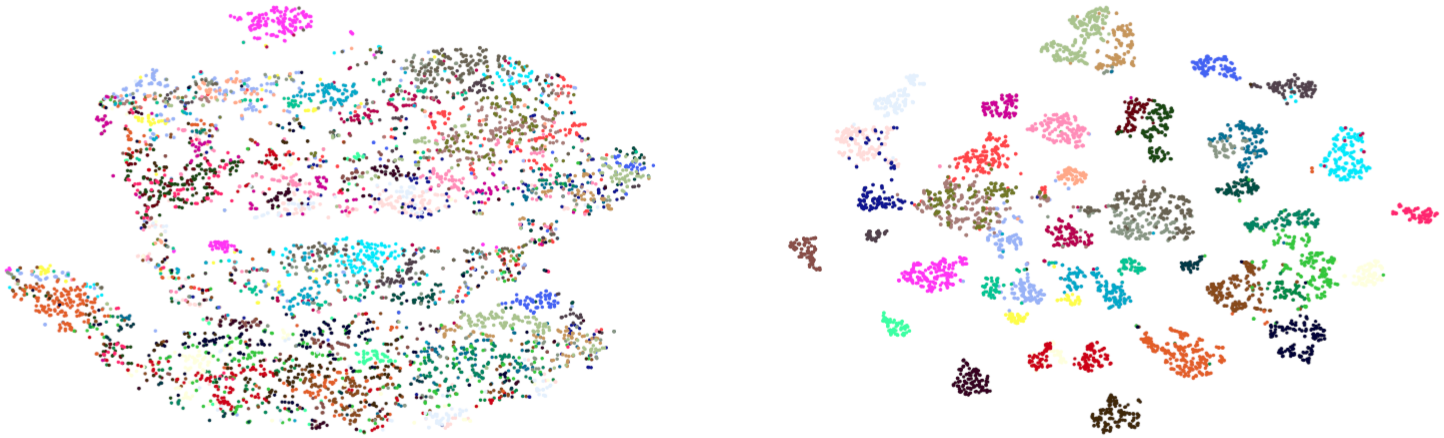
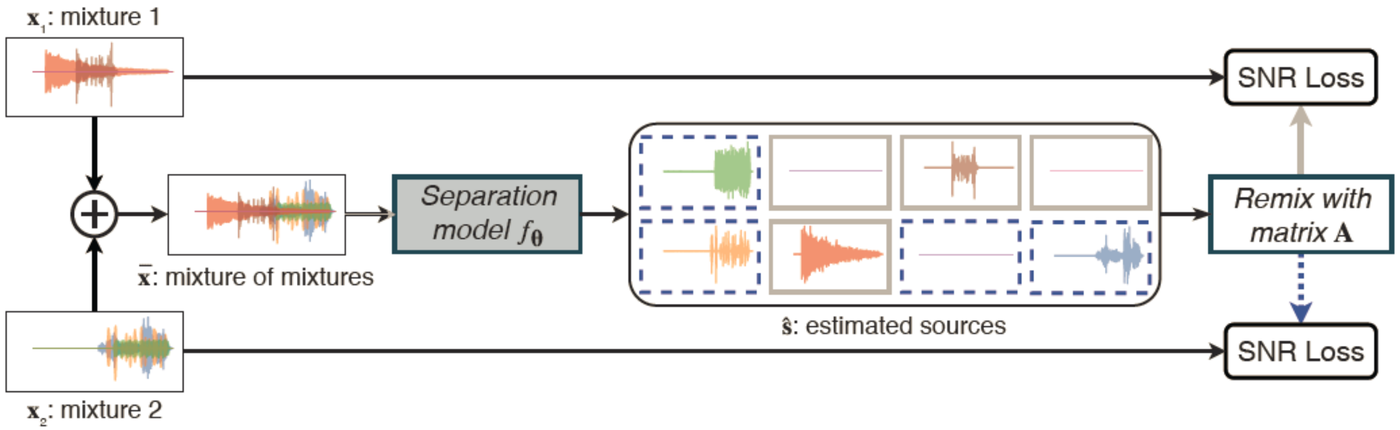


