1. INTRODUCTION

Anyone who has taken a language class at some point in their life has likely used the element of music to at least some degree throughout their learning of the target language (TL), so common has it become in language-learning curricula across the globe. As early as the mid-twentieth century, musical language-learning tools have been implemented with the intention of helping with memorization of features specific to the TL, ranging from individual sounds to colloquial phrases (Altessia 2022 and Engh 2013). Often, they are deemed quite effective to this end, which should not come as much of a surprise, considering the broad span of similarities between language and music. Indeed, when attempting to define music, ethnomusicologists often speak using linguistic terminology, arguing that like human languages, varieties of music constitute "self-contained systems" that can be both studied in terms of a broader social context as well as parsed into individual articulatory features (Nettl 2005:51). There is also published evidence from the field of psychoacoustics that goes to support this connection, one notable example being the theory of ABSOLUTE SPECTRAL TONE COLOR (ASTC). Introduced to the sound and voice sciences by voice pedagogist and researcher Ian Howell, ASTC demonstrates that humans conceptualize vowels as pitches by asserting that each frequency (i.e. pitch) within the human hearing range has a vowel-like "color", a color that remains constant throughout any other acoustic modifications that may be applied to it – just as a musical note played across instruments may vary in its timbre while still sounding the same note (e.g. a middle C being played on a string bass versus being played on a flute) (Irene & Harris 2022). However, while it may be considered common knowledge that spoken language and music production overlap in their utilization of both the physical and perceptual aspects of sound, there appears to be less known as to just how much having an affinity for processing music can help (or hinder) a person's ability to process language. It is from this region of uncertainty that my ideas for the following study spawned.

My primary intention for completing this project was to observe the extent to which individuals can perceive and reproduce English vowels as represented by SYNTHESIZED VOWELS – that is, composite pure tone frequencies generated by Praat computer software.[1] What I soon

became more curious to discover is whether there are any similarities, differences, or other patterns evident among individuals' perceptions when compared to not only the synthesized vowels themselves, but also individuals' own attempted productions of the synthesized vowels, and whether these patterns may be dependent on the individuals' level of musical affinity.

Using my own musical background as a baseline, I predicted that the higher level of musical background one has, the more likely their perception of a synthesized vowel is to align with their own production of the vowel, regardless of whether they perceive it as the vowel it was actually designed to emulate. This is by the rationale that a higher degree of musical skill generally includes a stronger ability to decipher and closely simulate pure tones. By this same token, I also predicted that even if a person with a higher degree of musical affinity can reproduce what they think they are hearing with more accuracy, it will be more difficult for them to accurately identify the sounds as the natural vowels that they are designed to represent. In the latter case, I proposed that a person's musicality would work against them in the sense that, because of their intensive training in homing in on individual tones within the context of music (i.e. having been conditioned to identify tones as musical pitches rather than specific spoken vowels), musicians may find it more difficult than non-musicians to perceive the synthesized vowels as the spoken vowels that they are intended to resemble.

2. METHODS

To accomplish my goal of observing individuals' ability to recognize and produce English vowels out of composite pure tones, I delivered a study wherein I first asked a variety of volunteers to listen to three different sounds, each created as an audio simulation of a spoken vowel. Afterwards, I asked volunteers to first determine and then attempt to produce the sound that they heard while I recorded their responses.

I was able to collect data from a total of twenty-one participants; however, in constructing my analyses, I ended up using only a portion of what was collected due to the

suspected impact of experimental error on some of the participants' results (see DISCUSSION AND CONCLUSION). Initially, I had intended to recruit the participants such that one third of them would have had some degree of advanced musical study (referred to as FORMAL MUSICIANS); another third would have had experience playing music, but not any background formally studying it (referred to as INFORMAL MUSICIANS); and the final third would have had no experience playing music nor any background in musical study (referred to as NON-MUSICIANS). While I did end up collecting data from an equal amount of formal and non-musicians, I was only able to retrieve data from a slightly lower number of individuals who fell into the informal category (five speakers as opposed to eight each for the formal and non-musician groups) due to time constraints and lack of accessibility to qualified participants. The category that each participant fell under was determined using knowledge of their occupation and/or field of study as well as the nature of the participants' identified extra-curricular interests – specifically, whether music is included. As far as other speaker characteristics are concerned, ages ranged from nineteen to sixty years with a mean age of about twenty-five years. Out of all who participated, about fifty-two percent identified as female. Neither age nor gender was considered as a potential affective variable on the results of this study. I also did not account for any specific dialectal differences between speakers while collecting and analyzing data; however, I did ensure that every participant self-identified as a native English speaker.

