</ A Real-Time Audiovisual Approach to Vowel Detection />

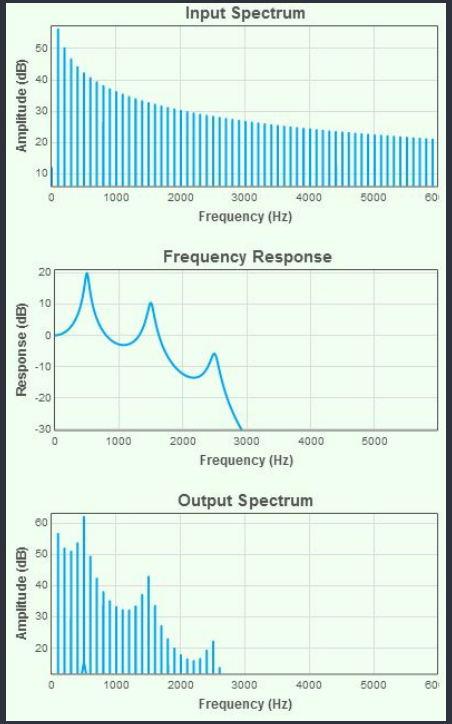
} /> [

Chris Morse, Francesco Papaleo, Tommaso Settimi

1 0 1 1 0 1 1 0 1 1 0 1 1 0 0 1 1 0 1 1 0 1 1 0 1 1 0 1 1 1 0 1 1 0 1 1 0 1 1 1 1 1 0 1

</ Background: Vowel Production

- Vowels: Speech sounds created without any stricture in the vocal tract
- **Source Filter Model:** Sounds are generated and independently shaped (filtered) by the vocal tract acting as a resonating tube
- The effect of vocal tract manipulation can be characterized by frequency response
- The frequency response of vowel sounds contains peaks, called **formants** which correspond to the resonant frequencies of the vocal tract
- Each vowel can be characterized by the relationship between formants irregardless of differences in the fundamental frequency.



1 0 1 1 0 1 1 0 1 1 0 1 1 0 0 1 1 0 1 1 0 1 1 0 1 1 1 0 1 1 0 1 1 1 0 1

</ Vowel Detection System

- Vowels form the basis of all known languages
- Correct vowel pronunciation is important for later success in more advanced linguistic tasks
- **Goal:** create a program which accurately detects and classifies different **Italian** vowels.
- Use the vowel detection to help non-native speakers better learn correct pronunciation.
- This could have possible applications in **language education, speech therapy and speech pathology**

1 0 1 1 0 1 1 0 1 1 0 1 1 0 0 1 1 0 1 1 0 1 1 0 1 1 0 1 1 1 0 1 1 0 1 1 0 1 1 1 1 1 0 1

</ System Overview

- **Three main components:**
  - Audio input/classification
  - Video input/classification
  - Graphical User Interface
- Audio Detection
  - Classification of formants associated with each value
- Video Detection
  - Classification of facial features associated with each value
- Graphical User Interface
  - Ability to select a specific vowel to practice
  - Accuracy rating feedback
  - Adjustable difficulty levels

1 0 1 1 0 1 1 0 1 1 0 1 1 0 0 1 1 0 1 1 0 1 1 0 1 1 0 1 1 1 0 1 1 0 1 1 0 1 1 1 1 1 0 1

</ Audio processing in SC (1/2)

- Audio input ->
- FFT and loudness tracking (percpetual model, summing excitation in ERB bands) ->
- Gate to de-activate the algorithm (send 0.0s) when no vowel is spoken ->
- Amplitude normalisation for make the vowel recognition independent from the amplitude ->
- HPFilter (over 100 Hz) and LPFilter (under 3000 Hz), since the 2 main formants for vowel recognition are between that range ->
- FFT -> take the local maxima and filter out (passes only bins whose magnitude is above a threshold and above their nearest neighbors) ->

</>

1 0 1 1 0 1 1 0 1 1 0 1 1 0 0 1 1 0 1 1 0 1 1 0 1 1 0 1 1 1 0 1 1 0 1 1 0 1 1 1 1 1 0 1

</ Audio processing in SC/Weki [2/2]

- IFFT ->
- BandPass filter bank with center frequencies at the formant frequencies of each of the first two formants for of the 5 vowels (high Q factor)
- RMS running sum and linear scaling for each couple of filters to measure the energy passing in each formant filters couple (5 output signals) ->
- OSC ->
- Wekinator 5 inputs ->
- Classification through NN (3 hidden layers), 5 outputs ->
- Max/MSP

