Alexandros Kontogiorgos-Heintz
NE360/HS361
Lab 10: Due 12/5/13

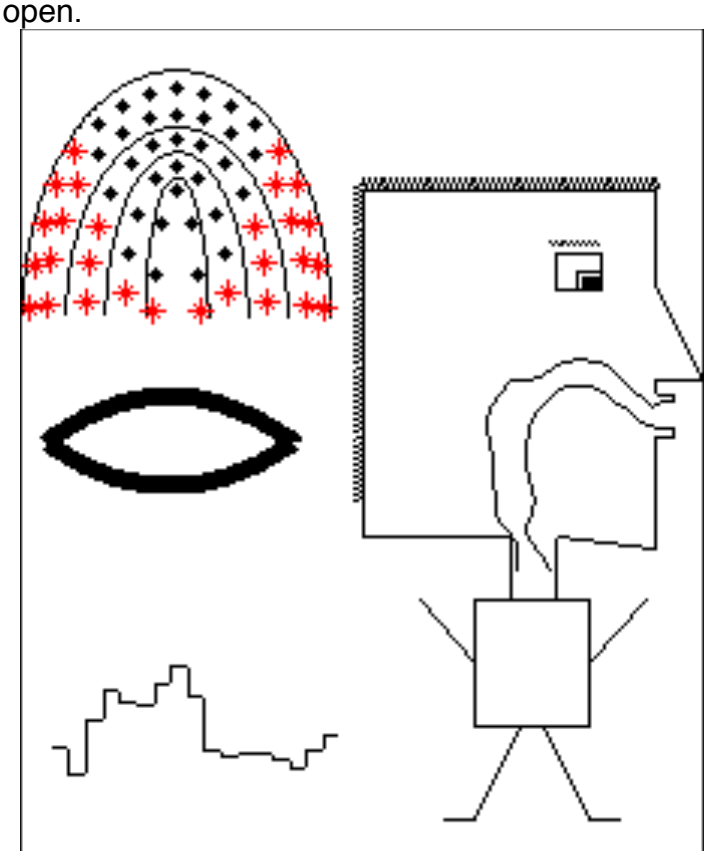
Abstract:

In this lab we experimented with the DIVA model. DIVA is a neural network model of speech production that is trained by babbling, then used to model motor movements used to create speech sounds. In this lab we experimented with its forward and inverse controllers to produce phonemes. We looked at how the model encodes auditory and somatosensory signals during learning, and how its accuracy increases with learning.

Results:

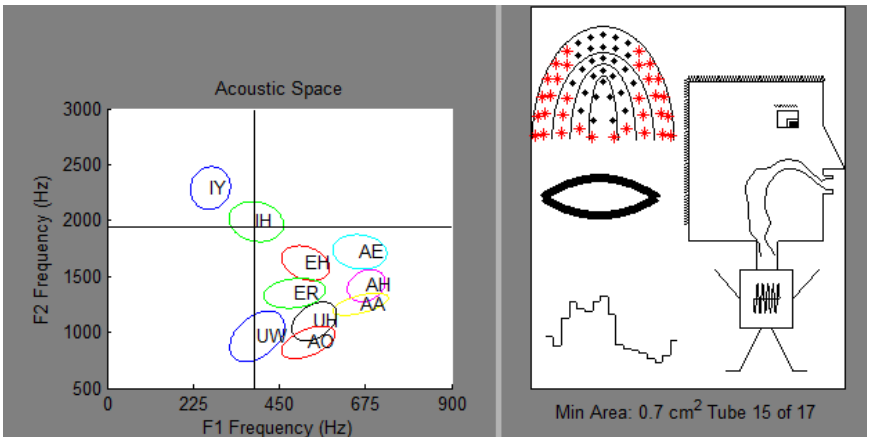
In part one of the lab we looked at the forward model in DIVA. We looked at the production of the baseline of the model, and modified specific individual parameters to observe their effects on the model.

