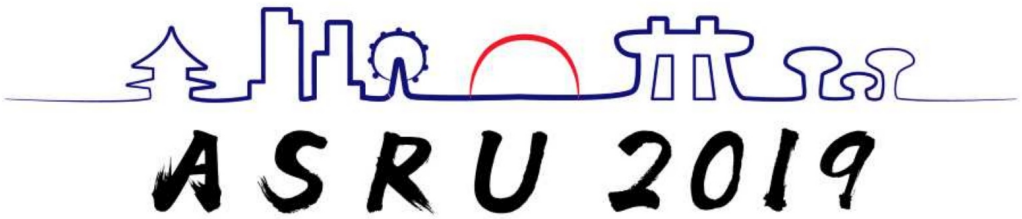
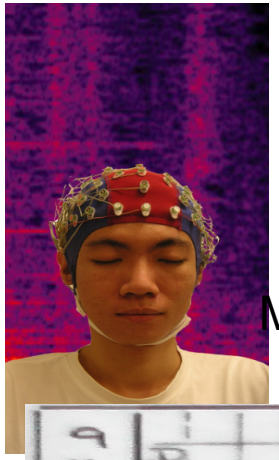
The Deep Kinship between Music and Speech

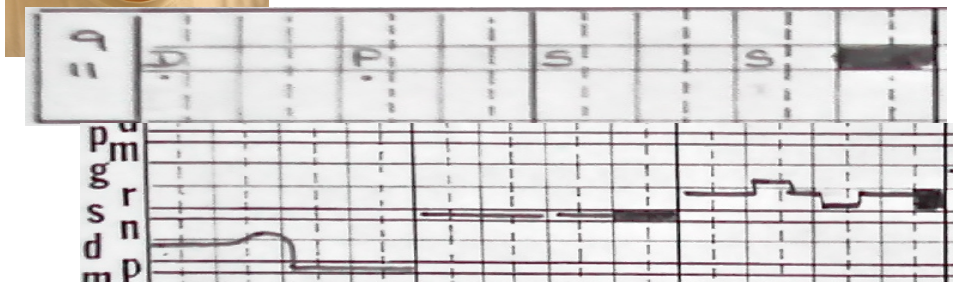
Lonce Wyse
National University of Singapore



Arts and Creativity Lab
National University of Singapore



Musical expectation

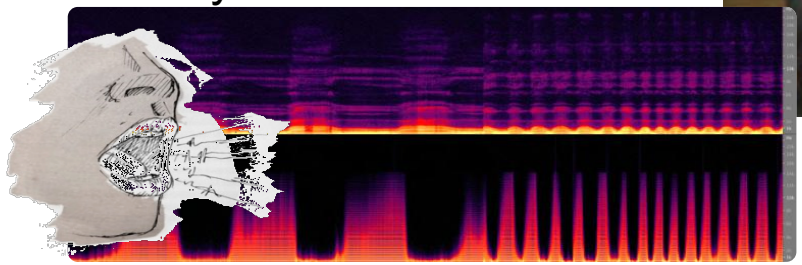


Modeling Gamakas in Carnatic Music

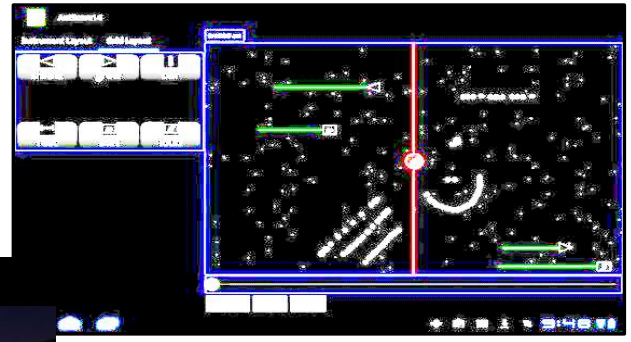
Sound Modeling

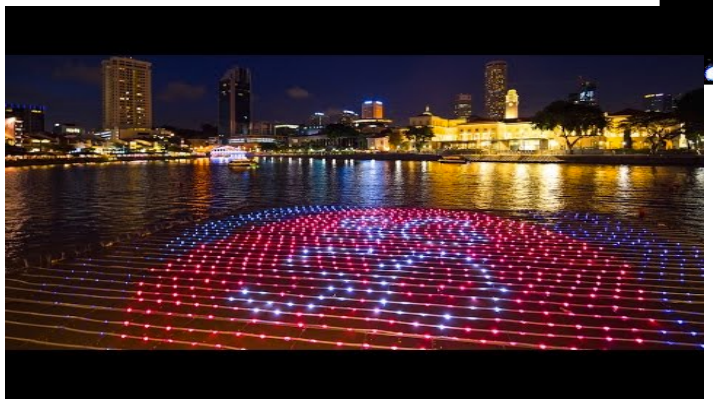
Eyes-free Games

Voice-controlled synthesis



Anticipatory Improvisation





Mobile platform for audience engagement

Vibrotactile Musical Experience for the Dea

Today

- Tell a story about two related domains
  - Review some of the commonality shared between speech and music
  - Review some musical developments of the past century
  - Reflect on “sense making” in music
  - Weave in some of my modeling work informs the narrative
  - Draw connections between sense making in music and speech.
  - **Ambitious Goal**: is find something about music that might make you think a little differently about speech.

Music: Auditory cheesecake

- Music is not adaptive, but rather an “exquisite confection crafted to tickle the sensitive spots of …. Our mental facilities.

Relationship between sound and music

Both are defining characteristics of humans

- Evolution
- Brain structures
- Hearing
- Generation
- Creativity (“generative”)
- Temporal
- Performativity

- Hierarchical
- Textuality
- Sociality
- Improvisation
- Multimodality
- Multi-channel (streams)
- Referentiality
- Combination (song, storytelling, sound poetry, Satie)

Speech

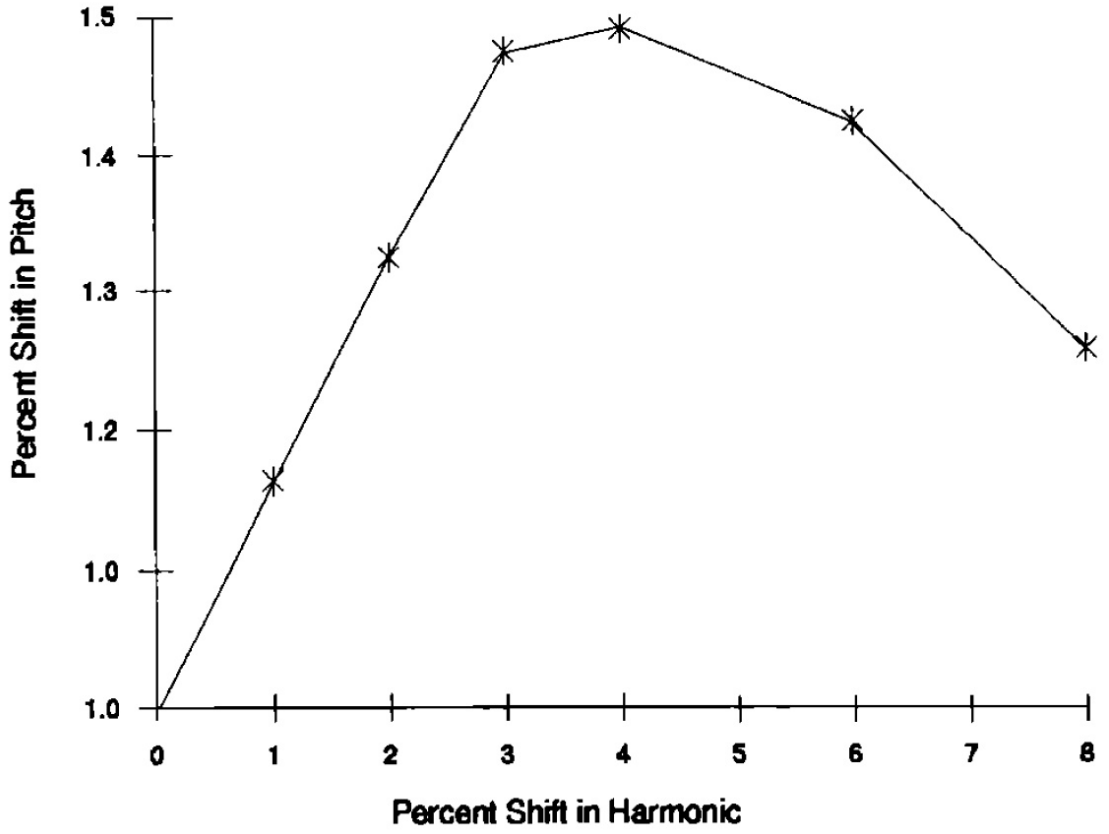
- Strings of words
  - referential semantics

- Prosody
  - Pitch
  - Amplitude
  - Rhythm
    - Timing
    - Stress pattern
    - Prediction (rhythm as a form of attention)

- Another word we use for that second group of qualities?
- How is meaning constructed in these sonic domains?

But pitch perception is not F0 identification

- Missing fundamental
- Narrow band noise, rippled noise
- Edge pitch
- Mistuned components
  - And grouping



https://sonicthings.org/stretchedKeyboard/

Expectation

- Fundamental process; survival value (prepare, disambiguate, respond more quickly)
  - Due to survival value, related to emotion (penalties & rewards in lieu of consequences)
- Well established as essential to emotional response to music
  - Leonard Meyer's Emotion and Meaning in Music (1956)
    - Without referential semantics
    - Expectations can be satisfied, violated, delayed, ambiguous thus manipulating emotion
- Language
  - Word "preactivation" facilitates comprehension
  - Difficulty of comprehension proportional to surprise in it context
  - Model building for reducing ambiguity of future events

Theories of musical meaning

- Focus on emotion
- Workhorse is EXPECTATION

IV I V { I } IV I V I

Tonic		Dominant
I	IV	V
Amin	Dmin	Emaj
Dmin	Gmin	Amaj

Theories of musical meaning

- Focus on emotion
- Workhorse is EXPECTATION

Tonic		Dominant
I	IV	V
Amin	Dmin	Emaj
Dmin	Gmin	Amaj

IV I V { I } IV I V I

But then came the 20th century ……

СВОБОДНАЯ МУЗЫКА.

Результаты примѣненія теоріи художественнаго творчества къ музыкѣ.

Въ первыхъ статьяхъ о теоріи художественнаго творчества я говорилъ объ ея могуществѣ, о томъ, что она можетъ сыграть роль магическаго жезла, ключа къ дверямъ, за которыми скрыто неизвѣданное счастье.

Сдѣлаемъ опытъ, попробуемъ проникнуть въ закрытыя палаты дворца музыки.

Естественная музыка.

Новыя возможности скрыты въ самыхъ источникахъ искусства, въ природѣ.

Мы—малые органы живой земли, клѣтки ея тѣла. Прислушаемся къ ея симфоніямъ, составляющимъ часть общаго космическаго концерта. Это—музыка природы, натуральная, свободная музыка.

Пора обратить вниманіе на естественное искусство и на законы его развитія.

Всѣ знаютъ, что шумы моря и вѣтра музыкальны, что гроза развиваетъ дивную симфонію, а музыка птицъ даже получила большое распространеніе въ обиходѣ обывателя.

Главнѣйшія изъ положеній о свободной музыкѣ уже опубликованы мною въ видѣ конспекта „Свободная музыка". С.-Пб. 1909 г. 12, 7 стр.

Nikolai Kulbin (1910) Free Music

100 years of musical innovation

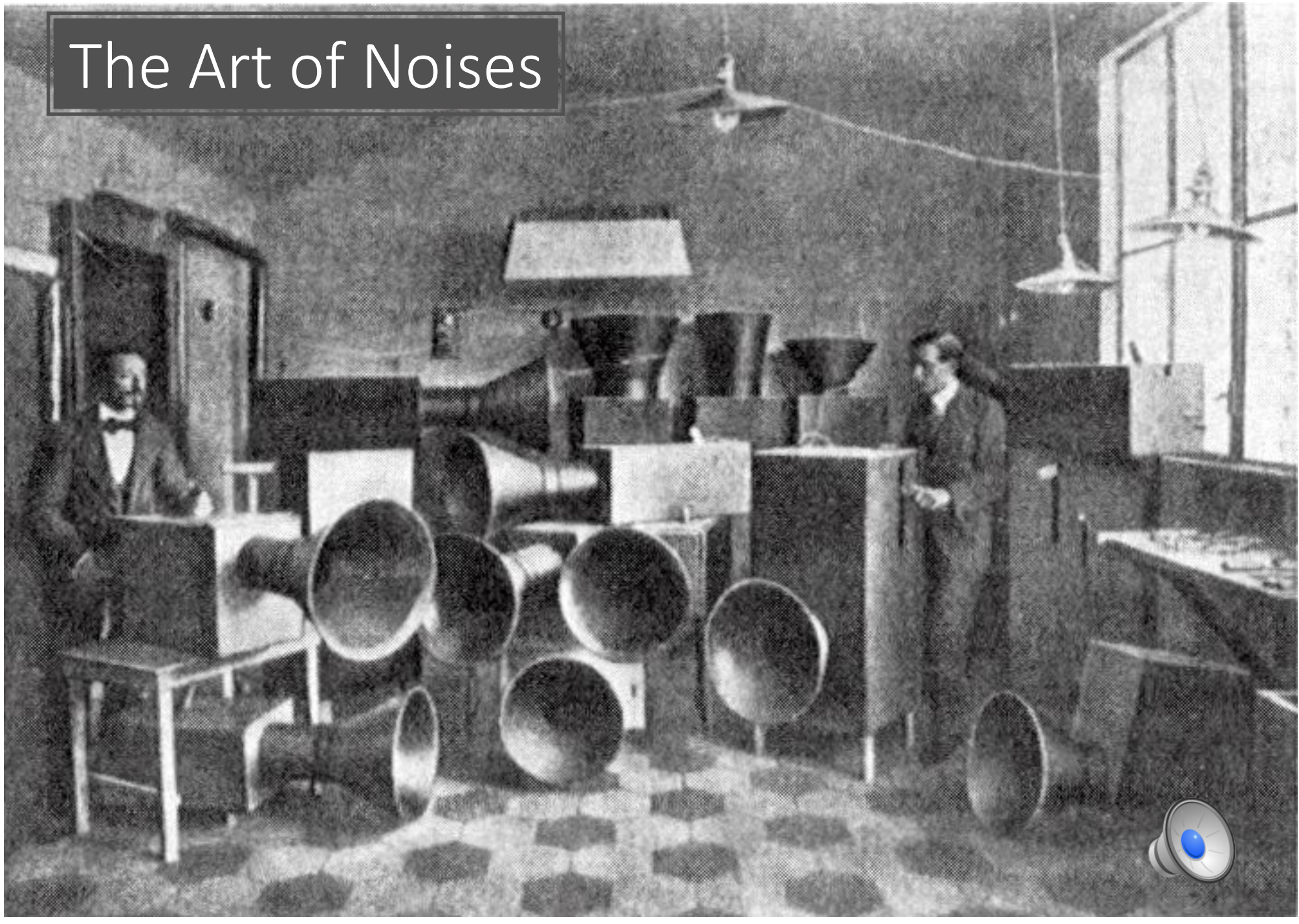
The symphony of the cosmic concert is the music of nature - the natural "free music". …

…everybody knows that the noises of the sea, wind, thunderstorm, makes a symphony as well as the music of birds - but right now, people exploit the music of nature according to the old laws - …

The Art of Noises

Beethoven and Wagner have stirred our nerves and hearts for many years. Now we have had enough of them, and we delight much more in combining in out thoughts the **noises of trams, of automobile engines, of carriages and brawling crowds**, than in hearing again the “Eroica” or the “Pastorale”.