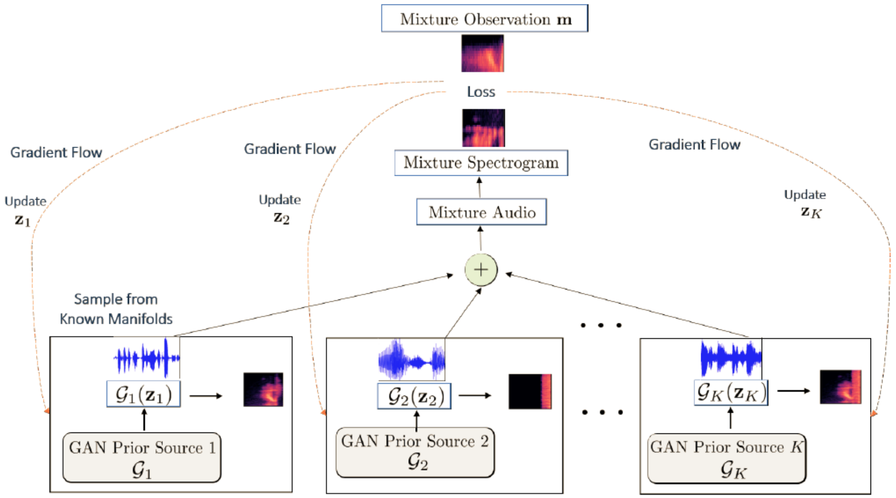
Fig. 2. Estimated attractors (k-means centroids) of test mixtures visualized by t-SNE. Each color represents a speaker. Left: Conv-DANet. Right: SANet.

[1] F. Jiang & Z. Duan, **Speaker attractor network: generalizing speech separation to unseen numbers of sources**, IEEE SPL, 2021.

Unsupervised Separation

- Humans do not listen to "parallel" data to learn to separate audio.
- When only mixtures available
  - Traditional: Independent Component Analysis (ICA), Computational Auditory Scene Analysis (CASA) methods
  - Self-supervised learning: Mixture Invariant Training [1]
- When clean sources (non-parallel to mixture) available
  - Traditional: Dictionary learning on these sources (e.g., NMF, sparse coding)
  - Impose GAN priors (e.g., WaveGAN) [2]