- The First 3 formants were F1: 490, F2: 1796, F3: 2500. The shape of the vocal tract is narrow at the glottis, and relatively straight and open until it reaches the tongue, which is raised at the back. The jaw is lowered, and the lips are open.

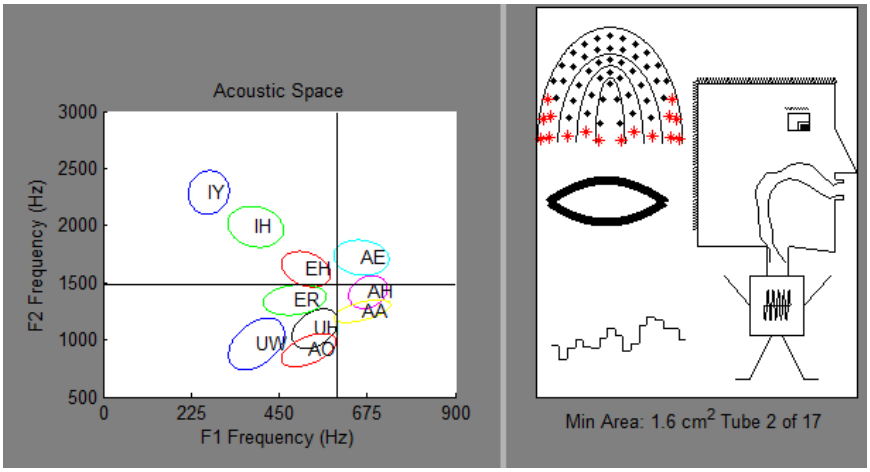


- Modifying the jaw height, tongue body height, and tongue body shape moves the position in acoustic space along the downward diagonal axis.

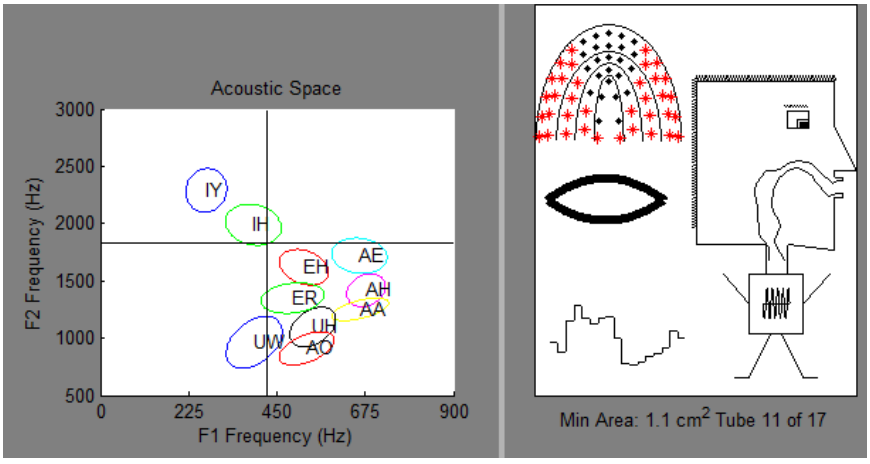
Modified jaw height to 1.50: sound was “ih”



Modified tongue body height to 1.50 sound was “ah”

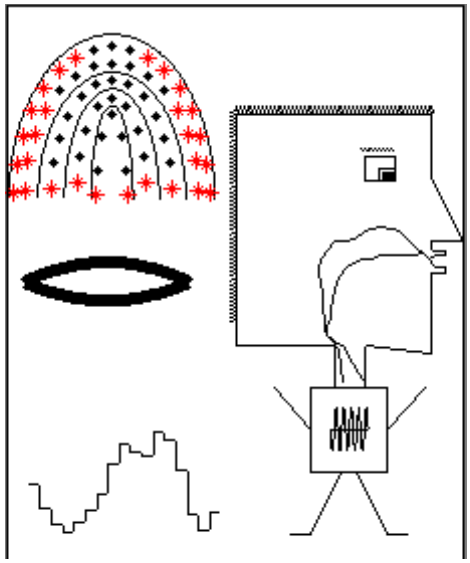


Modified tongue body shape to 1.50: sound was “ieh”