</>

1 0 1 1 0 1 1 0 1 1 0 1 1 0 0 1 1 0 1 1 0 1 1 0 1 1 0 1 1 1 0 1 1 0 1 1 0 1 1 1 1 1 0 1

</ Video Detection/Classification

- Face detection: FaceOSC
- Sends gesture features through OSC routing to intermediate interface
- Intermediate interface: Audiovisual Server (Python)
    - Reformatting of OSC messages in a Wekinator-friendly manner
    - Thread management to “ensure” sync
    - Audio + visual message packing

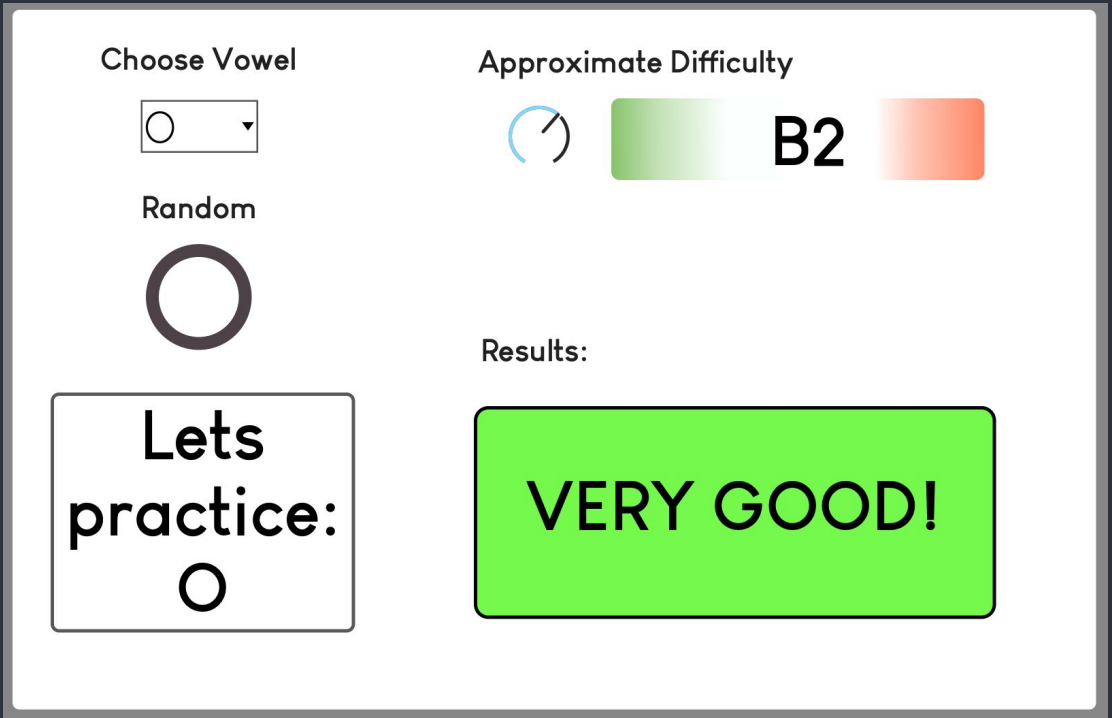
1 0 1 1 0 1 1 0 1 1 0 1 1 0 0 1 1 0 1 1 0 1 1 0 1 1 0 1 1 1 0 1 1 0 1 1 0 1 1 1 1 1 0 1

</ Graphical User Interface

- Designed in Max/MSP
- Vowel Selection
  - User can select one of 5 vowels (A, E, I, O, U) from a dropdown menu or press a button to randomly select one
  - The selected value informs the Wekinator output which floating point to pass to the evaluation system
- Evaluation system
  - Wekinator sends two floating point values corresponding to the Audio and Video outputs. Values are weighted according to their importance
  - Weighted floating point value is classified based on variable thresholds:
    - GOOD!
    - NEEDS WORK
    - NO >:(
  - Evaluation thresholds can be lowered to make achieving the GOOD! And NEEDS WORK categories easier