(Art of Noises, 1913)



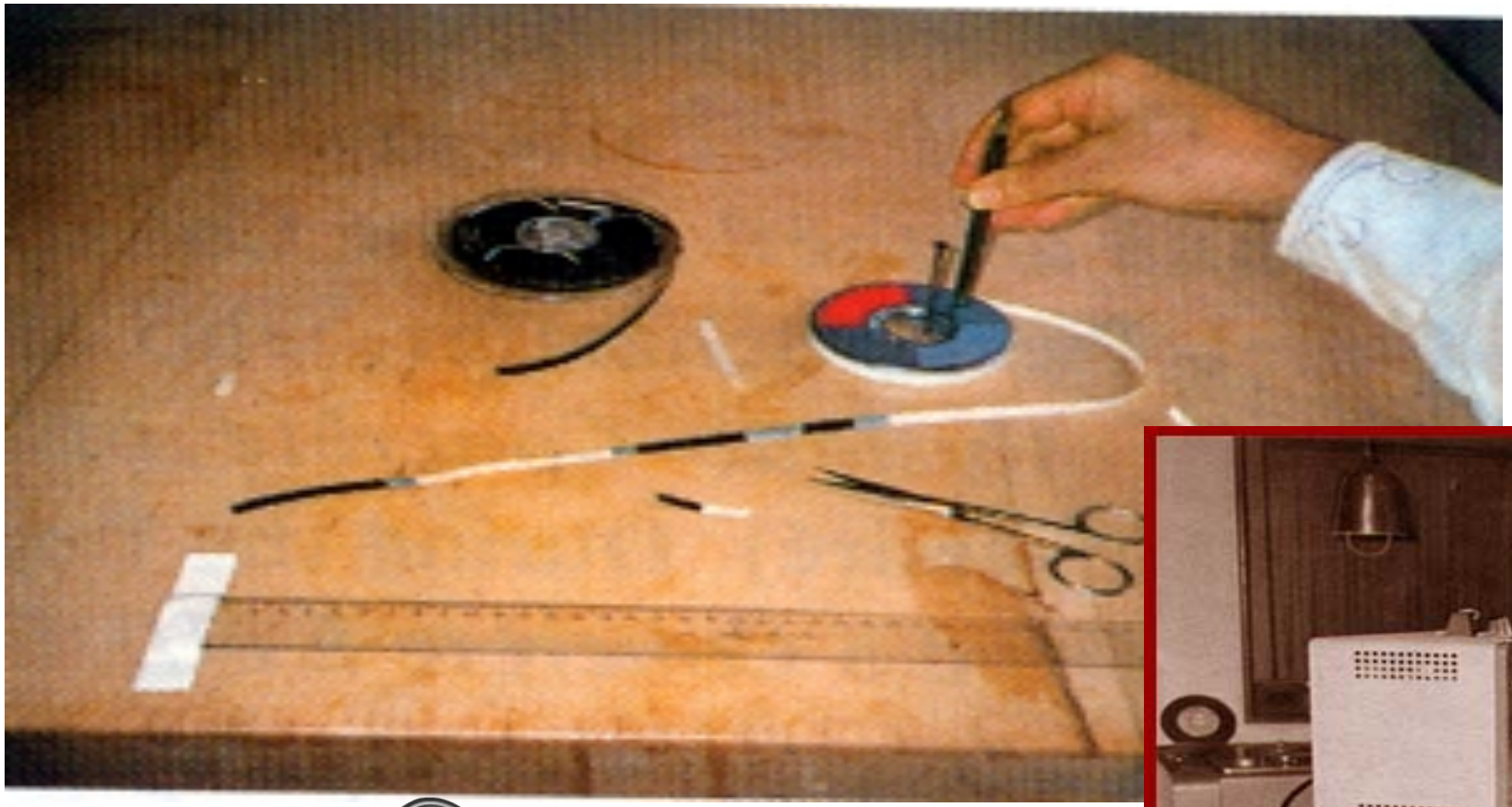
Musique Concrète

Pierre Schaeffer at his “chromatic phonogène” in 1953.

Ecoute reduite

Etude aux Chemins de Fer (1948)

Stockhausen



Electronic Music,
Spatial dimension
……

Gesang der jünglinge

New Music

- Extended vocal and instrumental techniques
- Recorded sound, referentiality
- Electronic sound (“sourceless”)
- Indeterminacy
- Notation (animated, graphic, real-time)
- All sound
- Interfaces
- Mediated communication

Not even talking about “sound art”!

- Big mistake to think of music as sequences and groupings of notes.

Aspects of sound not in the sample stream

- Sound is spatial (after it leaves the source)
  - Goes around corners
  - Can surround us
  - Is part of an “orienting system” (compared to visual)
- Evokes place
  - Soundscapes
- Is tactile

Reioji Ikeda

Aspects of sound not in the sample stream

- Bears a different relationship to objects than names or images.
  - Source “bonding”
  - Sound generally comes from the interaction of multiple objects (a “source” and an “exciter”)
  - Indicative of *events* as much as *objects*

Margaret Boden

“A creative idea is one which is novel, surprising, and valuable (interesting, useful, ...”

- P-creativity - novel for an individual
- H-Creativity − novel historically

Margaret Boden

"A creative idea is one which is novel, surprising, and valuable (interesting, useful, ..."

- P-creativity - novel for an individual
- H-Creativity – novel historically

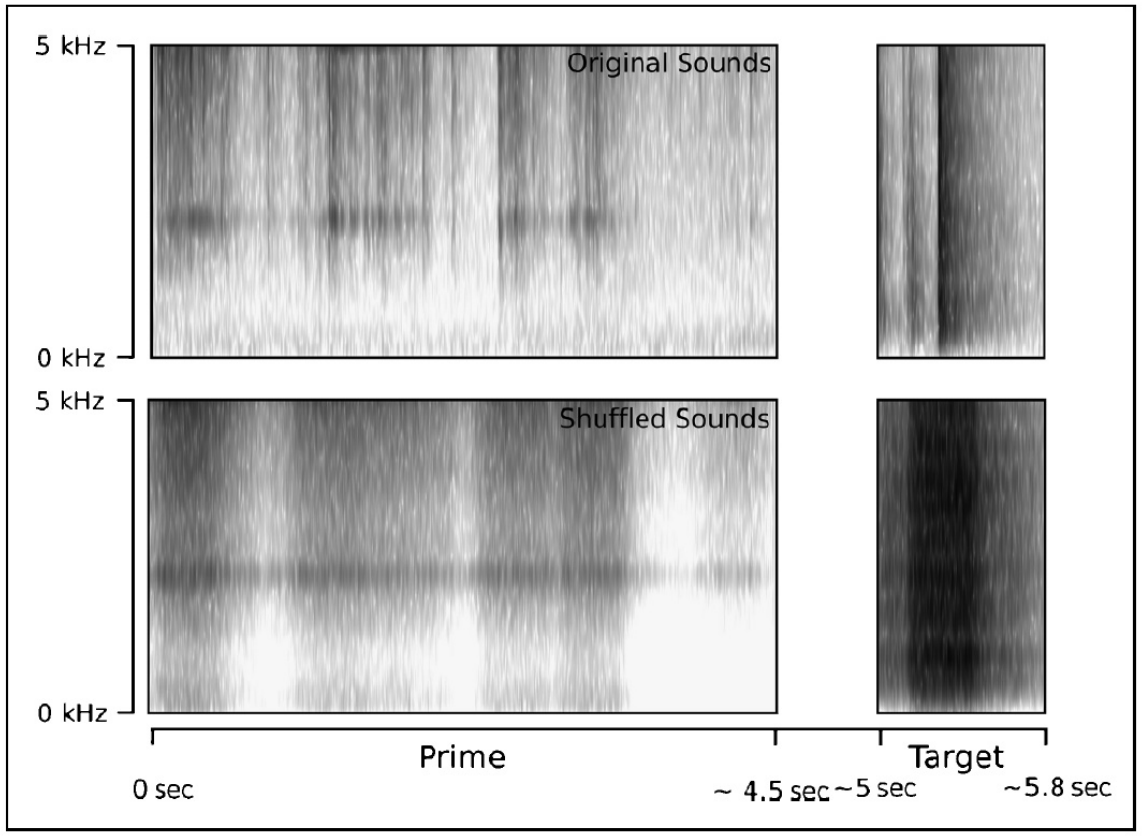
Do speech and music have different fluidity at these different time scales?

Expectation

- Priming in speech
  - Words prime words,
  - cross domain priming

- In sound
  - Previous research indicated exact repetitions produced faster and more accurate behavioral responses than different sounds. But perceptual and conceptual are conflated. Attempts to tease these two apart behaviorally were inconclusive.
  - Cross modal experiments (eg pictures priming sounds) suggest conceptual priming (but are difficult to interpret)

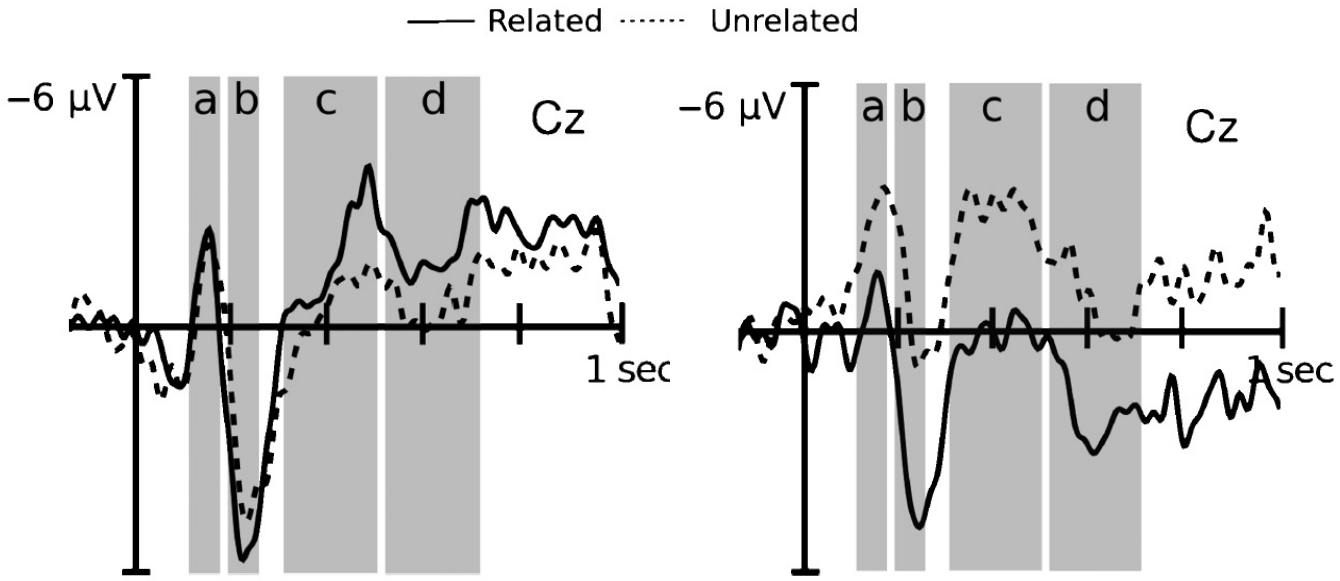
Sound Priming Sound?

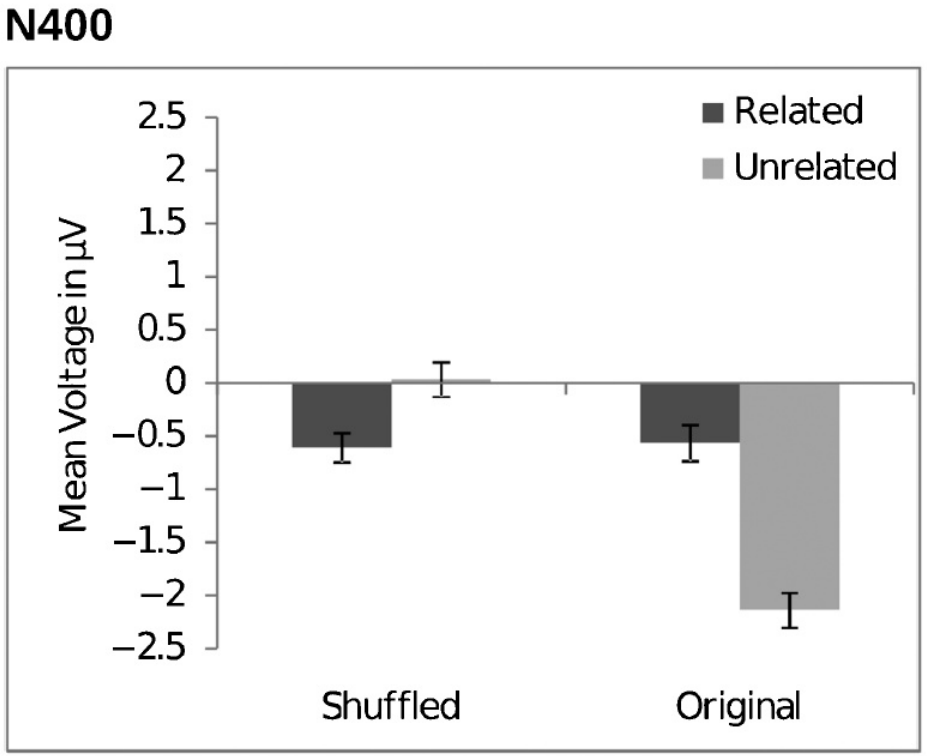


- Disentangle perceptual from conceptual priming
- Shuffle phases so mangled sounds have same frequency content, but are unrecognizable.

Schirmer, A., Soh, Y. H., Penney, T. B., & Wyse, L. (2011). Perceptual and conceptual priming of environmental sounds. Journal of cognitive neuroscience, 23(11), 3241-3253.

