

Collaborative Vocal Puppetry

Multi-User Performative Voice Synthesis on Distributed Platforms @ eNTERFACE 2011

Nicolas d'Alessandro¹, Thierry Dutoit², Maria Astrinaki²,
Johnty Wang¹, Onur Babacan², Àngel Calzada Defez³

¹ MAGIC - University of British Columbia (Vancouver)

² NUMEDIART Institute - University of Mons (Belgium)

³ La Salle - Universitat Ram3n Llull (Spain)

Performative Speech Synthesis

providing artificial speech technologies

vs. providing artificial speaking systems

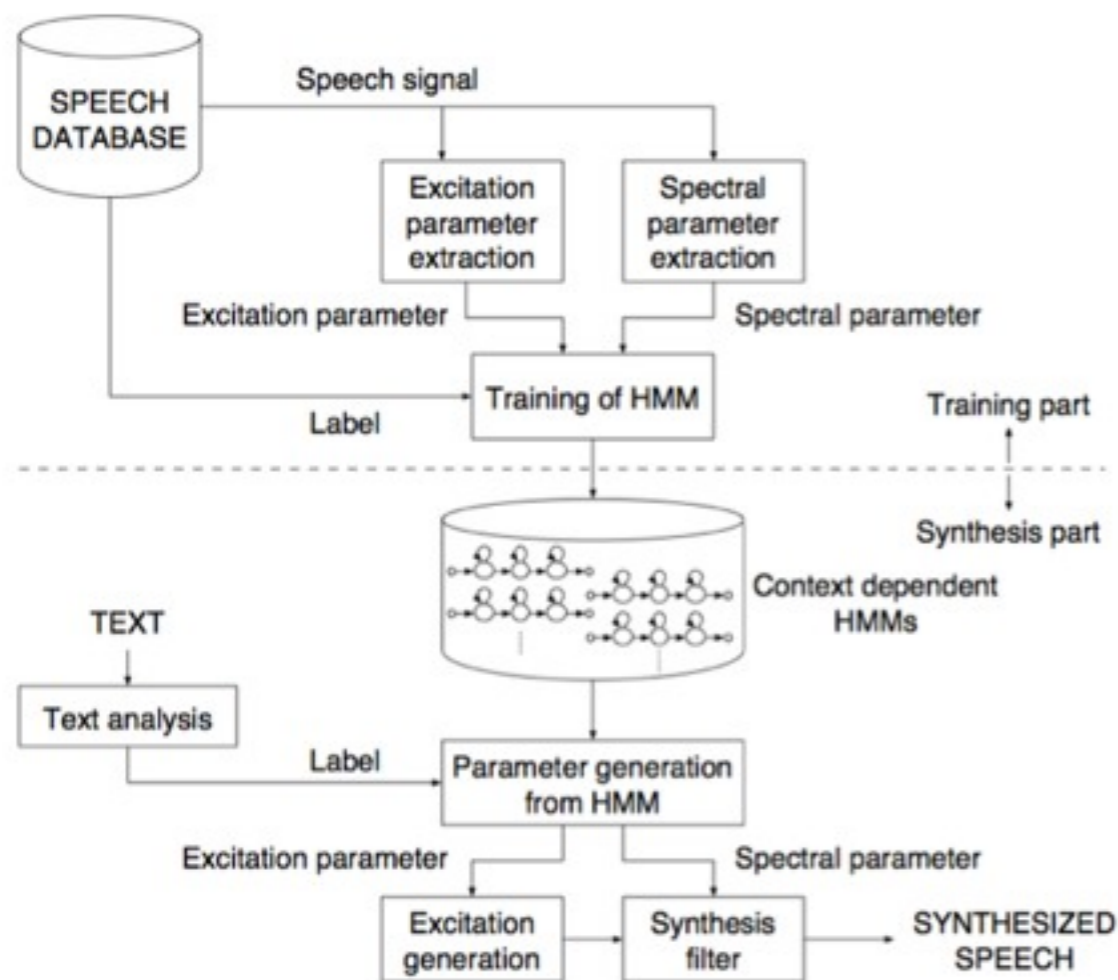
- text-to-speech (TTS) based on non-uniform unit selection (NUU) is the ultimate technology for providing high-quality speech waveforms - period, thank you :)
- it leads to a well-established but rather limited set of apps where the reading of some available text is the goal;
- but it does not make the conception of an *artificial speaker* much easier than back in the days of DECTalk, well actually it makes it harder (black-box design, time scales issues);
- the *artificial speaker* as a system: realtime, reactive, interactive, listener-specific, context-aware, ubiquitous,... many aspects which require a different perspective.

exploring... based on the two most promising approaches we have so far

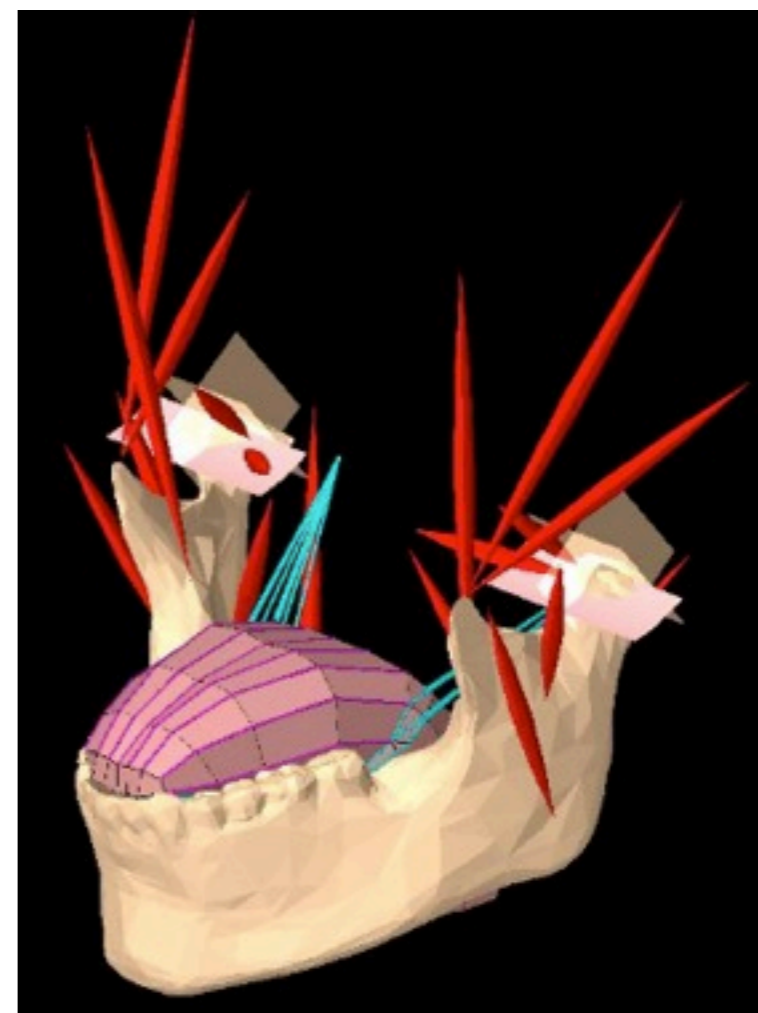
- 1) bio-mechanical modelling**
- 2) statistical parametric modelling**