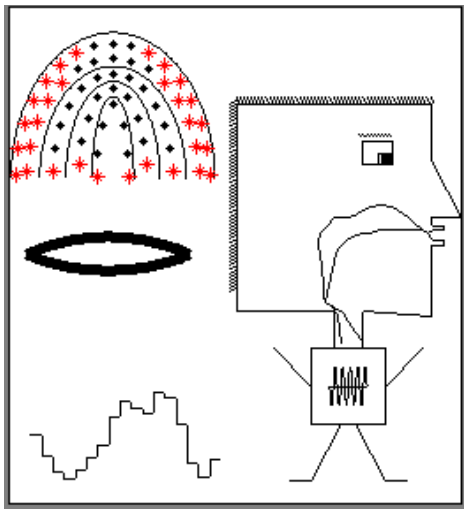


In part 2 of the lab we looked at the inverse controller in the DIVA model. This is the feature of DIVA that allows it to generate a sound by being given a set of coordinates in acoustic space. The model achieves this by learning the spatial relationships between different formant values during babbling.

- The “oow” sound was produced at Jaw Height: 0.85, Tongue Body Position: 1.72, Tongue Body Shape: -2.75, Tongue Lip Position: 2.42, Lip Protrusion: 0.51, Larynx Height: -2.45, Upper Lip Position: -0.99, Lower Lip Position: 0.99.