Results





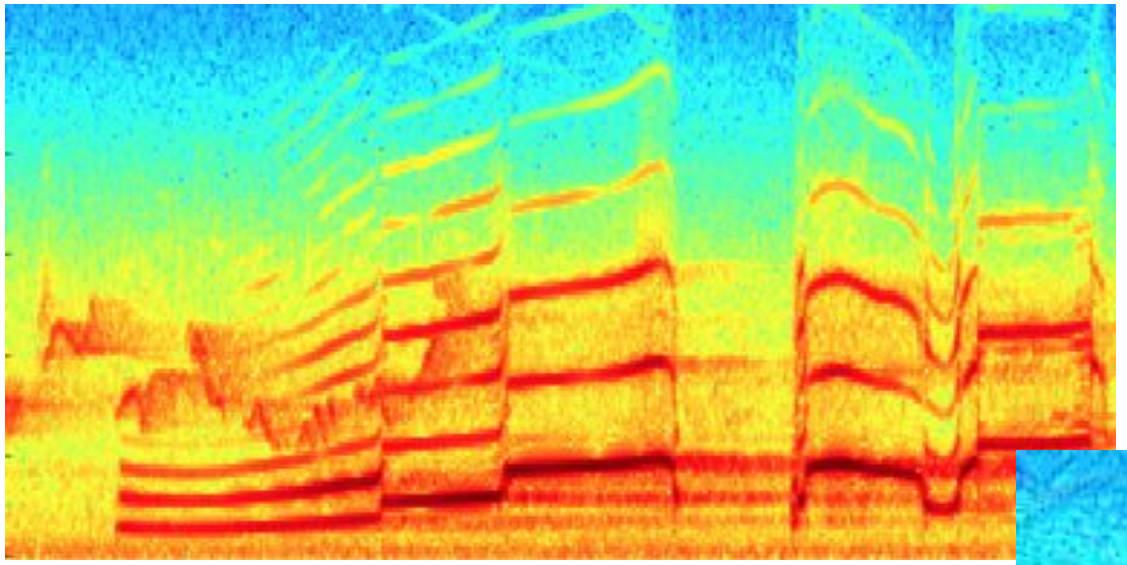
- Sonic context effects for both perceptual and conceptual aspects
- Specific N400 priming effect that suggests within-modal (sound) conceptual priming effect
  - (can't completely rule out verbalization, but doesn't appear to be present)

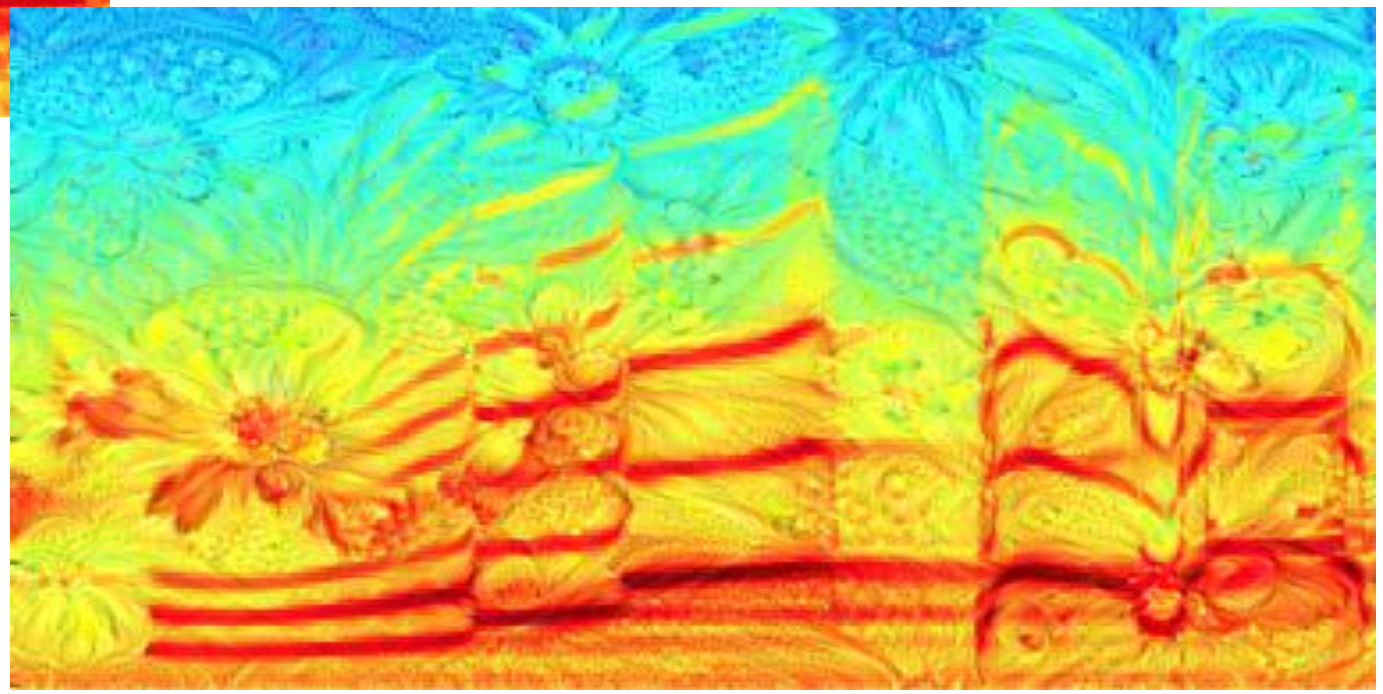
So new sounds for music, then!

- Representations
- Synthesis algorithms
- Sound space navigation
- Physical interaction

- Goals
  - Complexity of natural sounds (“realism”)
  - Real-time (not just a matter of speed)

Style Transfer? Spectrograms are 2D images





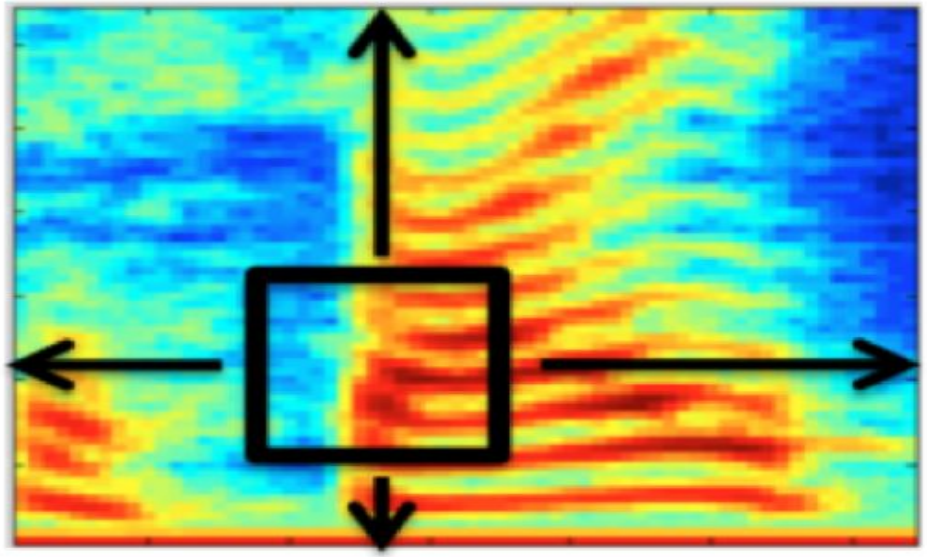
Issues?

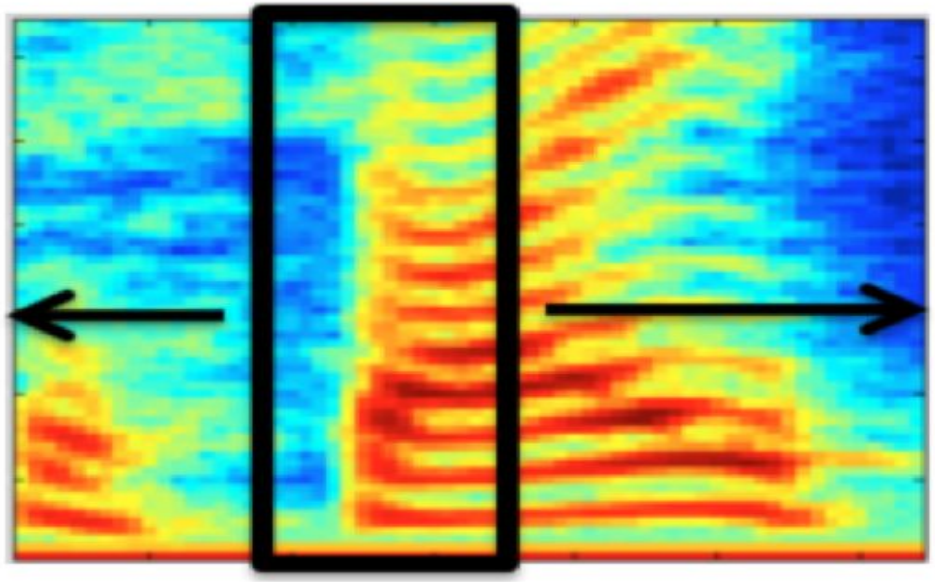
Representation

Time

Parameterization

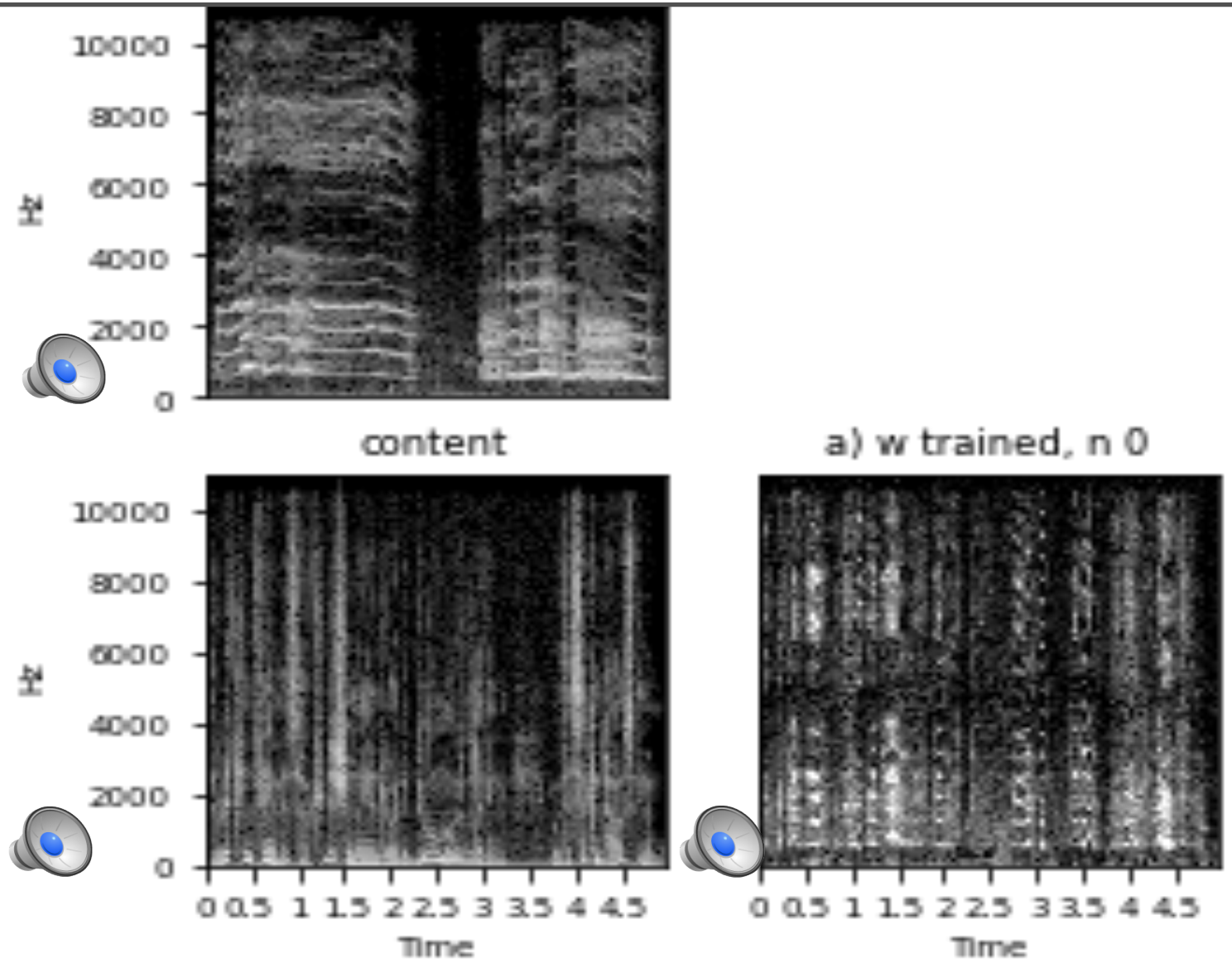
2D vs 1D convolution





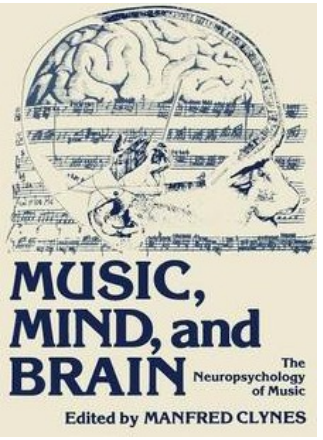
Anand, N. and Verma, P. (2016)

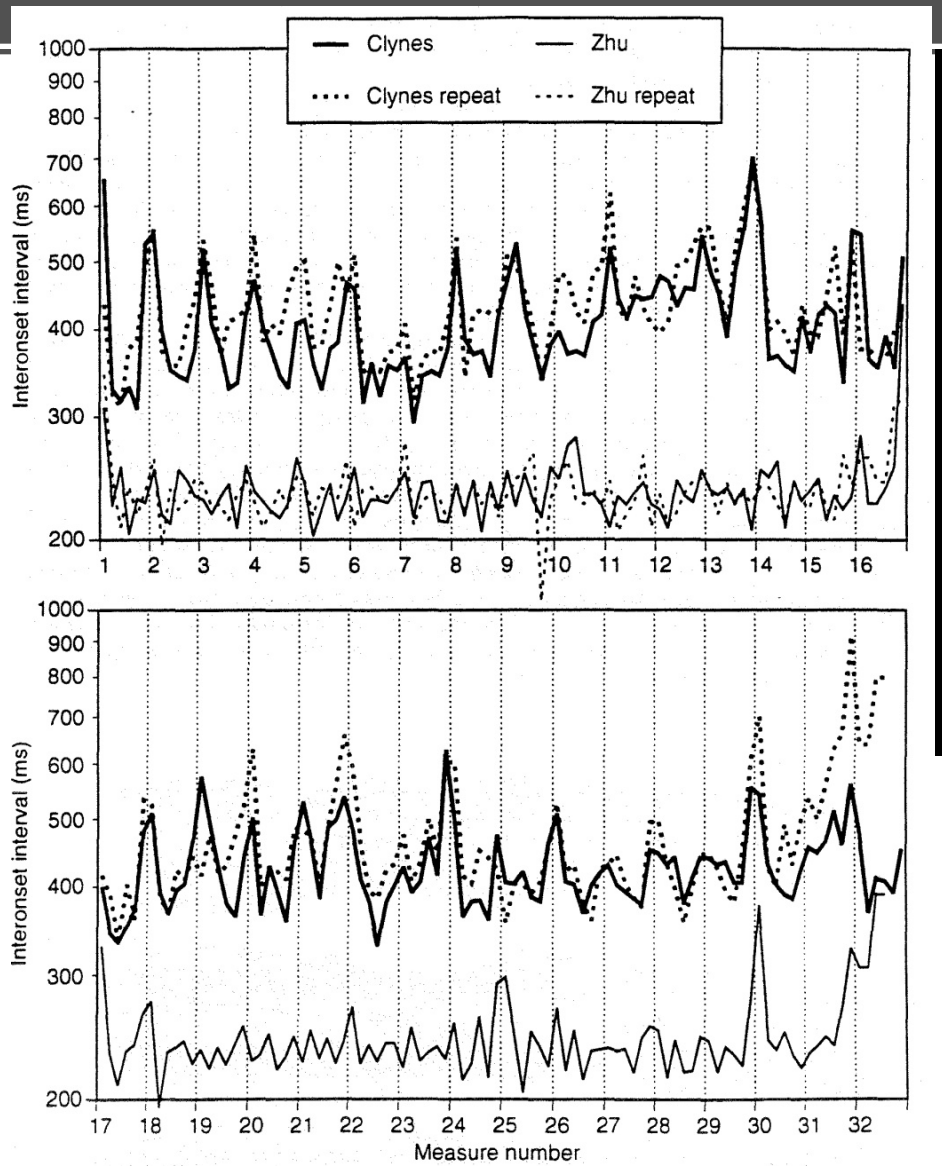
Using multi-layer audio-trained networks



But what is content and style in music?