Starting Points

two promising but non-adapted, totally non-reactive systems: HTS and Artisynth



HTS



Artisynth

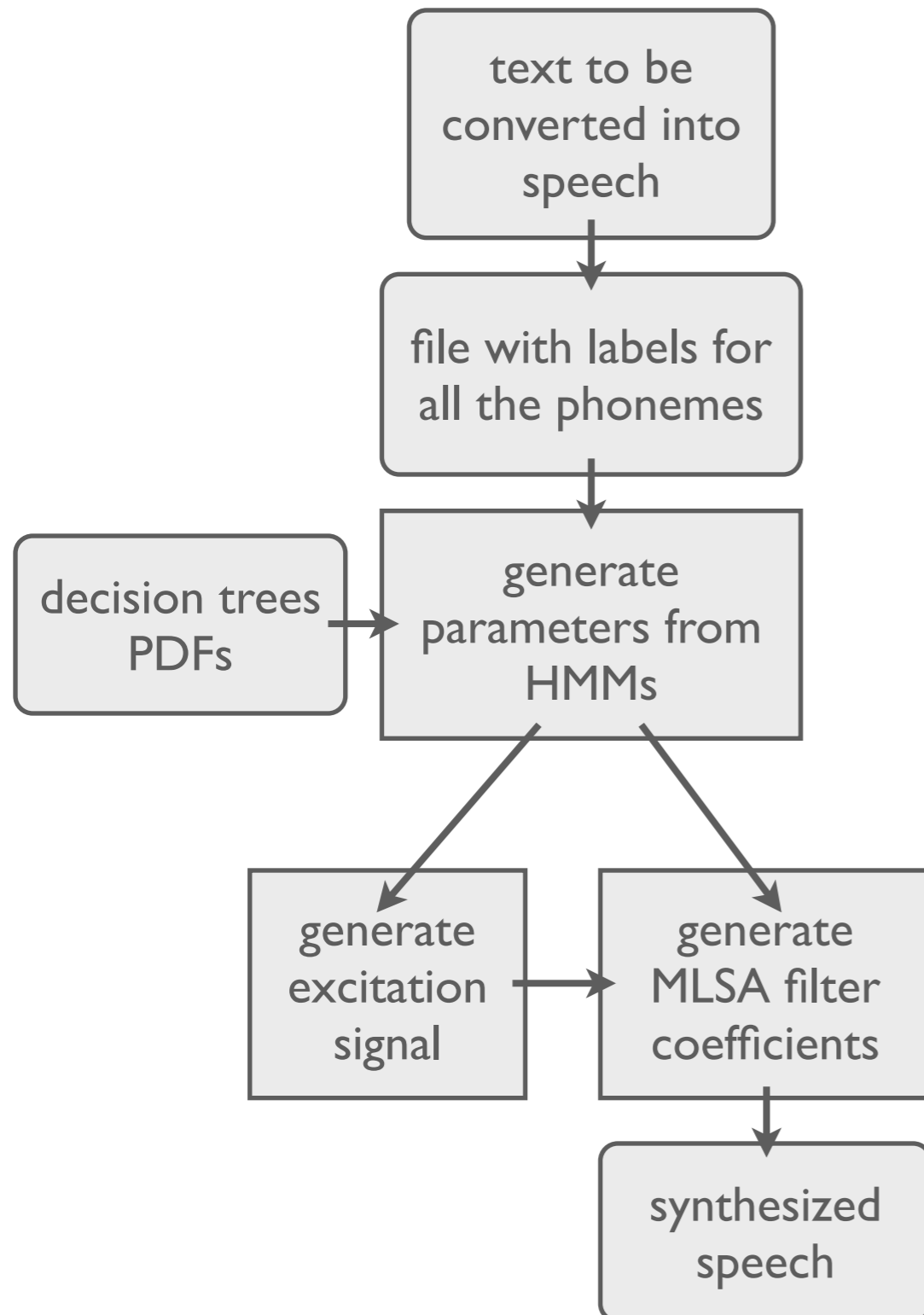
I think we went over the hill...
we reached convincing reactivity for both