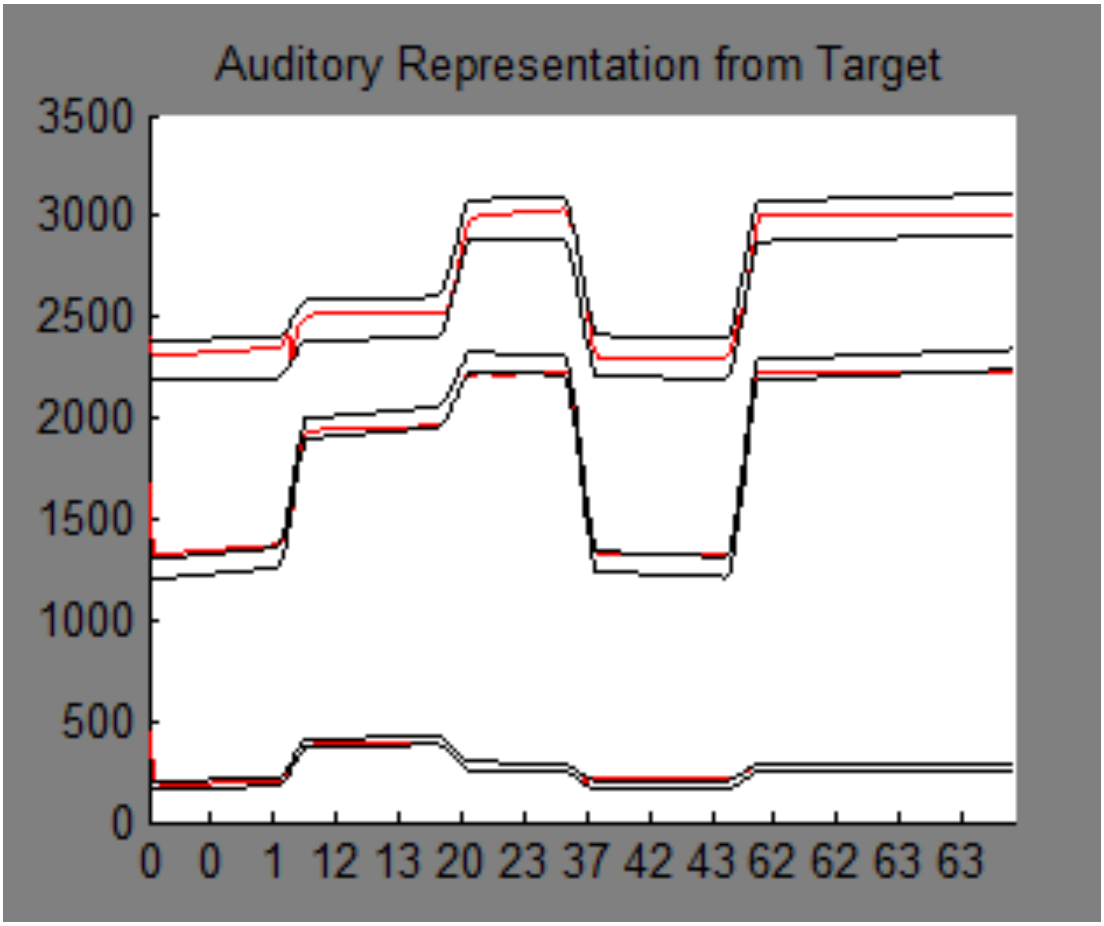


- The same sound, “oow”, was produced when the starting position was changed, but the vocal tract was slightly different; the tongue was slightly lower, and the lips where more tightly shut in the latter. The reason for this slight difference is that the positions, while still making effectively the same sounds, were easier/faster to reach from the different starting positions



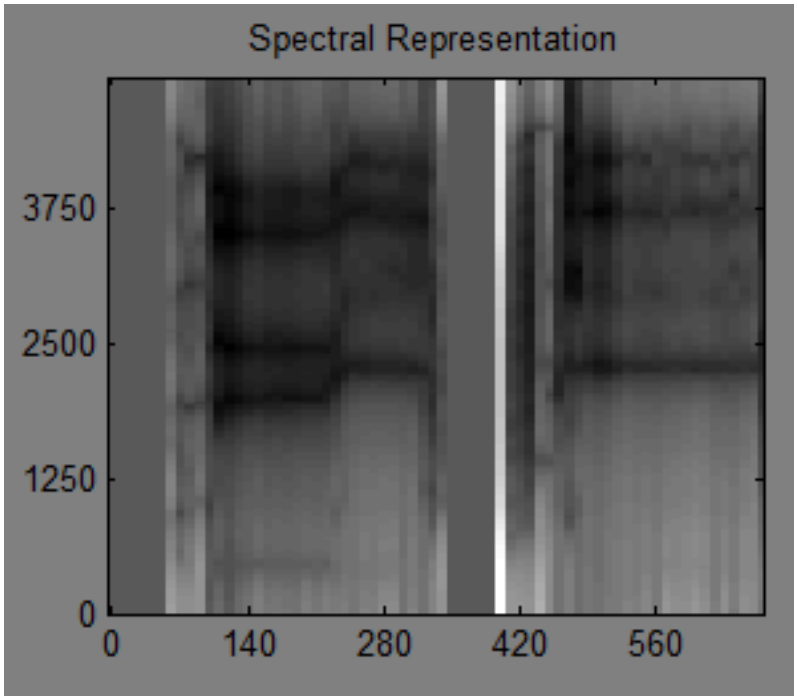
In part three of the lab, we looked at the feed forward control in the DIVA model. This feature allows for the creation of spoken words that have been practiced and learned.

- The lips open to form the “b” sound, and stay open for the following vowel sound. The vocal tract is mildly open (line in middle of chart) for the “b” sounds with the tongue high and the mouth open. For the vowel sounds the the tongue moves forward and the base of the tract expands. The “a” has higher tongue than the “y”. Based on the produced graph, which shows the spectrograph of the produced sound falling inside the accepted ranges, the model did a good job at producing the correct sound.