The Composer's Pulse, Manfred Clynes (1983 …)





Expressive timing profiles of Goldberg Variation 6, played by Manfred Clynes and Xiao-Mei Zhu

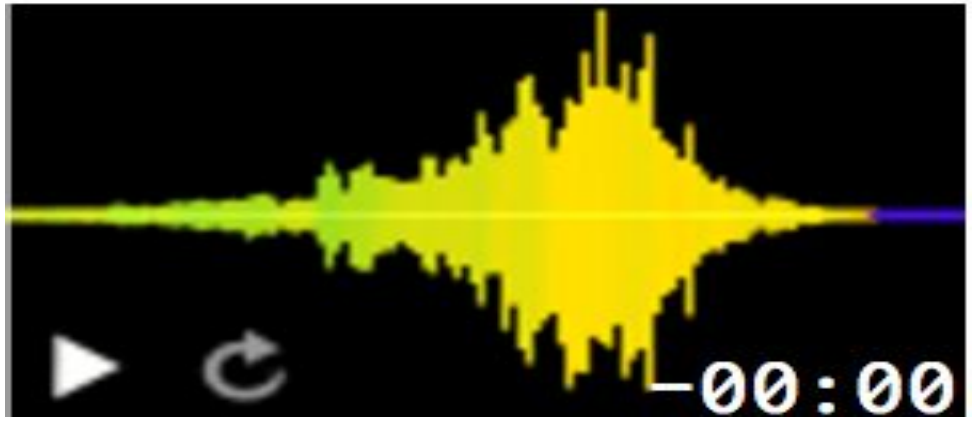
Style & content

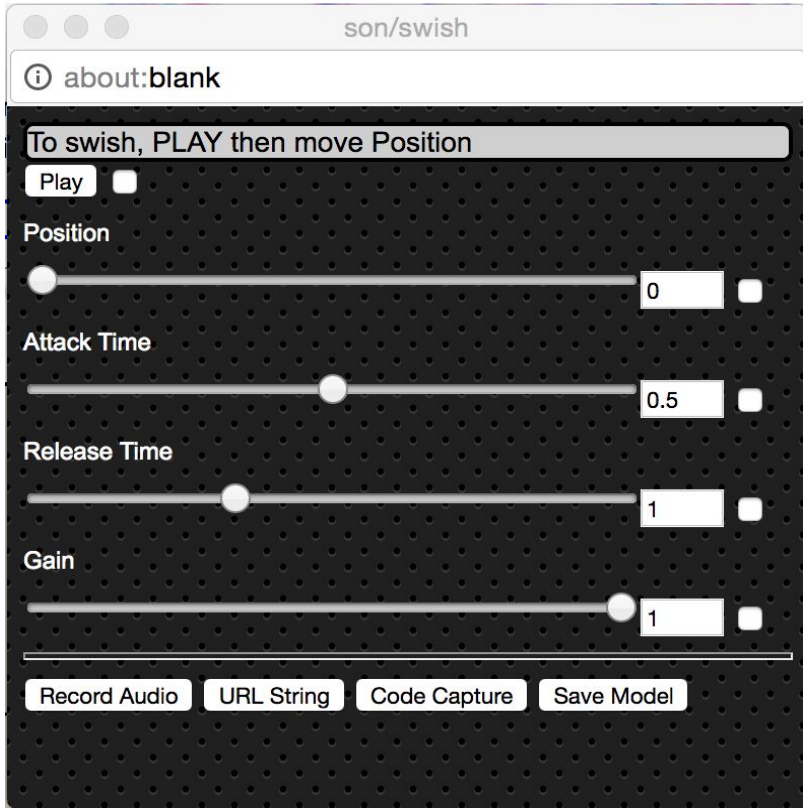
- Speech
  - Content − words (which stand in for their meaning)
  - Style - prosodic elements
    - Pitch, rhythm, amplitude, timbre

- Music
  - ?????
  - Arnold Schoenberg's timbre-structure "tone poems"
  - What are the "units"

Objective: Data-driven sound modeling

- Provide sound examples *and desired interaction*
- Get parameterized synthesis model





Lyrebird.ai – train on your voice,
Type text to synthesize

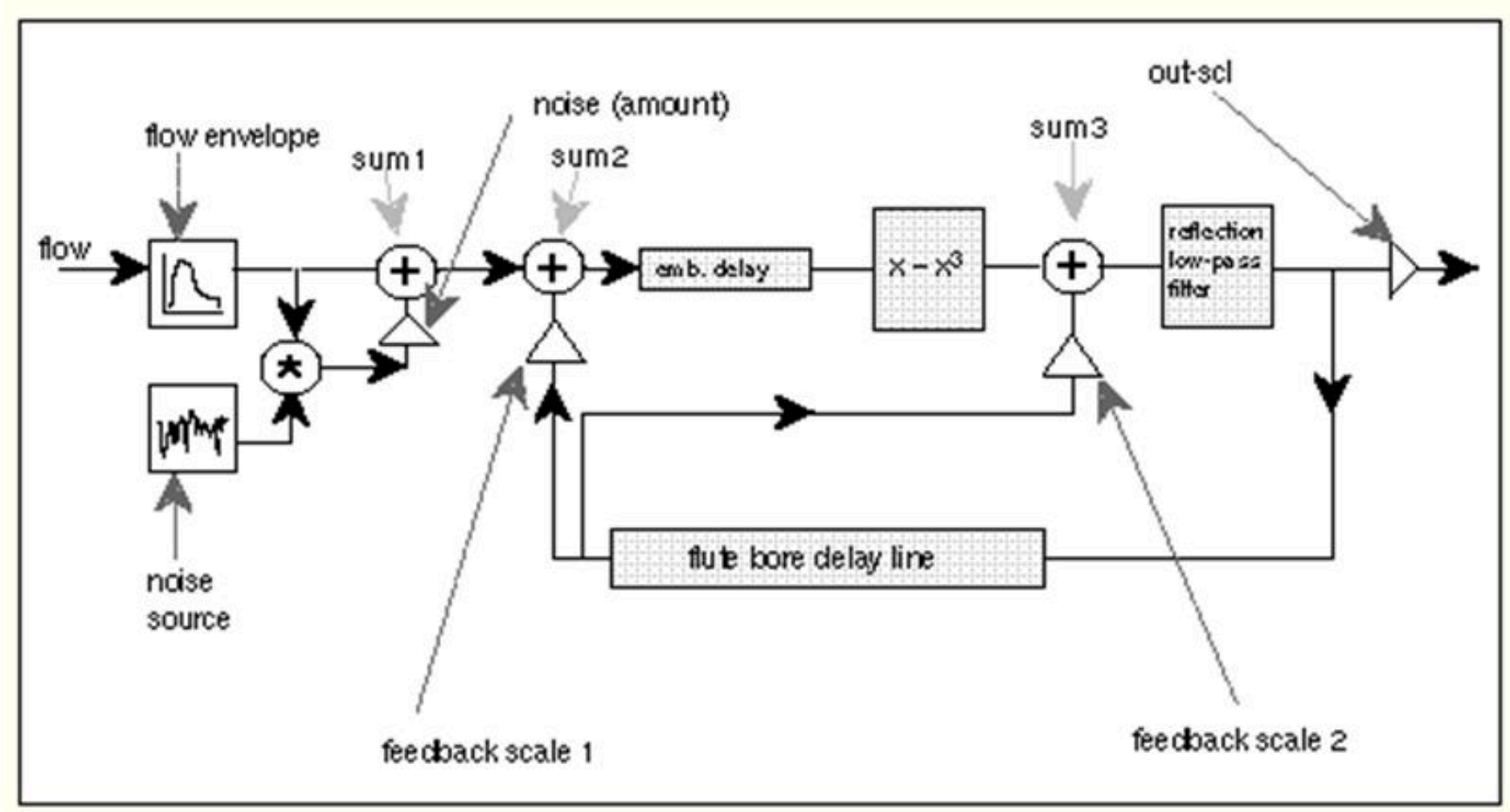
Traditional Sound Modeling similar to speech

- Physical modeling
- Acoustic modeling
- Concatenative synthesis

- However, for musical models/instruments,
  - Not just sonic space, but Control
    - Arbitrary (even configurable), many different kinds, real time,
    - Dislocation of causality
  - Want different models for different classes of sounds
  - Want different models for the *same* class of sounds
  - Will never have all the models we need for musical purposes (“composed instruments”)

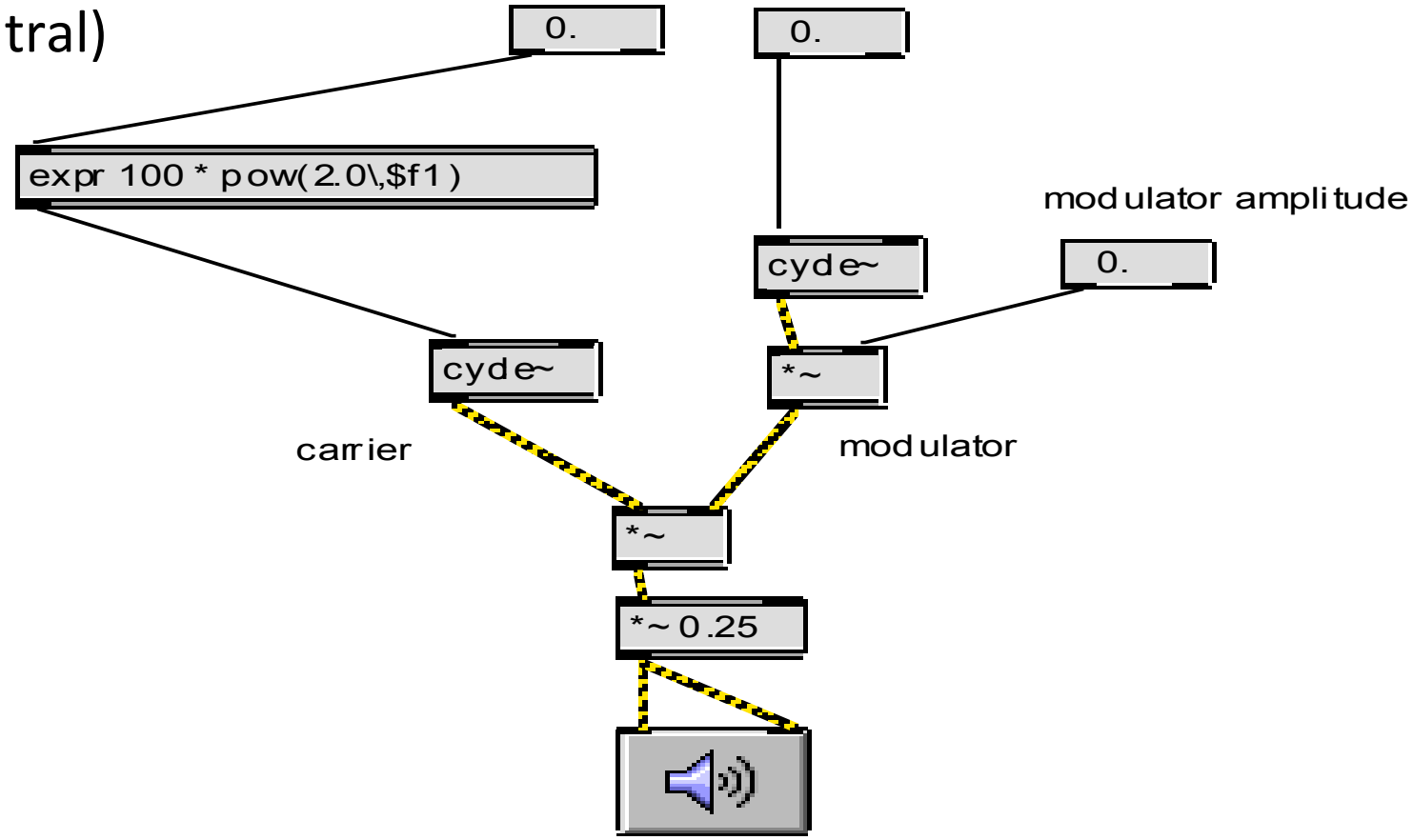
Waveguides

Physical Model of a Flute:

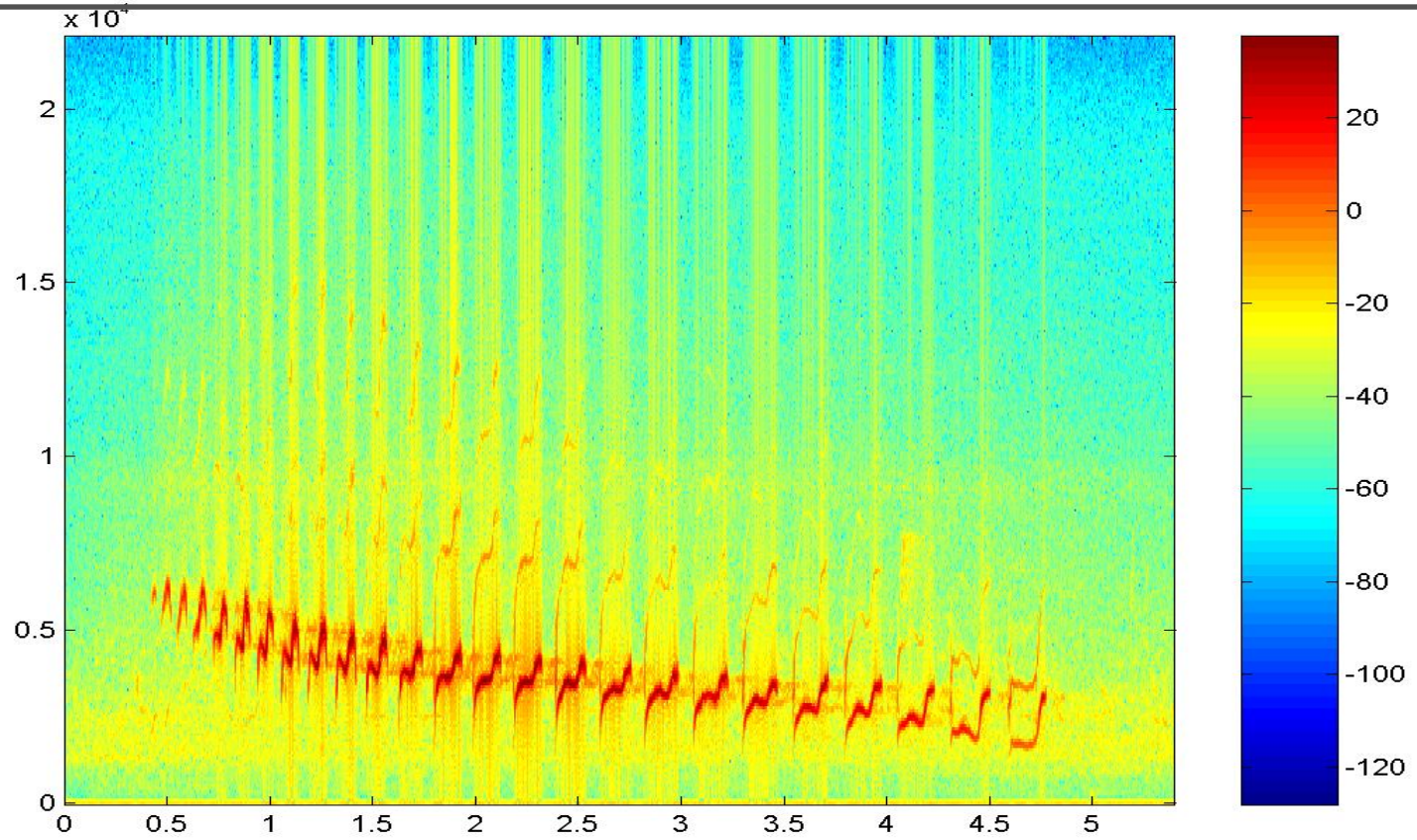


Acoustic Modeling

- Modular Synths
  - Max/MSP from Cycling ‘74
  - Tassman – (free at Harmony Central)
  - Synthedit – (free at Harmony Central)