MAGE



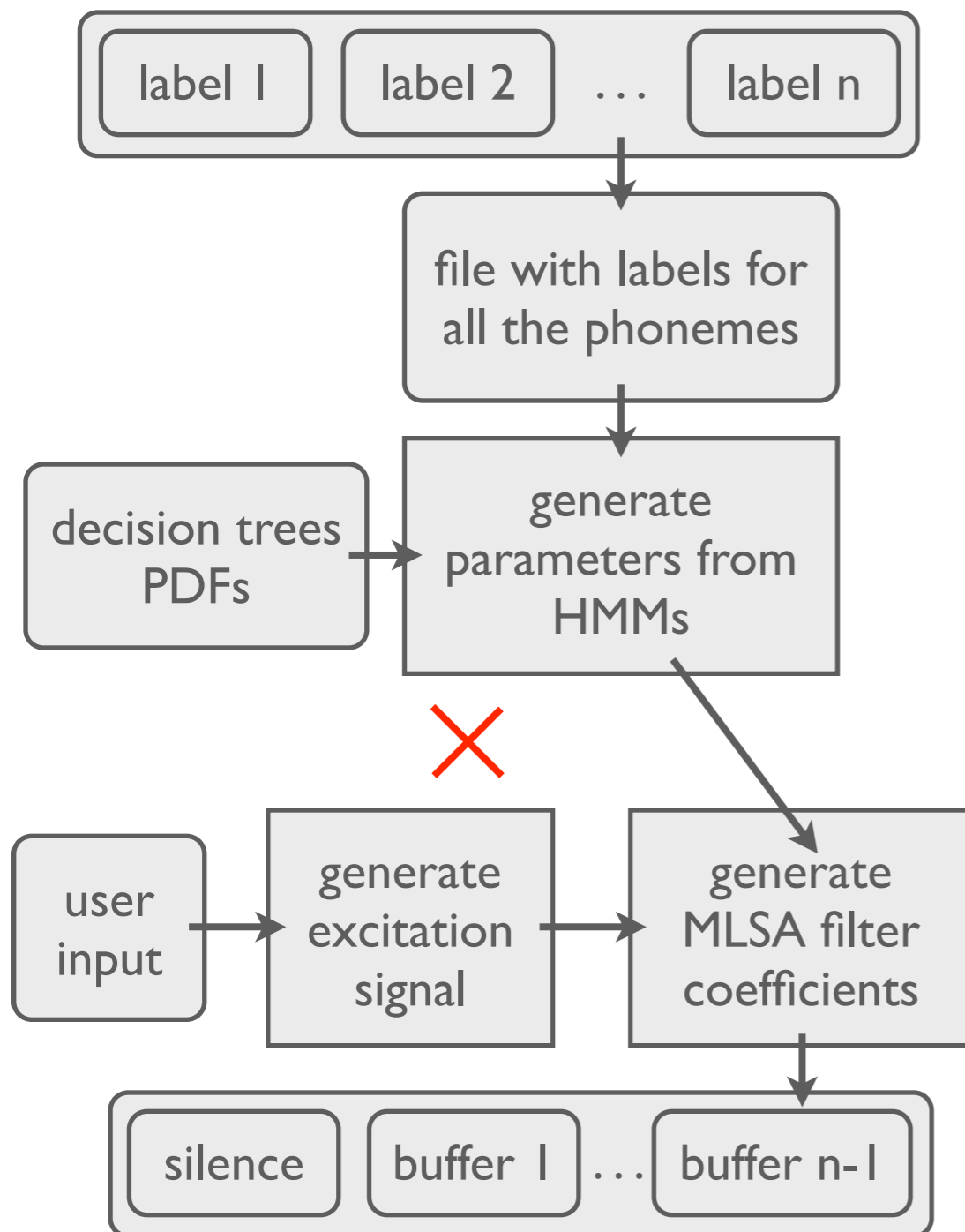
INVESTMENTS IN EDUCATION DEVELOPMENT

HTS: HMM-Based Speech Synthesis



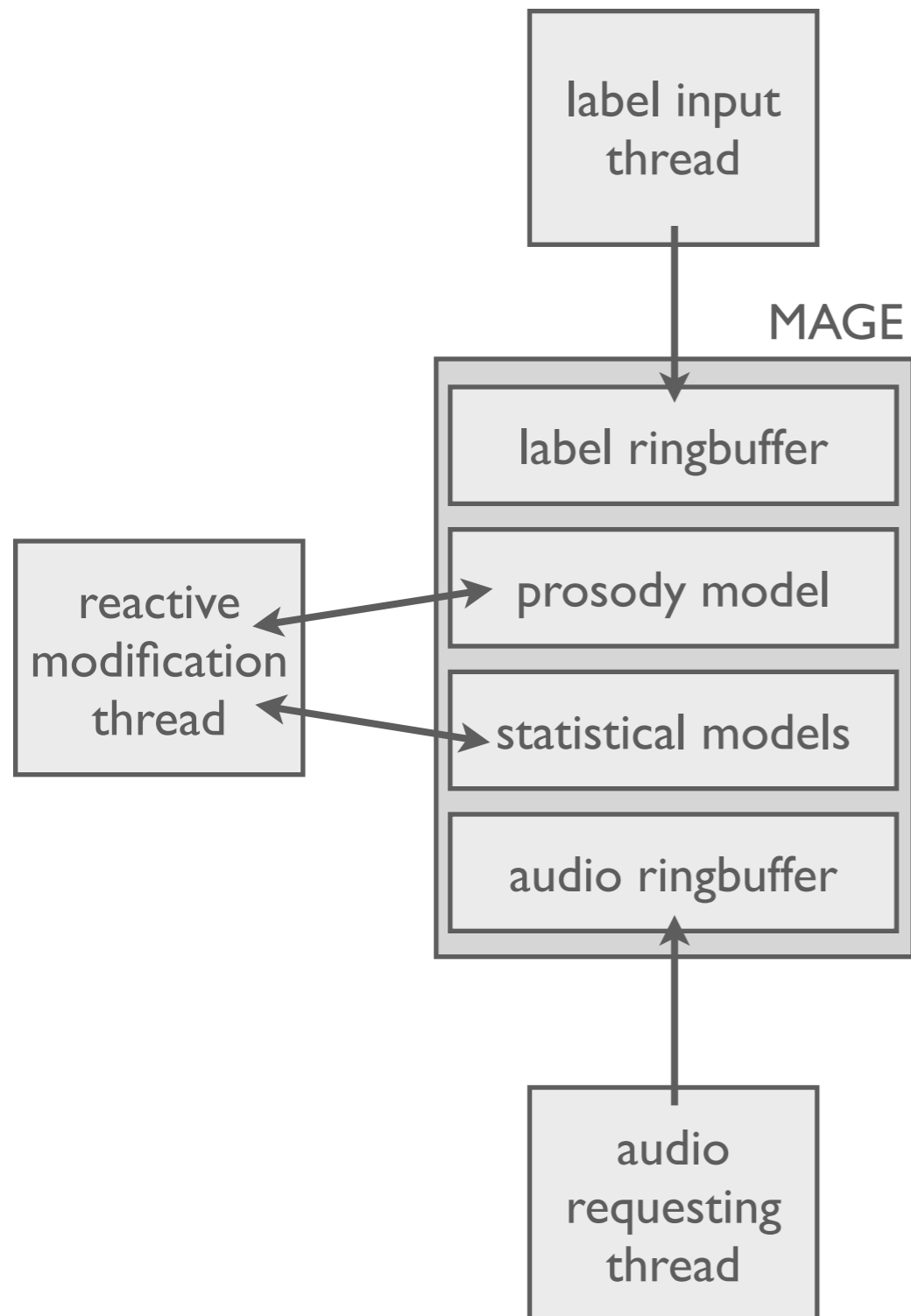
- input = full text to be converted
- parsing text into labels containing full contextual linguistic information
- select created decision trees and PDFs from the database
- generate spectrum, pitch and duration trajectories, optimized for the whole label target
- generate speech in a file