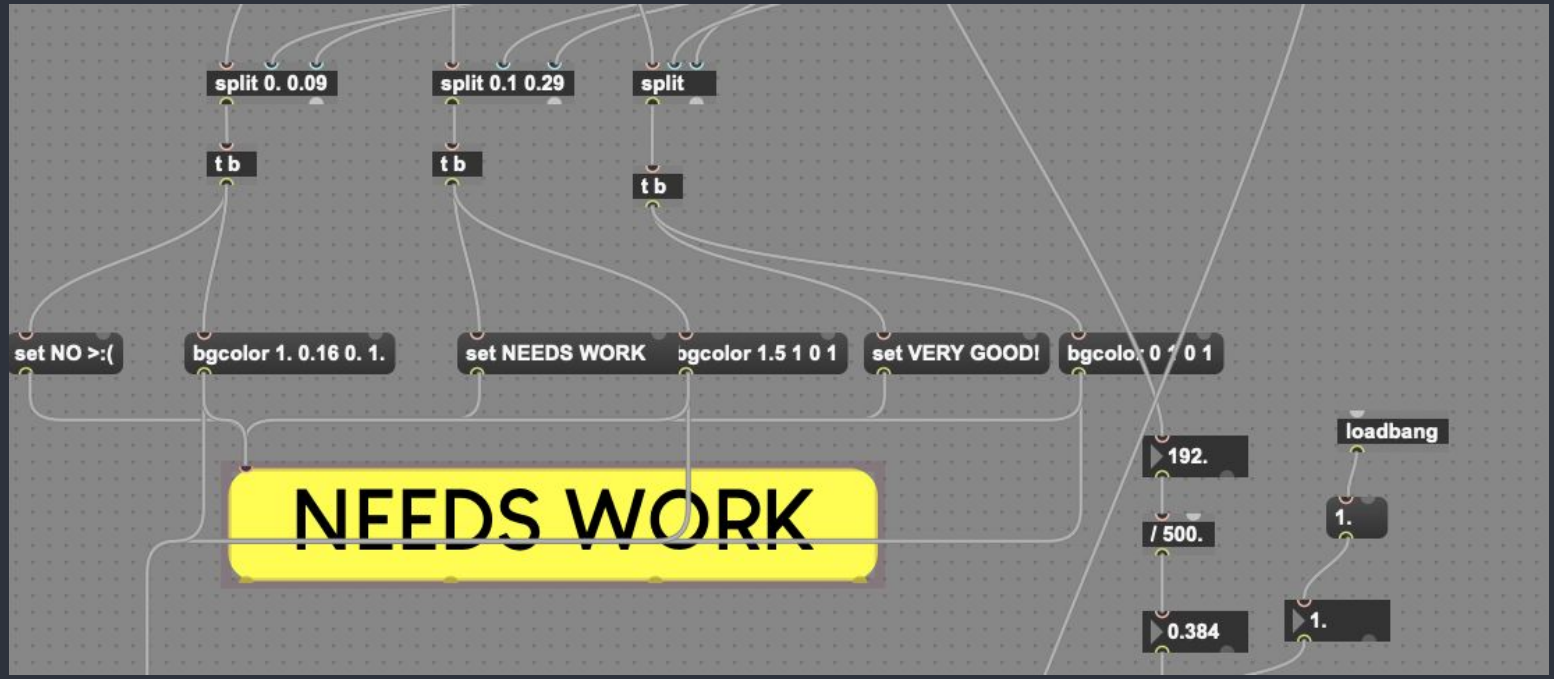
1 0 1 1 0 1 1 0 1 1 0 1 1 0 0 1 1 0 1 1 0 1 1 0 1 1 0 1 1 1 0 1 1 0 1 1 0 1 1 1 1 1 0 1

</ Graphical User Interface



1 0 1 1 0 1 1 0 1 1 0 1 1 0 0 1 1 0 1 1 0 1 1 0 1 1 0 1 1 0 1 1 0 1 1 1 0 1 1 0 1 1 1 0 1 1 0 1

</ Graphical User Interface



1 0 1 1 0 1 1 0 1 1 0 1 1 0 0 1 1 0 1 1 0 1 1 0 1 1 0 1 1 1 0 1 1 0 1 1 0 1 1 1 1 1 0 1

Citations

Huckvale M. PALS1004 Introduction to Speech Science. https://www.phon.ucl.ac.uk/courses/spsci/iss/week5.php

1 0 1 1 0 1 1 0 1 1 0 1 1 0 0 1 1 0 1 1 0 1 1 0 1 1 0 1 1 1 0 1 1 0 1 1 0 1 1 1 1 1 0 1