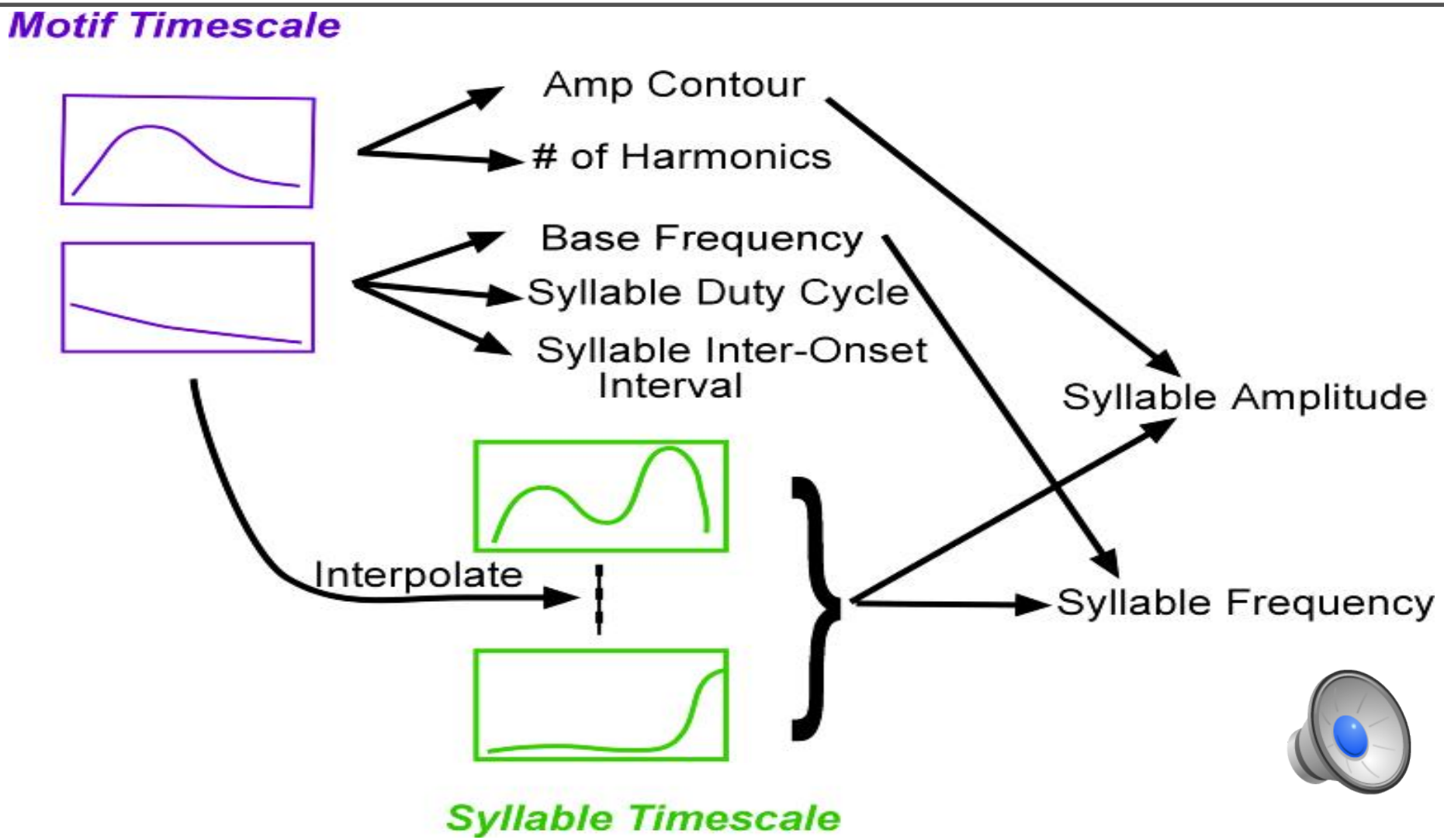


Canyon Wren Spectrogram

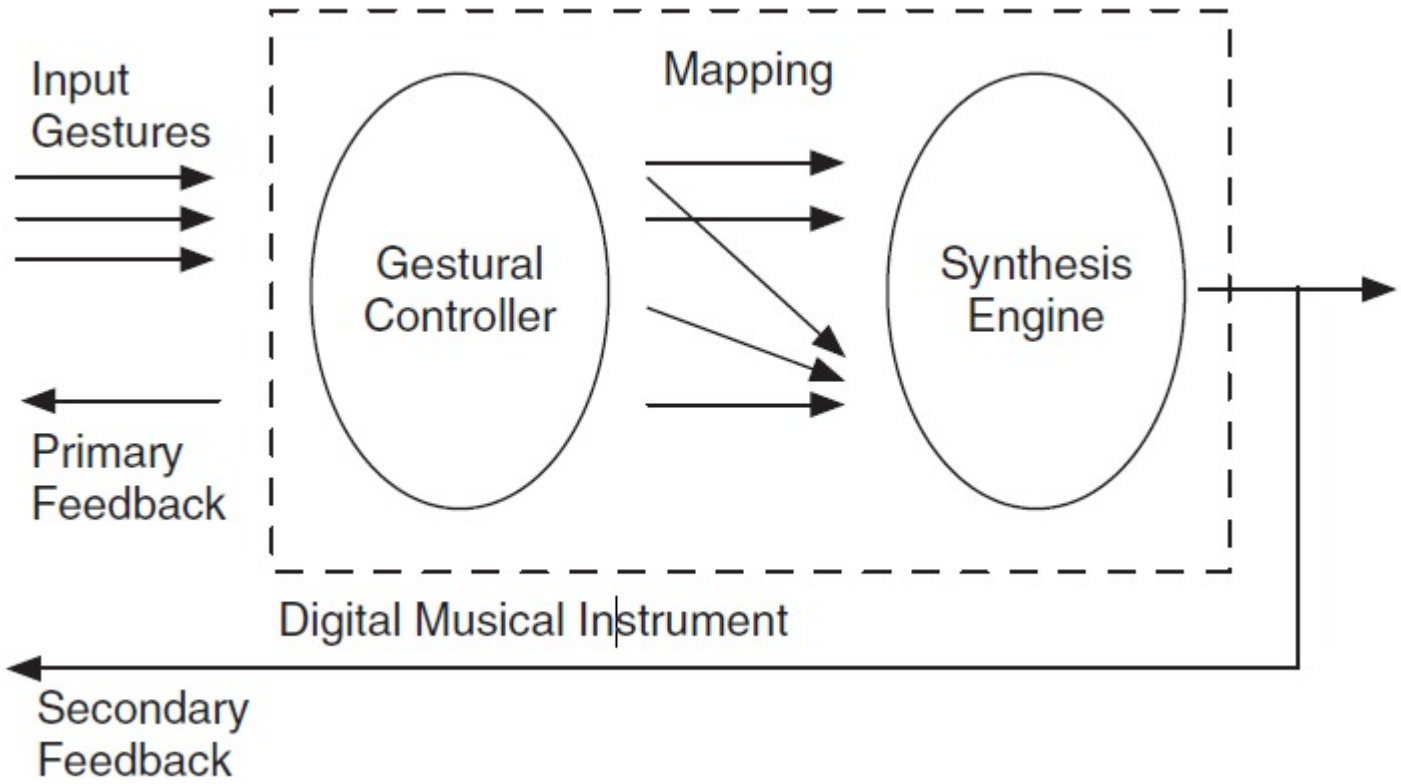


- Central pitch contour
- Harmonic contour
- Chirp frequency contour & “meta” contour

Model Structure



Gesture mapping



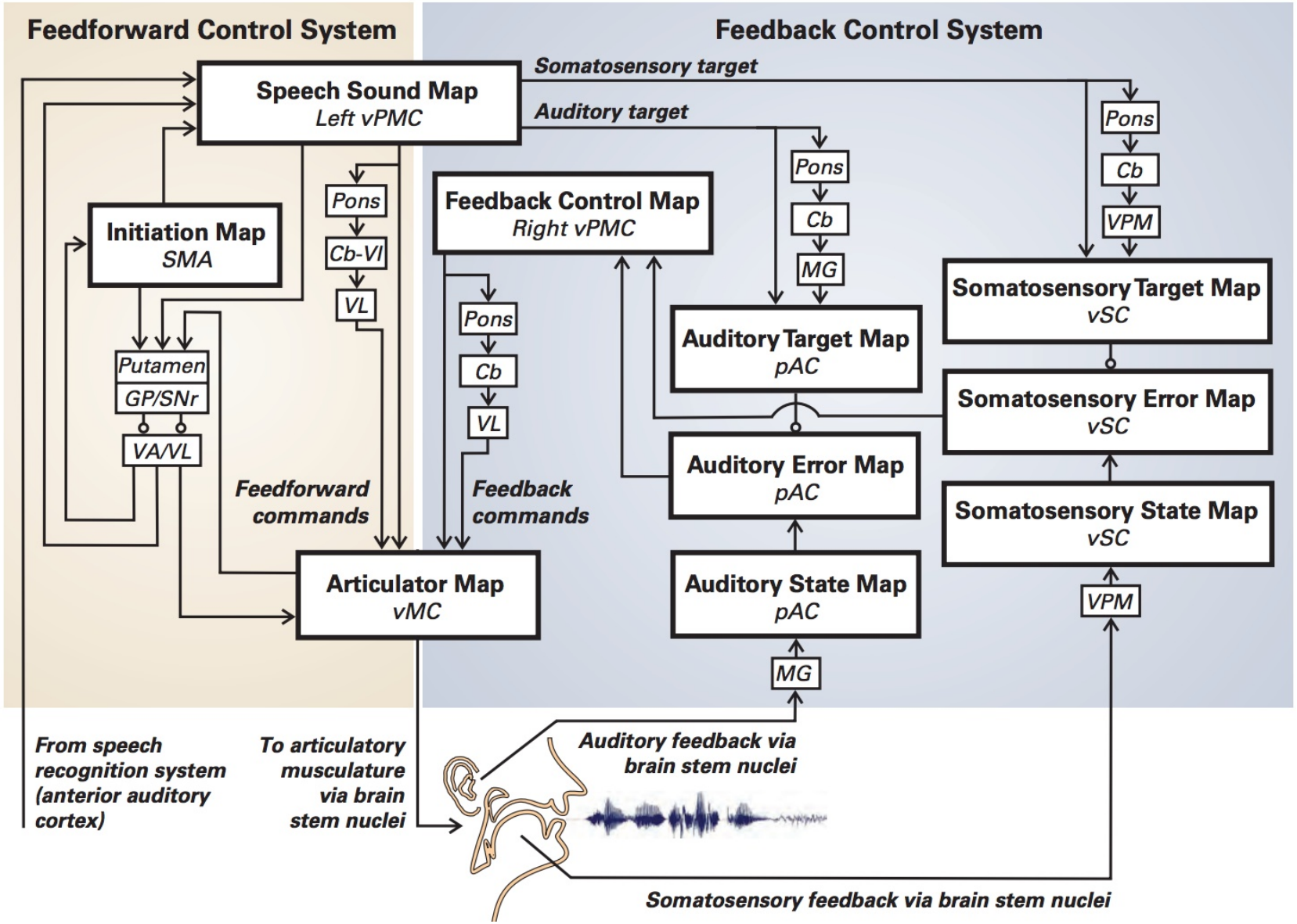
- Pitchification

- Wekinator

Miranda & Wanderly (2006)

Fiebrink, R. A. (2011). Real-time human interaction with supervised learning algorithms for music composition and performance. Princeton University.

DIVA and its decedents



Neural speech model.

Articulatory synthesizer.

Learns by babbling.

Frank Guenther, Boston University

Neural Control of Speech (2016)
MIT Press

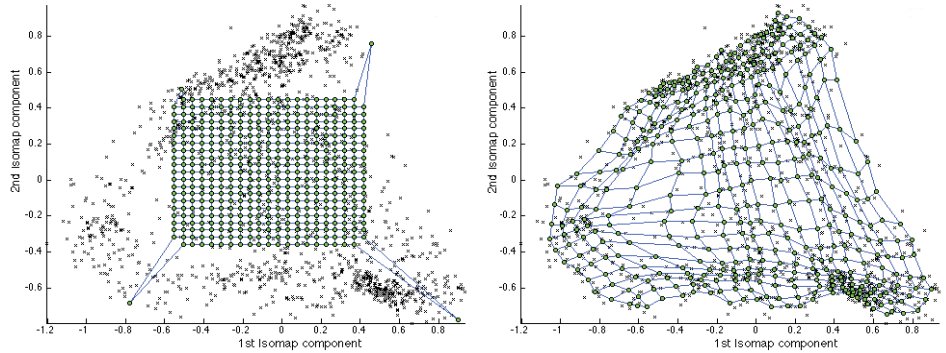
So many controls, so few hands!

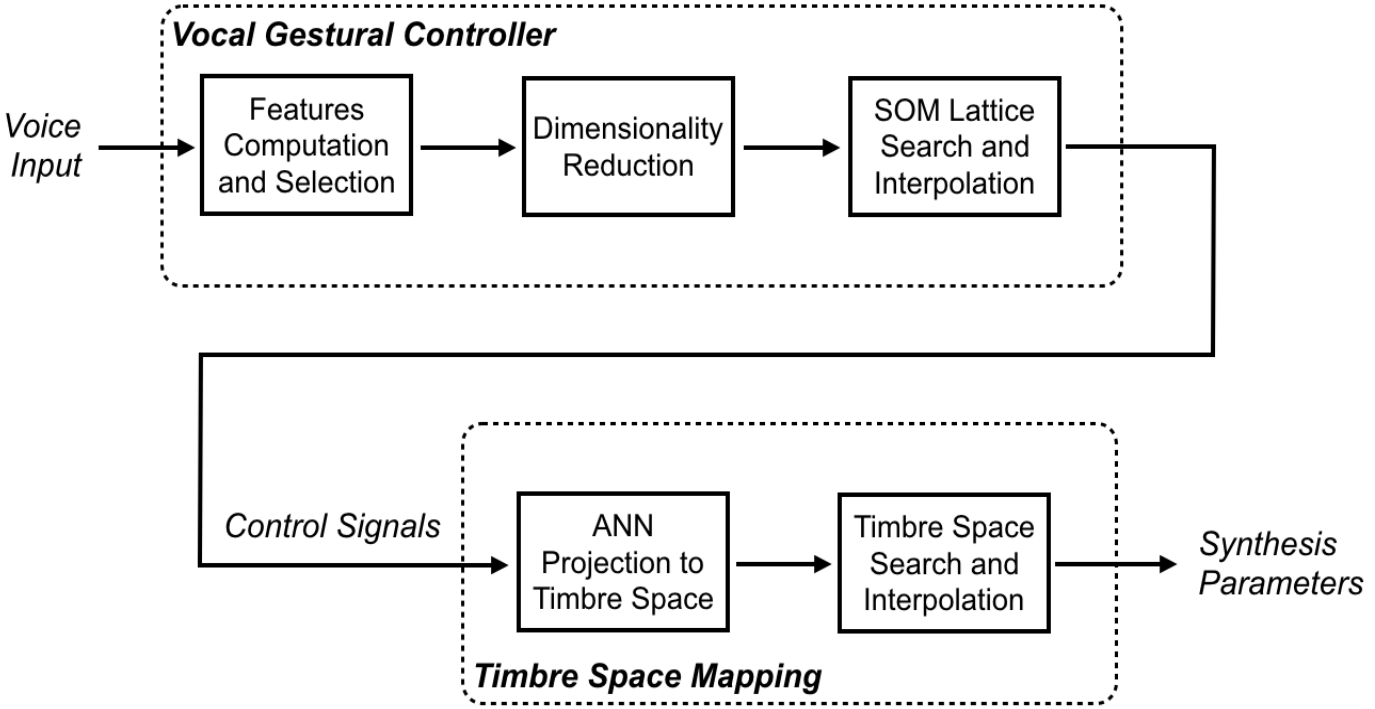
VCI4DMI

S. Fasciani and L. Wyse. "Vocal control of sound synthesis personalized by unsupervised machine listening and learning," Computer Music Journal, 24:1, 2018

Voice mapping

- User provides vocal sounds to be used
  - (voice and gesture customizable)
- Large set of features extracted, most robust chosen
  - Noisy features are discarded
  - Compute intrinsic dimensionality
  - SOM to cover space
- User provides synth (with params)
  - Learn gestures-> sonic features
  - Map sonic features->synth params