pHTS: Breaking the loop

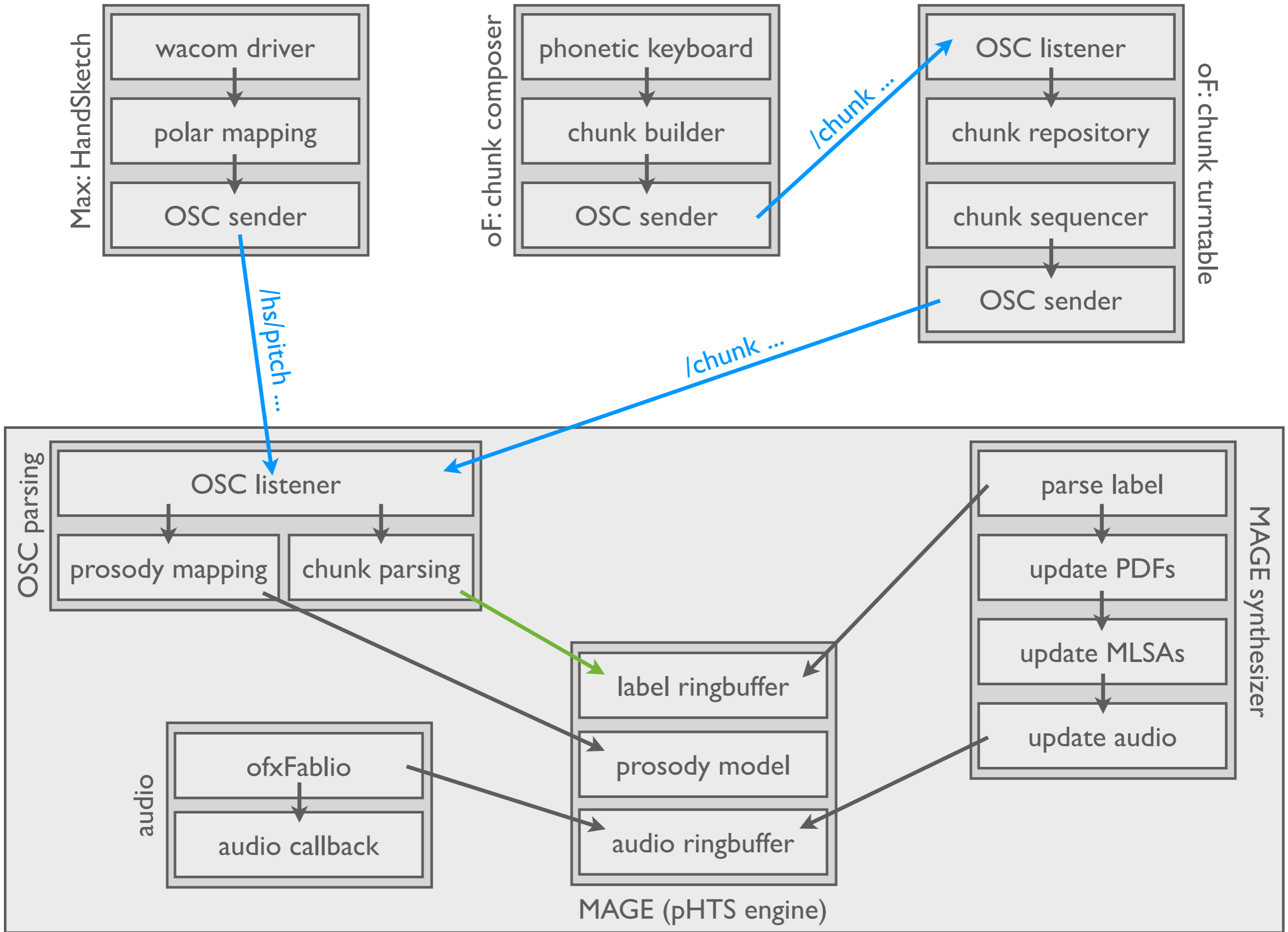


- sequential input of labels containing minimal contextual linguistic information
- delay of only one label to preserve coarticulation
- select created decision trees and PDFs from the database
- generate spectrum, pitch and duration trajectories for one label
- generate one buffer of audio

MAGE: Ready for reactive software



- thread-safe interface between the synthesizer and I/Os: labels and audio samples (lock-free)
- realtime access to the internal state of the engine: each step of the statistical modelling + prosody
- engine-independent



Reactive Natural Language Processing

HTS training

using full contextual linguistic information (from phonemes to complete utterance)

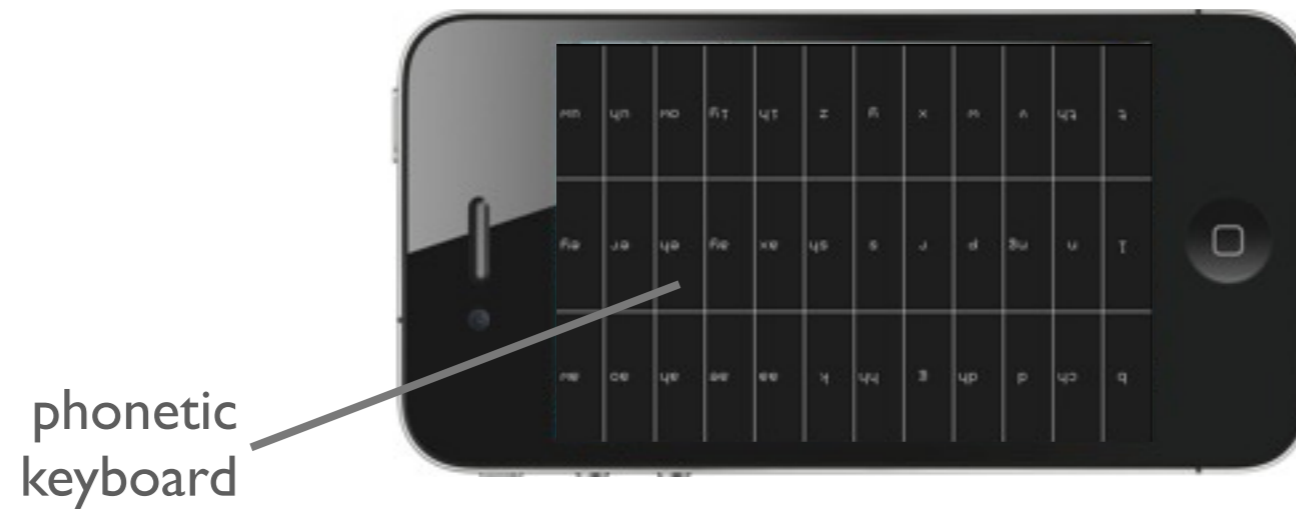
pHTS/MAGE training

using minimum contextual linguistic information reduced only to past and present phonemes and syllables