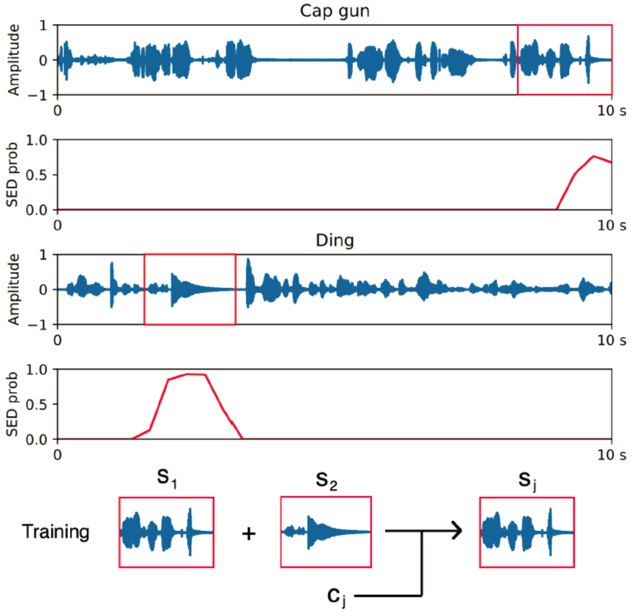


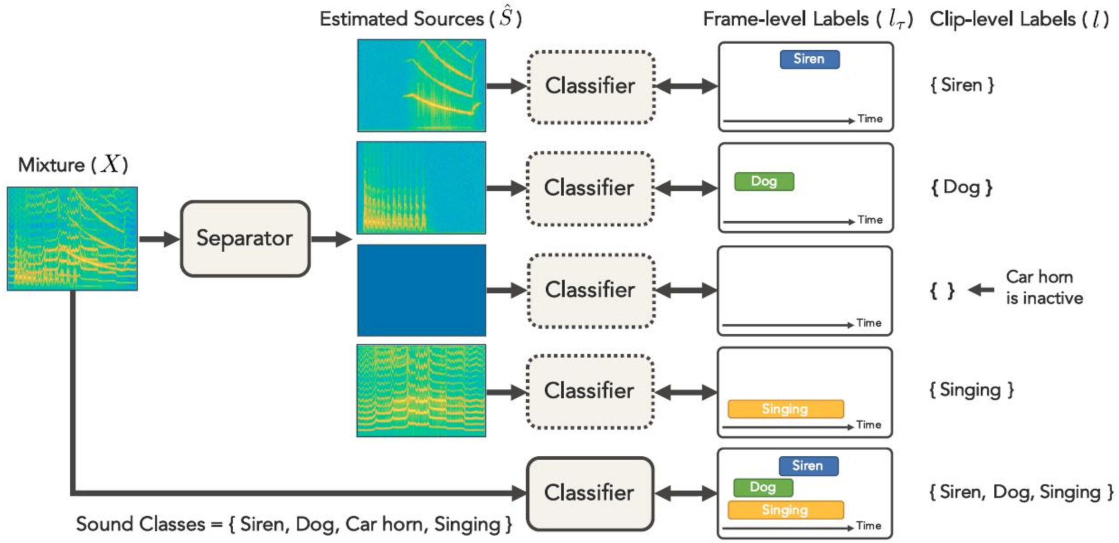


[1] S. Wisdom, E. Tzinis, H. Erdogan, R. J. Weiss, K. Wilson, & J. R. Hershey. **Unsupervised Sound Separation Using Mixture Invariant Training.** NeurIPS 2020.
[2] V. Narayanaswamy, J. J. Thiagarajan, R. Anirudh, & A. Spanias. **Unsupervised audio source separation using generative priors**. Interspeech, 2020.

Universal Sound Separation

- New task and dataset on separating general sounds (hundreds of sound classes) [1,2]
- Use sound event detection to generate training segments and weak labels [3]
- Use sound event detection to provide weak labels [4]





- [1] I. Kavalerov1, S. Wisdom, H. Erdogan, B. Patton, K. Wilson, J. Le Roux, J. R. Hershey, **Universal sound separation**, WASPAA, 2019.
- [2] S. Wisdom, H. Erdogan, D.P. Ellis, R. Serizel, N. Turpault, E. Fonseca, J. Salamon, P. Seetharaman, J.R. Hershey. **What's all the fuss about free universal sound separation data?**, ICASSP, 2021.
- [3] Q. Kong, Y. Wang, X. Song, Y. Cao, W. Wang, M. D. Plumbley, **Source separation with weakly labelled data: an approach to computational auditory scene analysis**, ICASSP, 2020.
- [4] F. Pishdadian, G. Wichern, & J. Le Roux, **Finding strength in weakness: learning to separate sounds with weak supervision**, IEEE/ACM TASLP, 2020.

Summary

- Fundamental research questions in audio scene understanding
  - Recognition, separation, de-reverberation, localization,
- Unique properties and challenges of audio scenes
  - Polyphonic, various timbre, rich structures, reverberation, difficult to annotate
- Inspirations from human auditory scene analysis
- Important tasks, state of the art approaches, and interesting directions
  - Automatic music transcription
  - Sound event detection
  - Audio source separation

- My questions for you:
  - Do you find audio scene understanding helpful in vision tasks?
  - Can you find novel ways to use visual information to help audio understanding?


Thank You!