In action

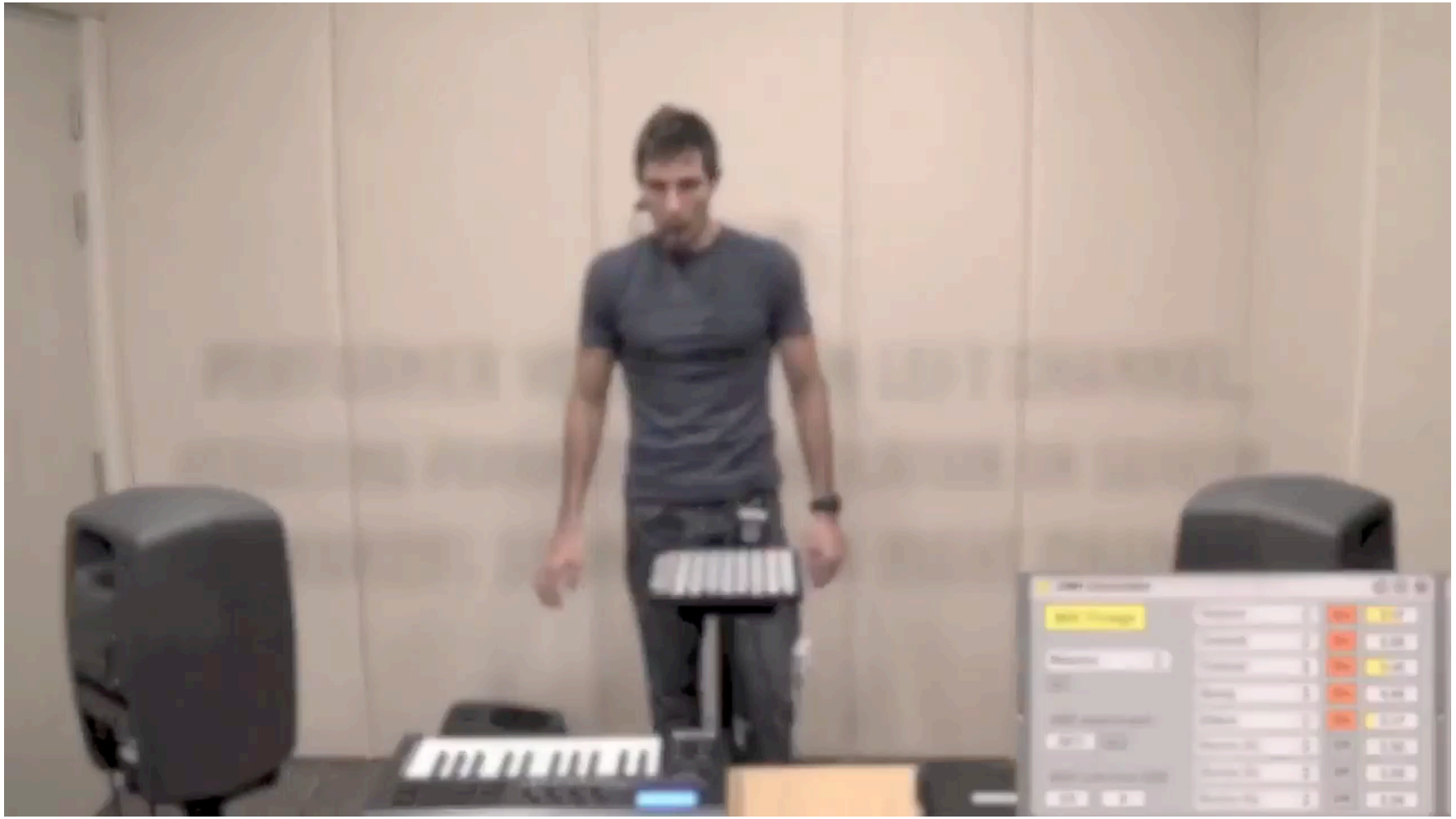
- Hands+voice
- Voice only
- Exposed voice

In action

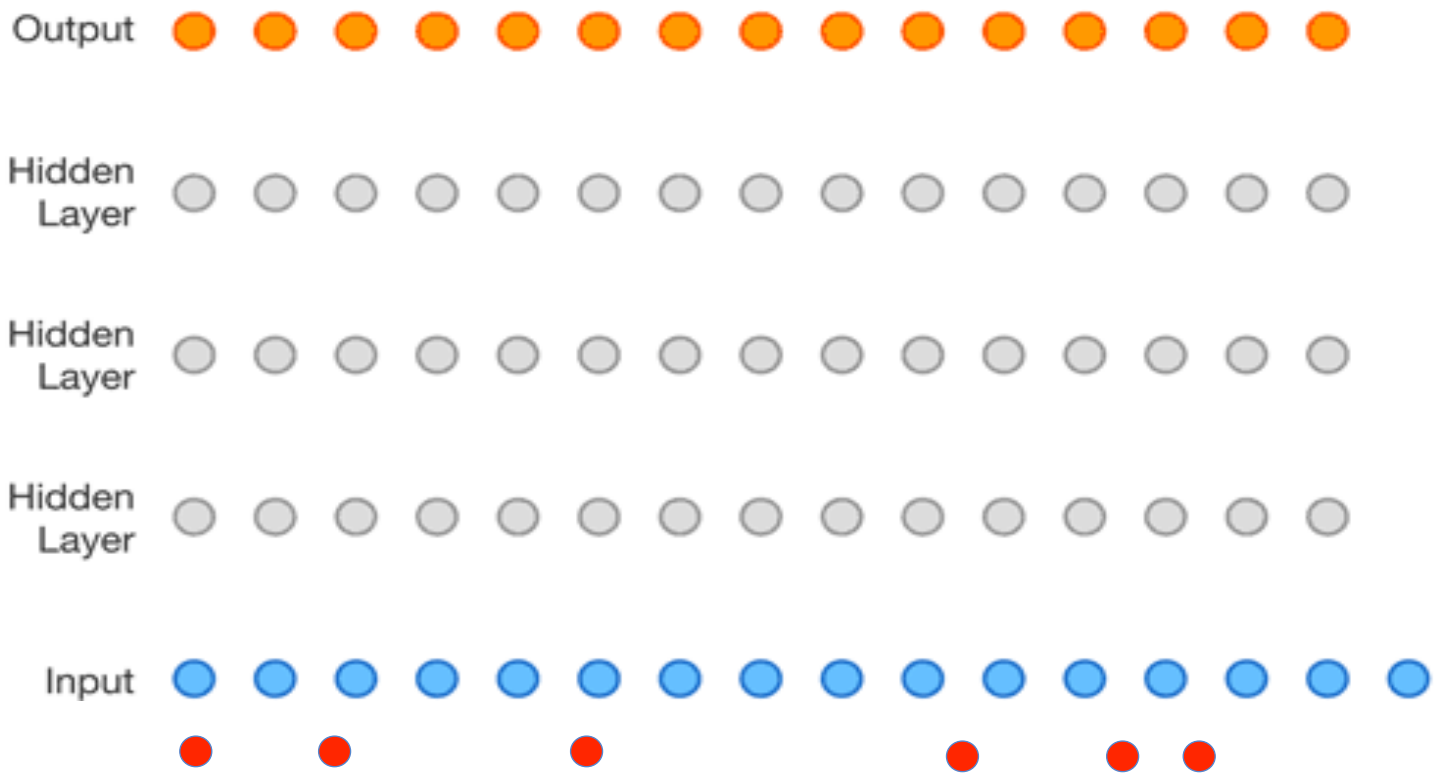
- Hands+voice
- Voice only
- Exposed voice

In action

- Hands+voice
- Voice only
- Exposed voice



WaveNet



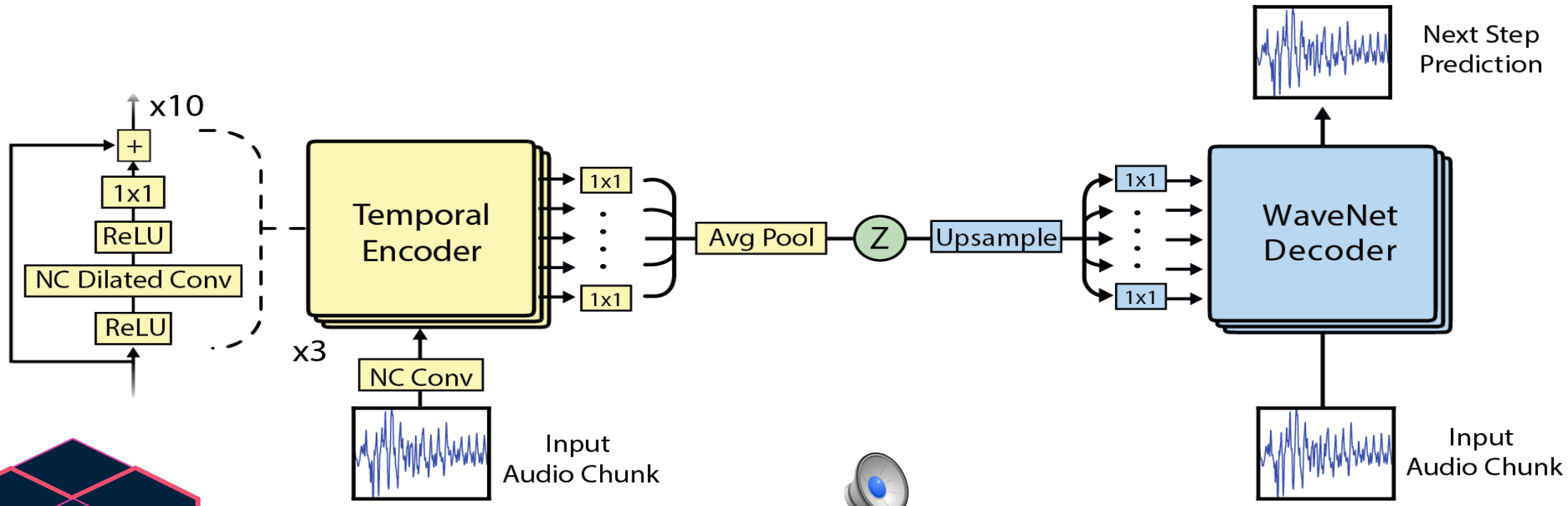
Additional “conditioning” input

[Piano, based on samples only]

Van den Oord et al. (2016)

[Conditioned on speaker only]
[Conditioned on speaker & phonemes]

NSynth

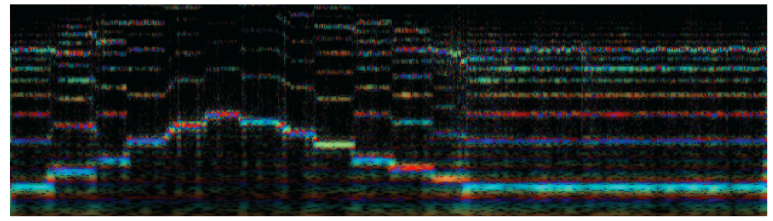


magenta

Plucked bass

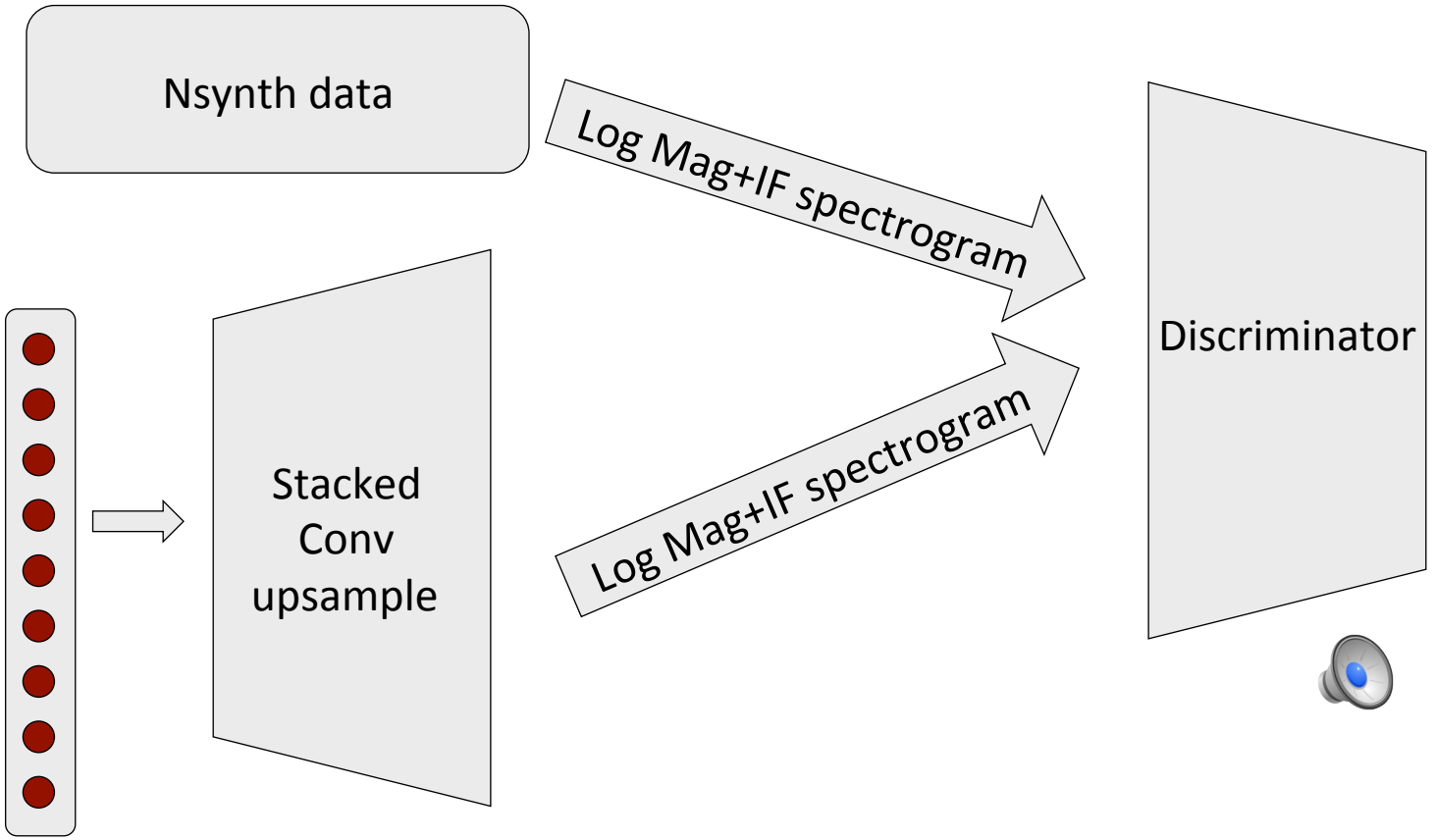
Flute

Bass/flute combination



https://magenta.tensorflow.org/nsynth

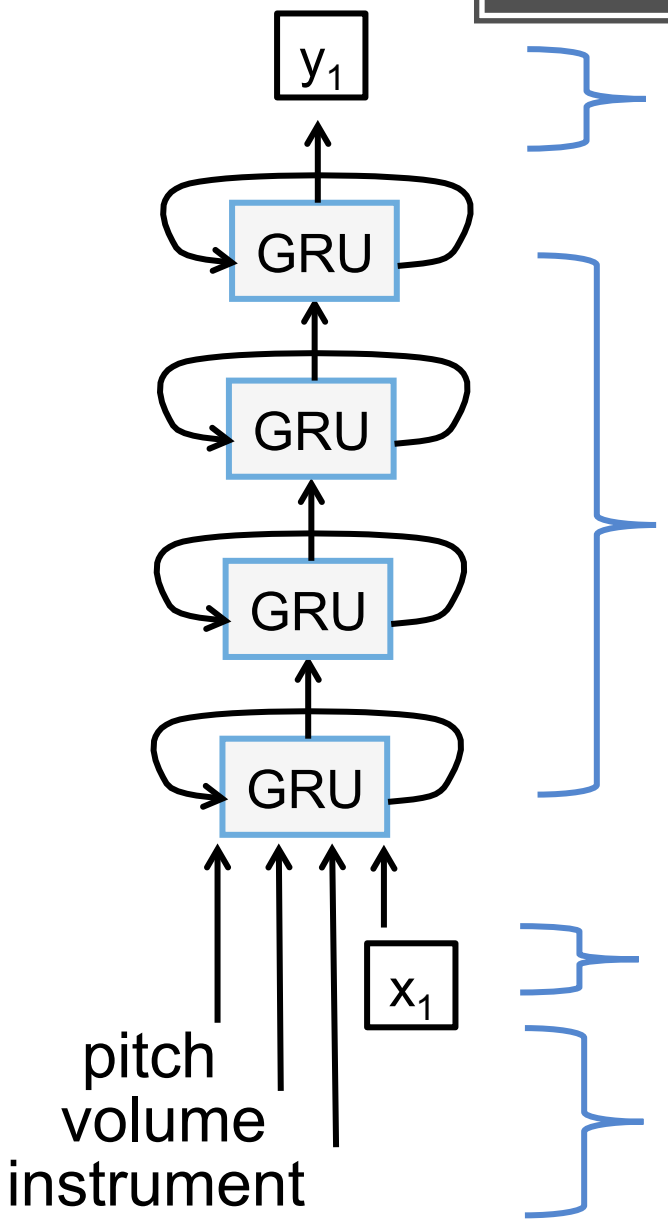
GANSynth



256 noise + one-hot pitch label

Engel, J., Agrawal, K. K., Chen, S., Gulrajani, I., Donahue, C., & Roberts, A. (2019). Gansynth: Adversarial neural audio synthesis. arXiv preprint arXiv:1902.08710.