Reactive Natural Language Processing

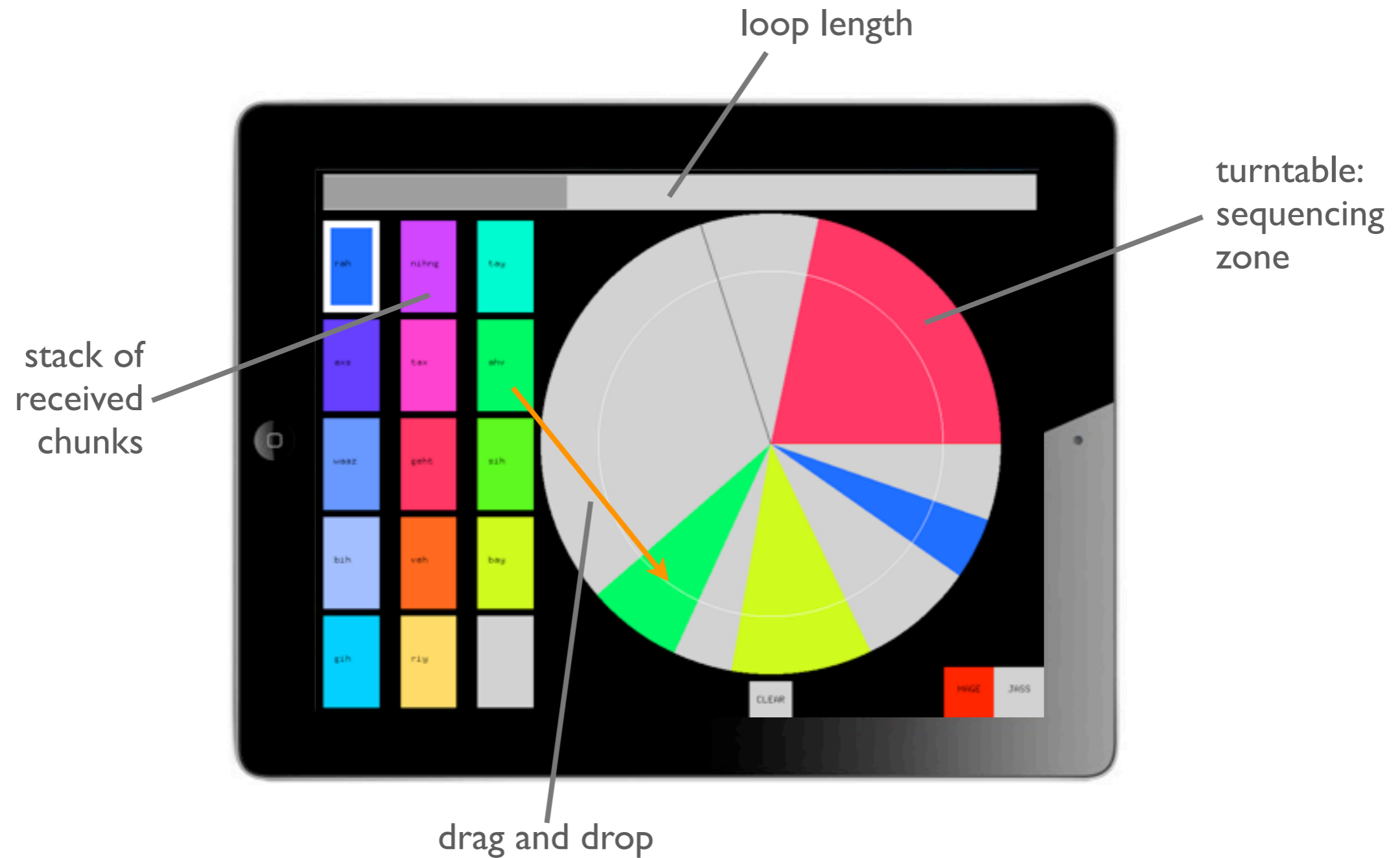
- phoneme:
 - {before previous, previous, current, next, next next}
 - position of current phoneme in current syllable
 - syllable:
 - number of phonemes at {before previous, previous, current, next, next next}
 - accent of {before previous, previous, current, next, next next}
 - stress of {before previous, previous, current, next, next next}
 - position of current syllable in current word
 - number of {preceding, succeeding} stressed syllables in current phrase
 - number of {preceding, succeeding} accented syllables in current phrase
 - number of syllables {from previous, to next} stressed syllable
 - number of syllables {from previous, to next} accented syllable
 - vowel within current syllable
 - word:
 - guess at part of speech of {preceding, current, succeeding} word
 - number of syllables in {preceding, current, succeeding} word
 - position of current word in current phrase
 - number of {preceding, succeeding} content words in current phrase
 - number of words {from previous, to next} content word
 - phrase:
 - number of syllables in {preceding, current, succeeding} phrase
 - position in major phrase
 - ToBI7 endtone of current phrase
 - utterance:
 - number of syllables in current utterance
- phoneme:
 - {before previous, previous, current, next, ~~next next~~}
 - position of current phoneme in current syllable
 - syllable:
 - number of phonemes at {before previous, previous, current, next, ~~next next~~}
 - accent of {before previous, previous, current, next, ~~next next~~}
 - number of syllables {from previous, to current} accented syllable
 - number of syllables {to next} accented syllable : random
 - vowel within current syllable
 - word:
 - No information
 - phrase:
 - No information
 - utterance:
 - No information

Chunk Composer



double-tap sends the chunk

Chunk Turntable



double-tap stresses the vowel
chunk is sent when head hits

HandSketch



fan diagram: control
through angle and radius

angle ~ pitch variation
radius ~ speed variation

pen: pressure
and tilt + buttons

pressure ~ volume
tilt ~ vocal tract size
button #1 ~ abs/rel pitch
button #2 ~ gate tilt

FEM Tongue

Artisynth vocal tract [/Users/johnny/Documents/workspace/artisynth_2_0]

File Models Edit View Help

scale 1023 0 0. 1. scale 1023 0

prepend /mo

prepend-GGP

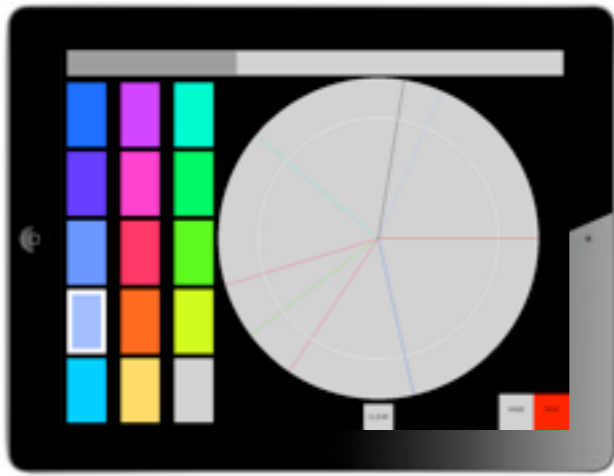
Tongue Exciters

GGP	0.500000
GGM	0.000000
GGA	0.000000
STY	0.000000
GH	0.000000
MH	0.000000
HG	0.000000
VERT	0.067900
TRANS	0.000000
IL	0.000000
SL	0.000000