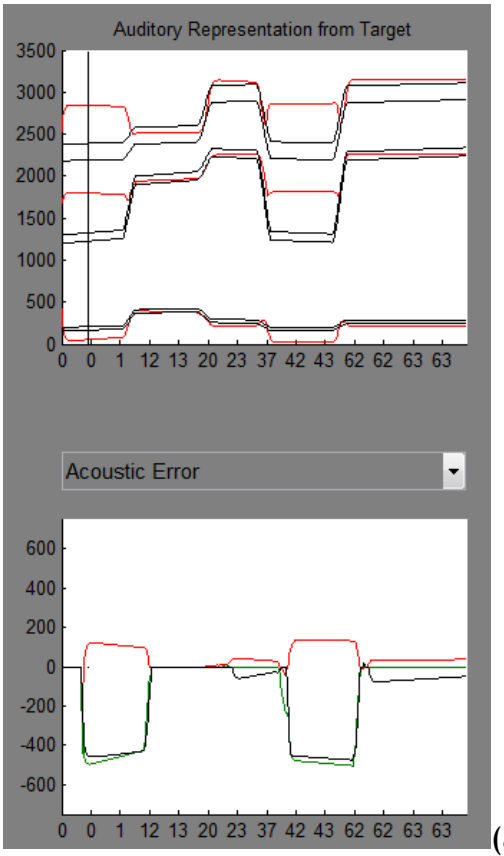
- The lines jump back and forth around 0, never going higher than .2/-.2, for somatosensory error. For acoustic error, the lines are steady around 0 for the entirety of the sound, except a dip to -400 at the very beginning in end of the channels. The DIVA model works on a negative feedback loop, so when there is a non-zero acoustic or somatosensory error it causes the model to readjust to negate the error.

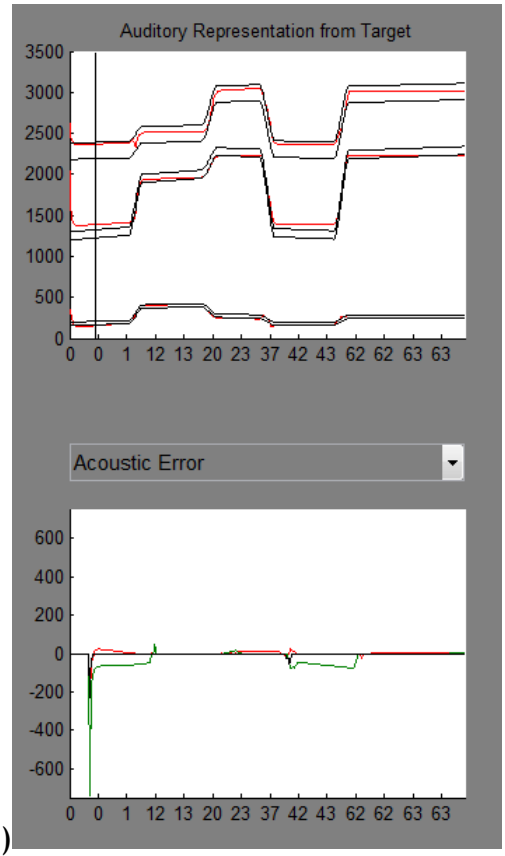
- In the regions of the produced spectrogram that are clearly shown (when the mouth is not completely closed for the “b” sound), the graph is very similar to the Auditory Representation from Target graph. They both show the two groups of two step like bands followed by straight bands, and both have relatively straight bands at the bottom, though it is very faint in the Spectral Representation.



In part four of the lab we looked at the process by which DIVA can learn words. We retrained the model to learn the word ‘baby”, the word used in the previous section.

- After 2 practicing the production of the word “baby” twice the Auditory Representation from Target graph shows that the produced sound is generally close to, if not within, the desired regions. This is with two main exceptions though. It is at theses two points where there is a major acoustic error shown in the lower plots.



(above)


(below)

- With more learning, the more the error decreases and the produced word sounds more accurate. The error though, does stay high at the very beginning of the sound production.

Questions:

1) M(t) is the current position, M(0) is the initial position, aff∫Mff(t)g(t)dt is the total feedforward signal over a given time and afb∫Mfb(t)g(t)dt is the total feedback signal over a given time.

2)

Conclusion:

In this lab we examined the DIVA model, and learned its capabilities. We looked at how it can produce sounds based on a given vocal tract formation, how given a sound it can create a vocal tract formation, and how it is able to learn speech sounds. In doing this we related the output of the software back to the principles (the feedback loop) that the model uses to create given sounds.