RNN Architecture



One-hot, 256

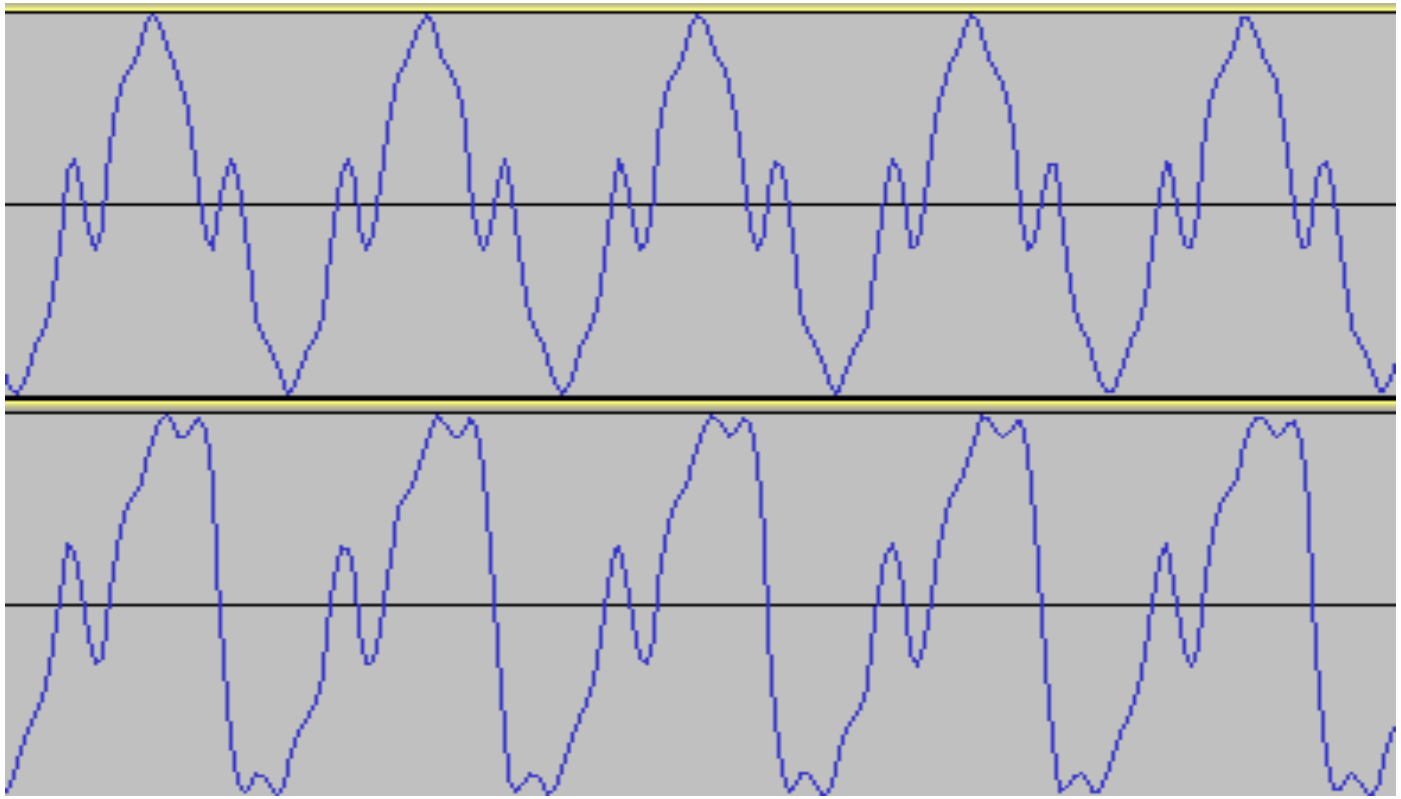
4 layers, 40 units each

Mu-law, real in [0,1]

real in [0,1]

Wyse, L. (2018) **Real-valued parametric condition of an RNN for interactive sound synthesis.** In *Proceedings of the Proceedings of the 6th International Workshop on Musical Metacreation, ACM Conference on Computational Creativity*. Salamanca, Spain, June, 2018.

1st: Synthetic signals

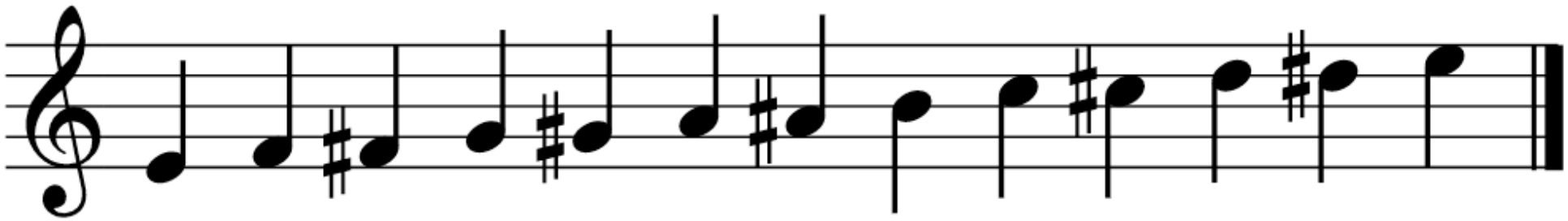


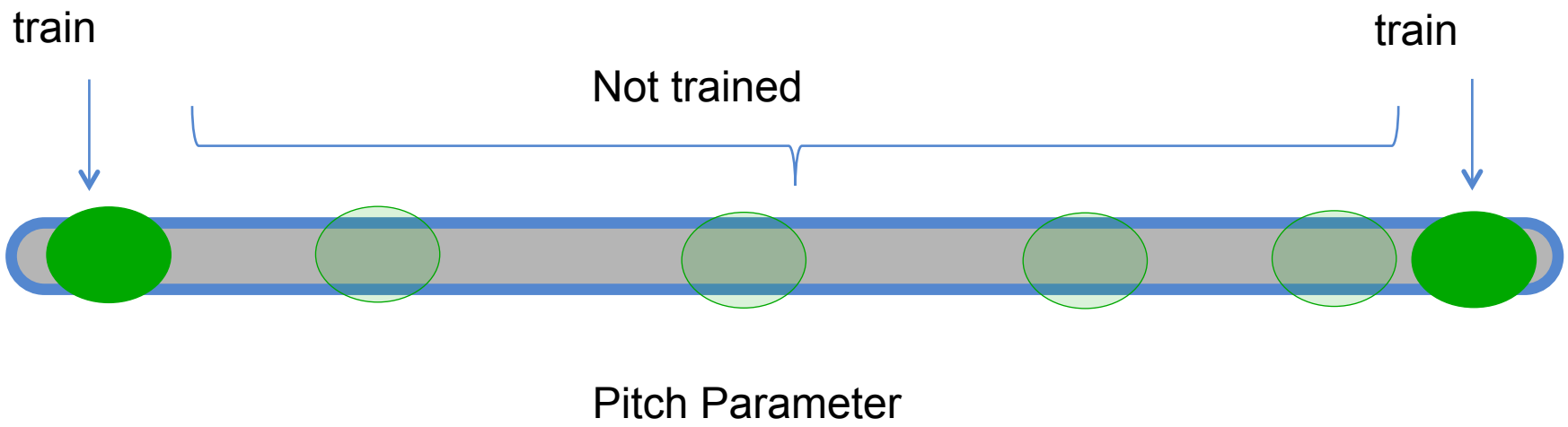
- Odd harmonics (“clarinet”)