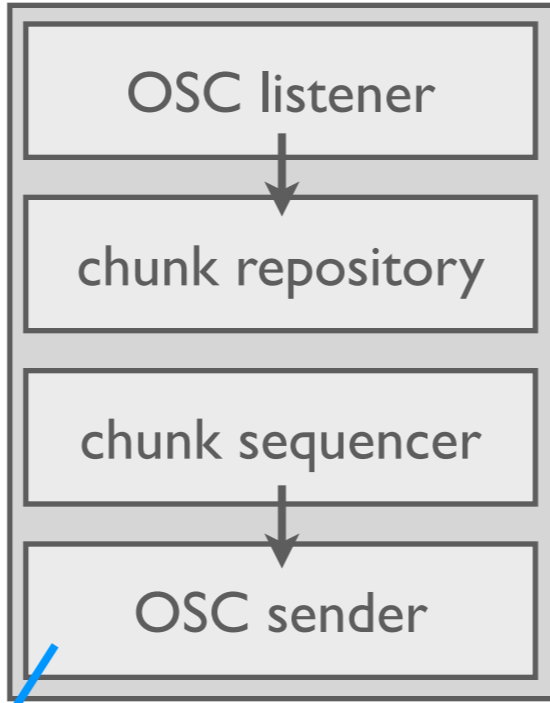
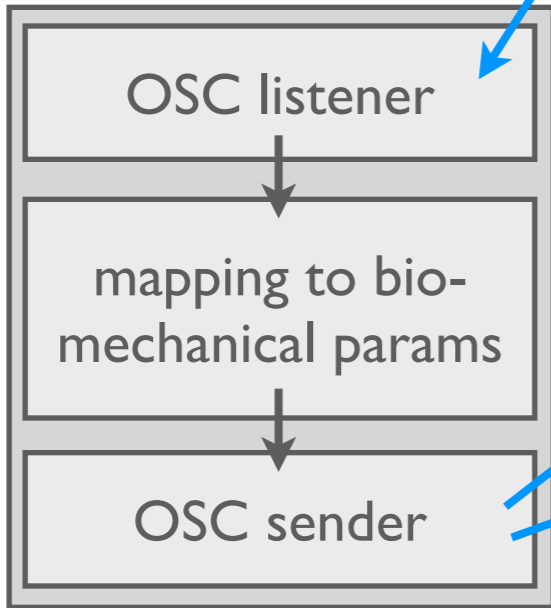
Vocal Tract

u_xx mult1
u mult1.0
wall coeff
lipCf 1.0
length 0.17
A(0) 0.1256
A(1) 0.1256
A(2) 0.1256
A(3) 7.0685
A(4) 7.0685
A(5) 7.0685
A(6) 5.6977
A(7) 4.7289
A(8) 3.8210
A(9) 3.0548
A(10) 2.706
A(11) 1.797
A(12) 1.355
A(13) 1.209
A(14) 0.842
A(15) 1.420
A(16) 2.149
A(17) 3.890
A(18) 2.360
A(19) 0.502
A(20) 1.539
A(21) 0.502

models/tongue/bundles/STY_R/fibres/64

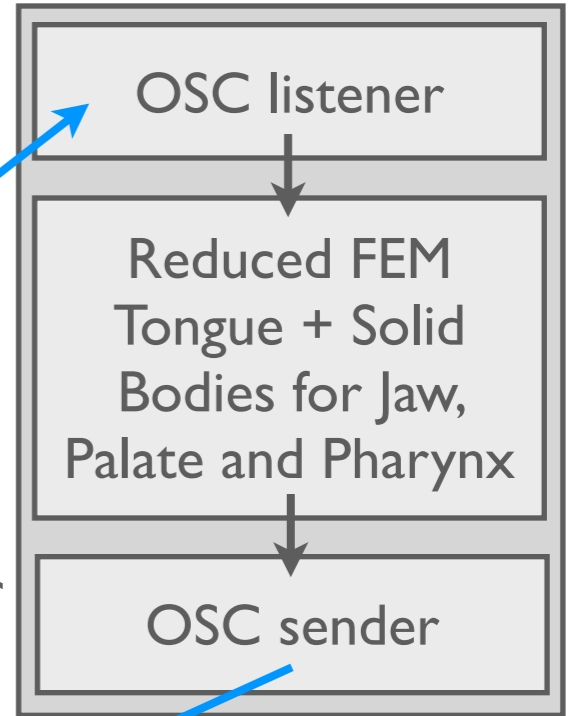


Max: mapping patch

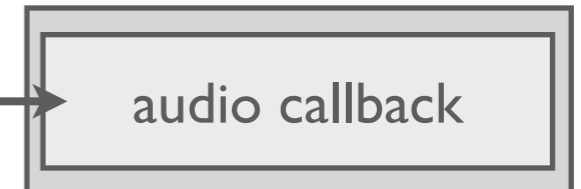
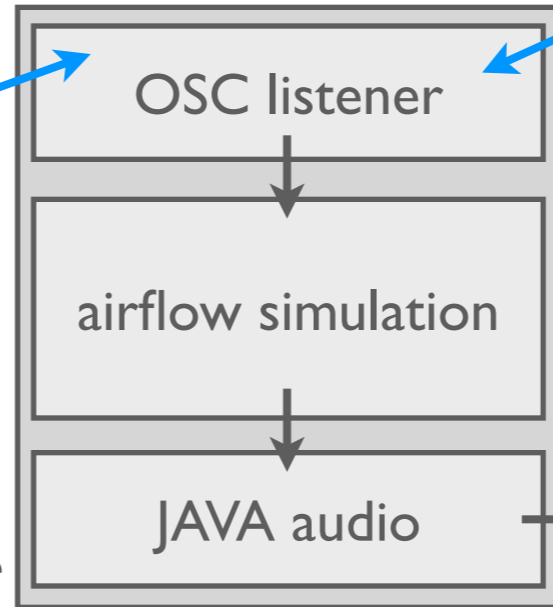


of: tract turntable

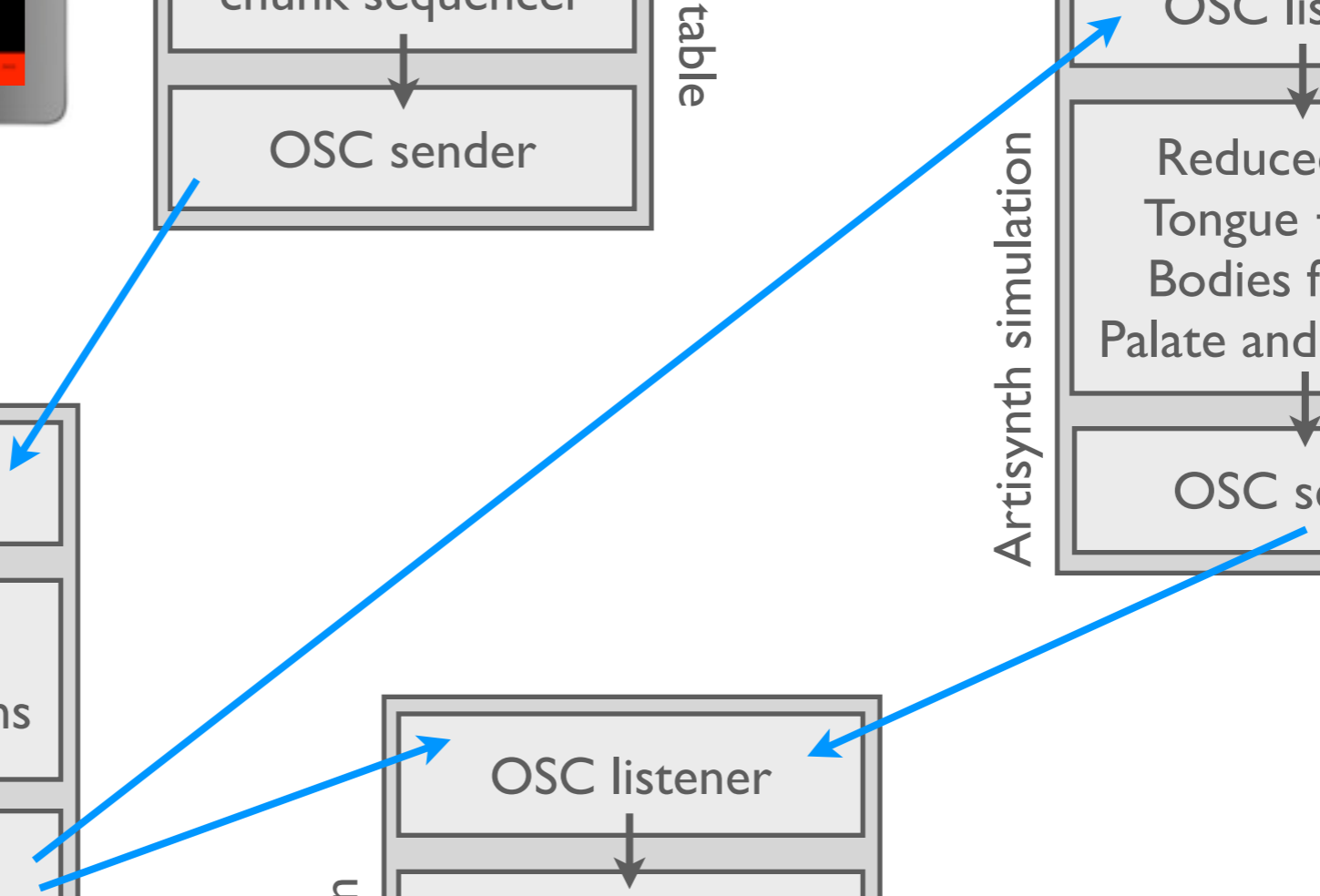
Artisynth simulation

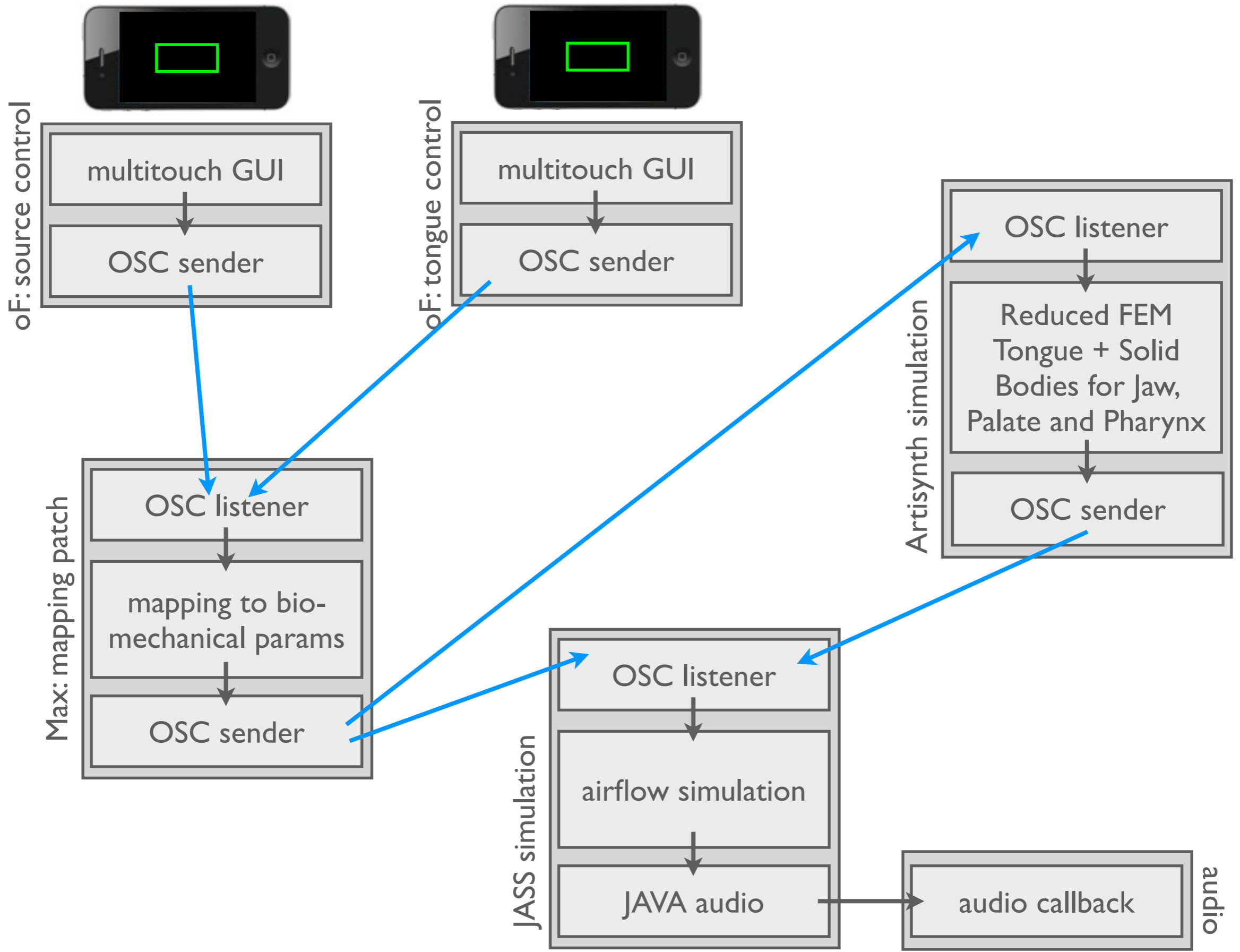


JASS simulation

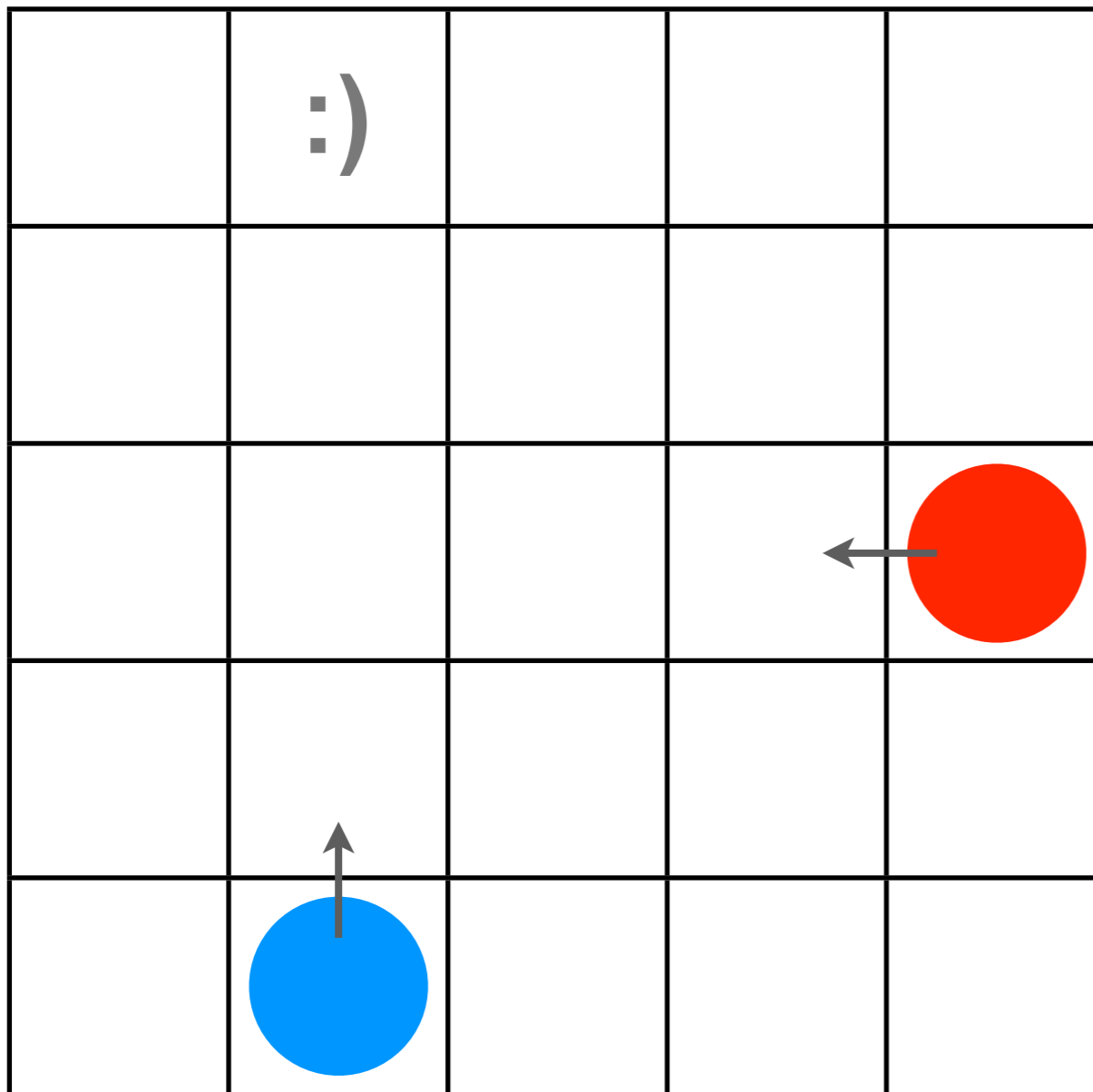


audio





Game



- simple language to tell directions, colours and validate action or not
- one user is instructed to move coins by other who is talking through the device
- we record all trajectories achieved by the “instructor” and time to achieve the task

Game

WORD	MEANING	GESTURE
ale	red	back open/middle closed/front open