- Even harmonics

Waveform looks the same at all pitch values

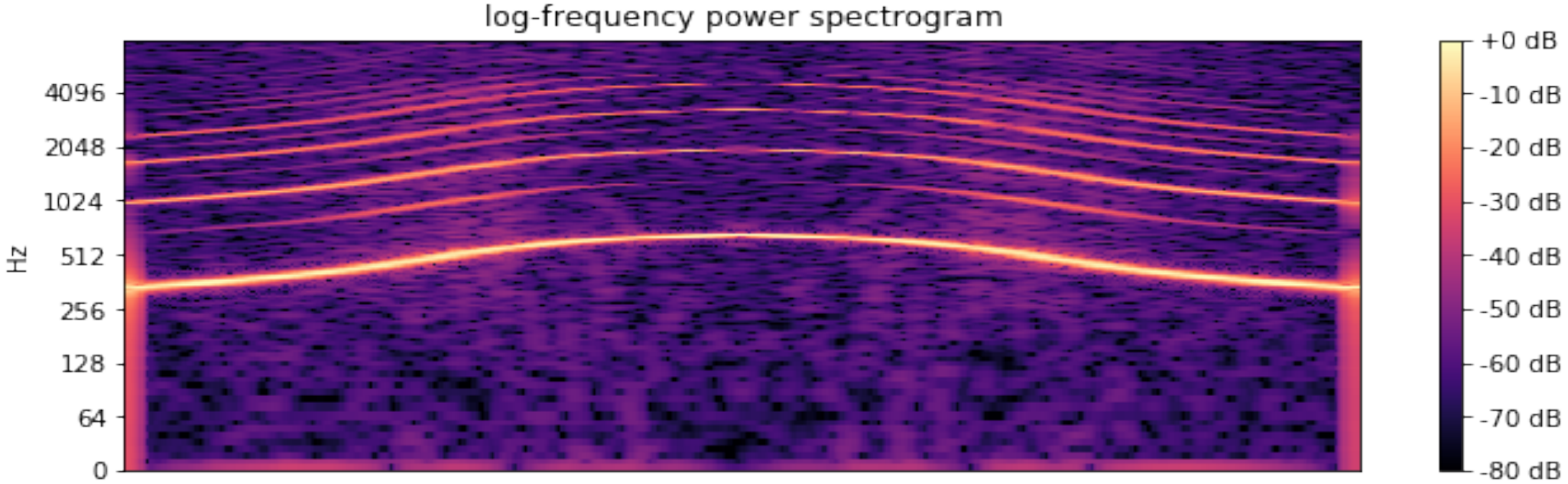
Extreme generalization test



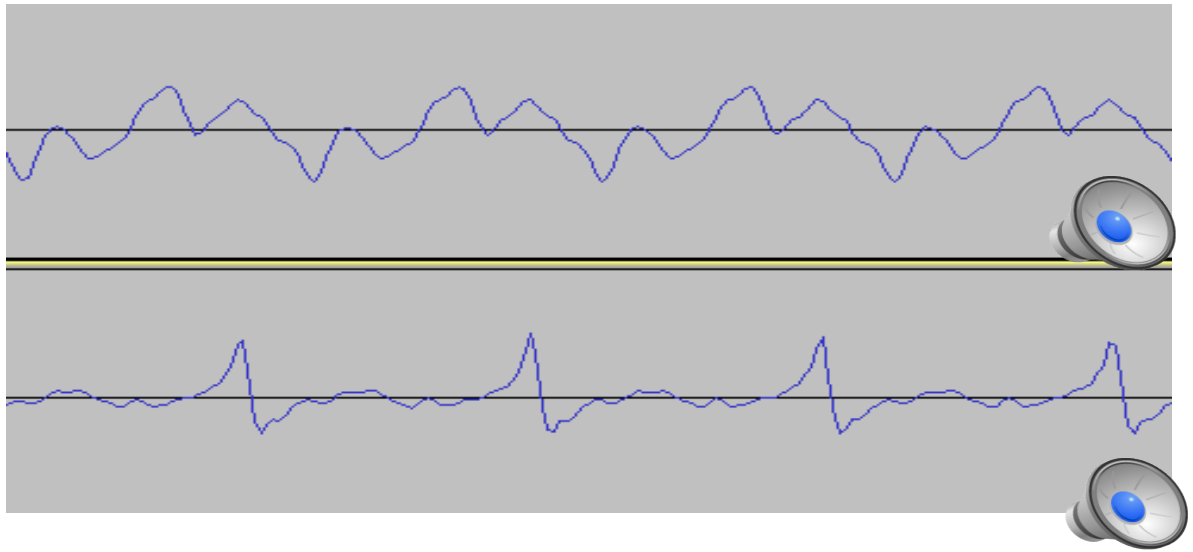


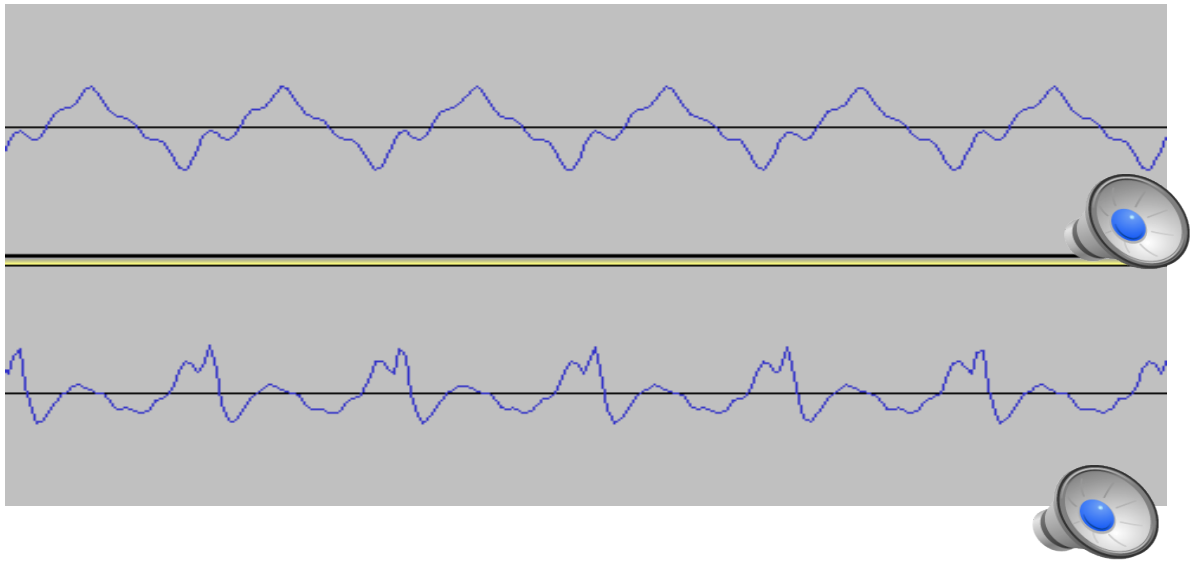
Generalization

Train: two synthetic instruments,
Two pitches, an octave apart: E4, E5



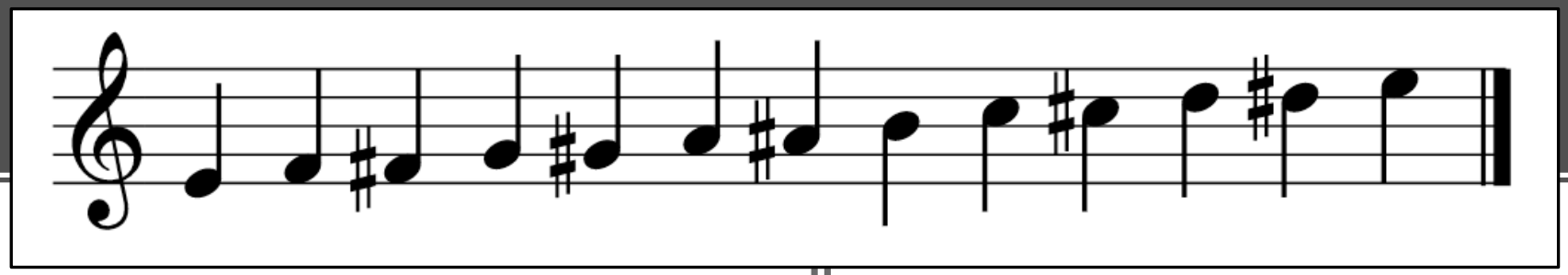
Nsynth acoustic data



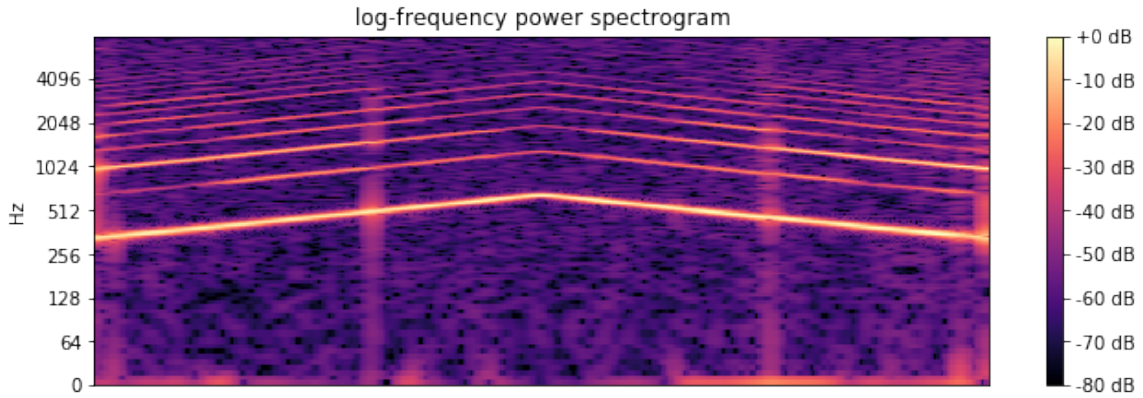


- Waveforms change with instrument
- Waveforms change with pitch

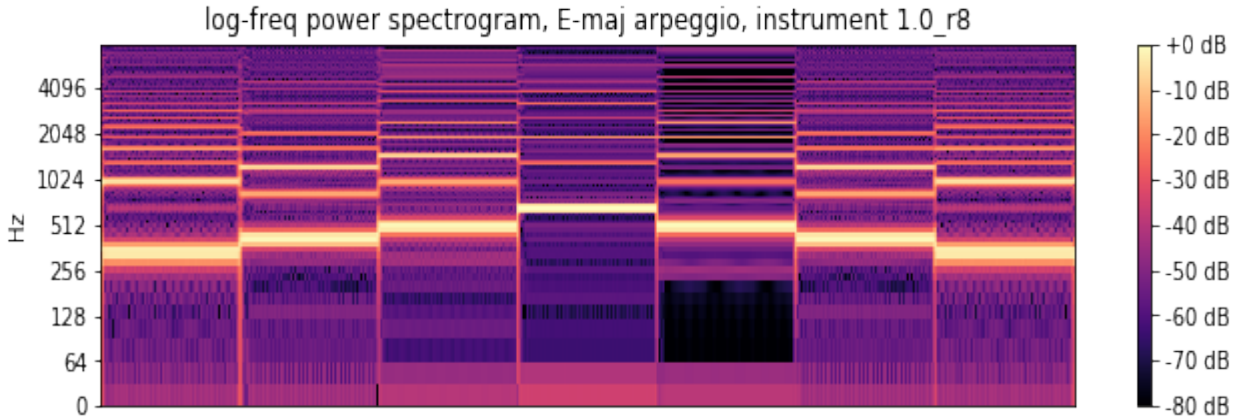
Training on steady-state pitches

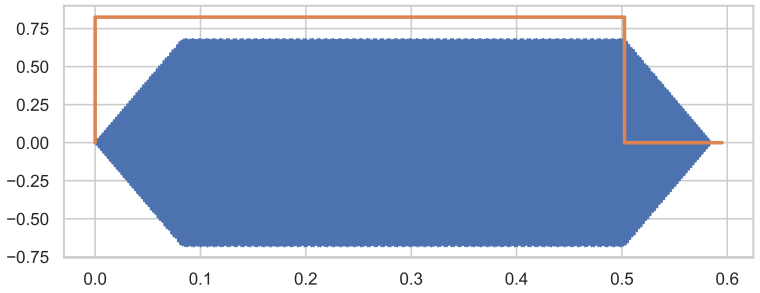


Generalization
(synthesize between chromatic pitches)

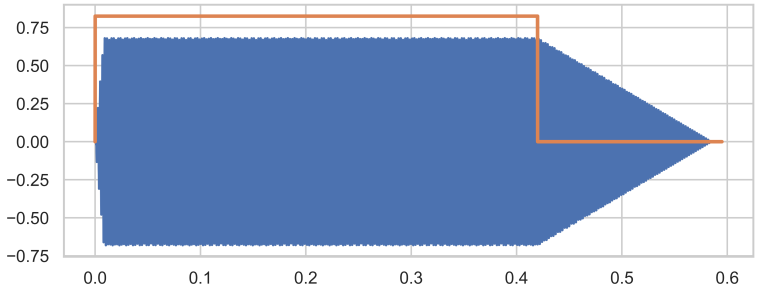


Responsive
(unseen sequences)





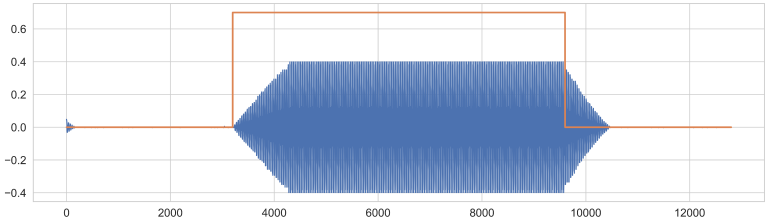
SynthEven

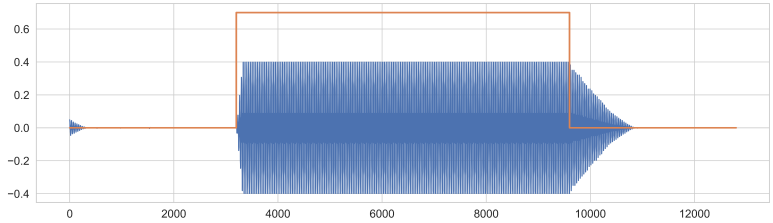


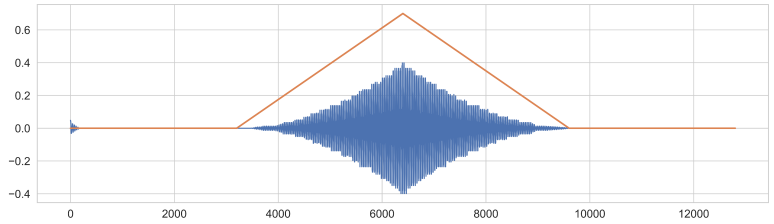
SynthOdd

Transients

- **Implicit** – Not directly controlled by a parameter but is a response to a change in (volume) parameter
- **Parameter disconnect** – Non-instantaneous response to control parameter - forms over time







Synthesis results

Next up : textures

- Noisy
- Difficult distributions
- Parameter identification and labeling challenges (and opportunities)
- But a huge class of musically useful sound

Challenges for sense making

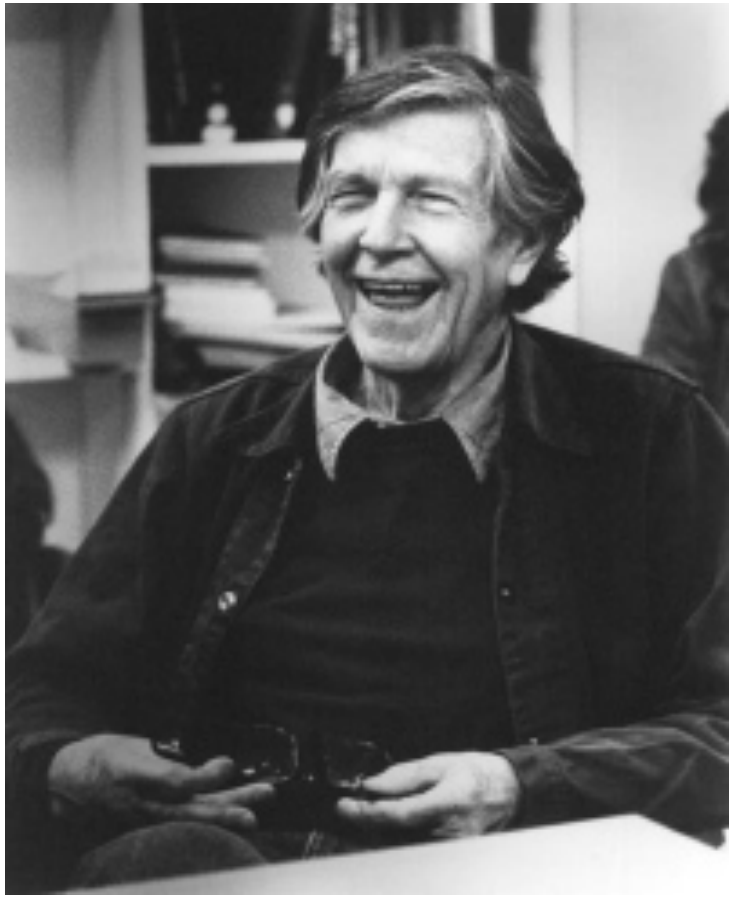
- **No "units"** or "atoms" (like notes) to break everything down in to. A structural element might not even be "a sound".
- No objective typology
- New listening strategies
  - What is pertinent (and what irrelevant) depends upon the listening strategy adopted.
- Analysis needs to reveal the inner mechanics of a work "in text", but also its relationship to the outside world.

Models based Listening

- Spectromorphology
- Gesture surrogacy
- Michel Chion listening modes
  - Causal,
  - Semantic,
  - Reduced
- Transformational (variations)
- Models based Listening
  - When a listener engages in model building as a listening strategy, the model is “generative”. There are sounds/behaviors they could make but haven’t yet.
  - Perhaps more importantly, sounds/behaviors they would not make.

John Cage

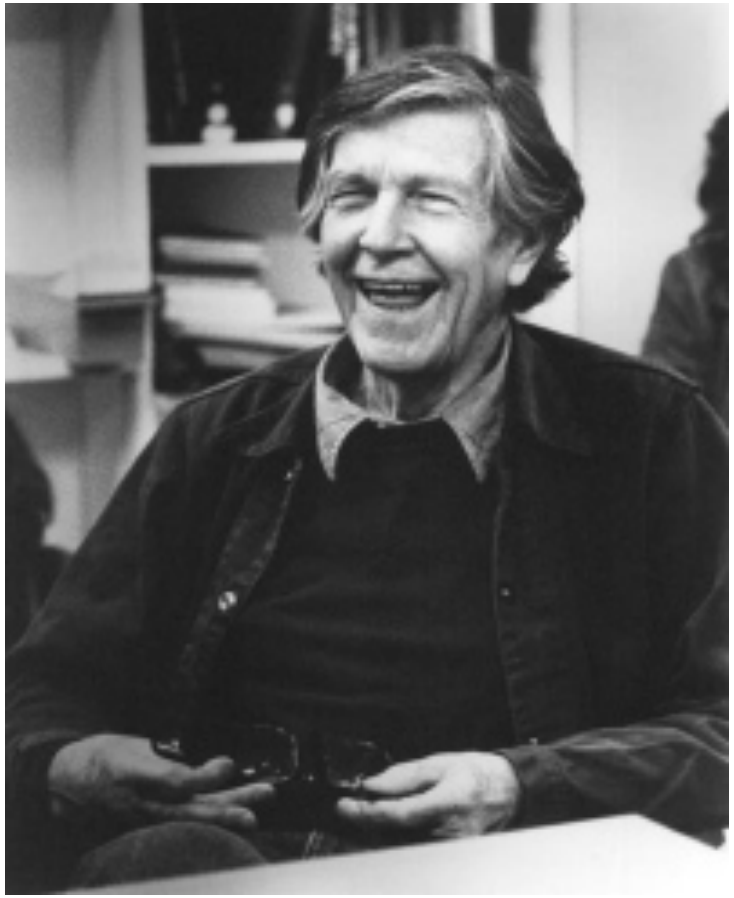
(1952) 4’33”



Turing test?

John Cage

(1952) 4’33”



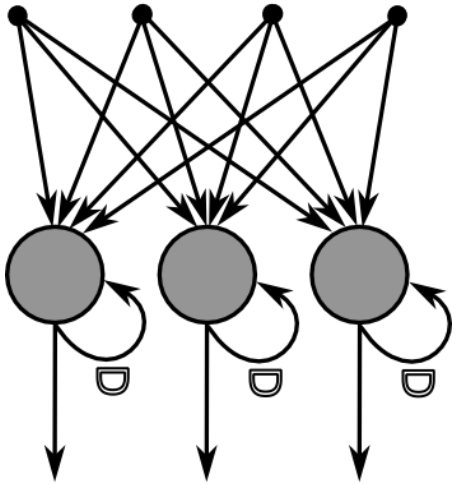
Turing test?

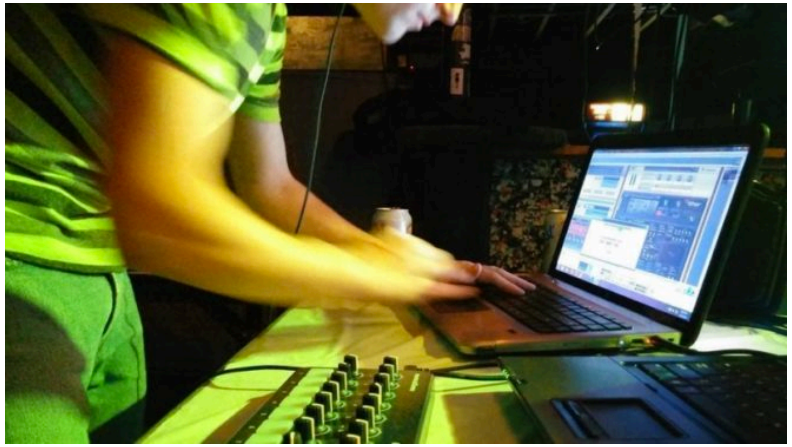
Talk about “out of distribution”

Metaphor and sense-making

- Meaning as an active process
- The origins of language
- *Start* speech understanding with the prosodic/musical elements only, add the words later….

Giambattista Vico (1668-1744) : “The origin of language is metaphor”





Thank you

Please contact me about

PhD funding opportunities

Lonce.wyse@nus.edu.sg