ela	blue	front open/middle closed/back open
lau	up	back closed/back open/back closed
lua	down	back open/back closed/back open
lea	left	front closed/front open/back open
lae	right	back closed/back open/front open
ea	yes	front closed/back open
ua	no	back closed/back open

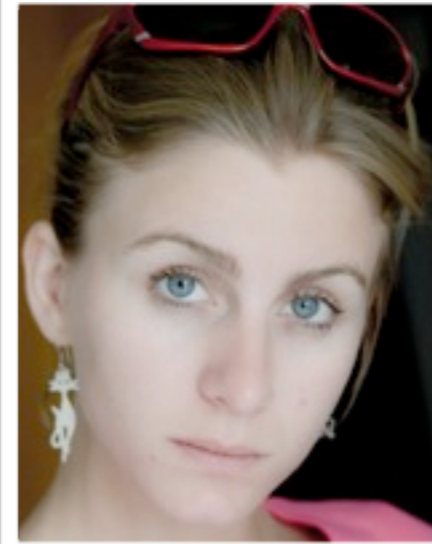
Team



Nicolas
Alessandro



Thierry
Dutoit



Maria
Astrinaki



Johnty
Wang



Àngel
Calzada



Onur
Babacan

thank you
guys !!!

thanks for your attention
come and see/try our demos