Created using Praat, the sounds I provided for my study participants to hear were simply constructed combinations of PURE TONES layered on top of one another. The pure tones spoken of are equivalent to musical pitches of the same frequency (e.g. middle C, E, and G on a piano). Depending on the relationships between them when played simultaneously, these tones have the potential to create a variety of composite sounds, ranging from musical chords (e.g. C major and C minor) to spoken vowels (e.g. the vowels present in the English words "beet" and "bat"). Like musical chords, vowels are constructed of multiple tones of specific frequencies layered on top of one another, with the most observable frequencies being referred to as FORMANTS. Which frequencies are labeled as formants depends on a variety of factors, including the shape of the human vocal tract upon the vowel's production (SoundBridge 2019). Conversely, it is a vowel's empirically prescribed formant values that help distinguish it from other phonemes, including other vowels as well as consonants. (see APPENDIX for a comparison between spectrograms of

musical chords and spoken English vowels, as well as a comparison between the spectrogram of a vowel with that of a consonant). Due to Praat's lack of humanistic qualities in its production of sound, combining the pure tones chosen for this study amounted to highly digitalized-sounding composite tones that were perhaps initially more representative of analog signals than spoken vowels. Nevertheless, I chose to create three of these so-called "composite" tones, each which is composed of individual pure-tone frequencies that correlate with the spoken vowel it is intended to emulate. To create each composite tone, I started by generating three "new" sounds in Praat, choosing the option of "Create sound as pure tone". Upon selecting this option, I was prompted to edit the acoustic features of the tone, including the number of channels used in playback, start and end times, sampling frequency, tone frequency, amplitude, and fade-in/fade-out duration. For the purposes of this study, the only property I chose to manipulate was the tone frequency, which I adjusted to reflect the formant frequencies recorded in prior research for the vowels I chose to emulate (EdUHK 2021[2], Piché 1994-97, and SoundBridge 2019). Out of all the formant values available, I used only the first three formant values listed, as it is these formant values that are known to characterize a vowel the most prominently and consequently help distinguish it from others (EdUHK 2021). After generating the three individual tones according to the frequencies from the referenced formant list, I combined the three by creating a new sound file, done by simply highlighting the three tones and selecting Praat's "Combine to stereo" option. This ultimately created three digitalized composite sounds meant to emulate naturally spoken vowels, including the high front unrounded vowel /i/, the middle back rounded vowel /o/, and the low front vowel /æ/ (see APPENDIX for spectrograms of the three simulated vowels, as well as a table showing each vowel's first three formant frequency values). I chose to emulate these particular vowels mainly on the basis of their variation relative to each other in terms of articulation, with each accounting for a different region of the vocal tract (see APPENDIX). I thought that including as wide as an articulatory variety as possible when playing the study sounds for the participants would yield the best results in terms of the participants' likelihood of perceiving the sounds as differing in quality, thereby presumably increasing their likelihood to assign vowel qualities to them.

Before listening to the study sounds, each participant was given a document with a list of common English vowels and instructions to indicate which sound they heard for each sound that

is played (see APPENDIX). The list of vowels also included an "other" (fill-in-the-blank) option as well as an "unsure" option. Each participant was also given additional oral instructions before beginning the activity: after assuring that they had received the aforementioned document, I explained that I would play three different sounds for them, and after each sound, I would ask them to select a sound they thought they heard using the document as a reference. I also let them know that after selecting a sound, they would then be prompted to produce the sound that they heard as I recorded them using the sound recording function of the virtual meeting space Zoom.[3] A few additional details that I mentioned to almost every participant included the assurance that they would be given unlimited listens for each sound; no choice that they made would be considered "right" or "wrong"; and upon producing the sounds, to not feel compelled to exactly replicate the sound, but rather produce whatever they thought the sound might be trying to emulate, encouraging them to use their normal speaking voice. After all three sounds were played and the participant's attempts to produce each sound were recorded, the participant was asked to send their document with their written choices for each sound back to me for data input.

3. DATA AND RESULTS

To analyze speakers' ability to accurately identify English vowels from the frequency composition of the three synthesized vowels, comparisons were drawn between sounds that participants perceived/produced and the actual sounds that were emulated. Cases wherein participants' perceptions did not match with their own productions of the sounds were addressed as well, leaving room for the consideration of any possible connections between an individual's ability to accurately identify the target sound and their ability to produce what they perceive.

3.1. IDENTIFICATION OF SOUNDS

			Synthesized Vowel 1: /i/			
Sound	**Perceived vs. Produced**	**Total [Formal Musicians]**	**Total [Informal Musicians]**	**Total [All Musicians]**	**Total [Non-musicians]**	**Total [All Participants]**
/i/	Perceived	3	3	6	4	10
	Produced	3	3	6	5*	11
/u/	Perceived	2	1	3	---	3
	Produced	2	1	3	---	3
/e/	Perceived	---	---	---	2*	2
	Produced	---	---	---	1	1
/ʌ/	Perceived	1	---	1	---	1
	Produced	1	---	1	---	1
/ɑ/	Perceived	---	---	---	1	1
	Produced	---	---	---	1	1
/ɪ/	Perceived	---	---	---	1	1
	Produced	---	---	---	1	1
?	Produced	---	---	---	1	1

TABLE 1. Study participants' identification of synthesized vowel /i/. The asterisk * reflects one case wherein a participant's perception did not match their own production: **/e/ (perceived sound) → /i/ (produced sound)**

As indicated by TABLE 1, most study participants both perceived and produced Synthesized Vowel 1 as the sound it was designed to resemble, the high front unrounded vowel /i/. The sound with the second highest perception and production cases for this synthesized vowel was the vowel /u/. In these cases, the height of the synthesized vowel was perceived correctly, but backness was not; rather, it was both perceived and produced as a high-back as opposed to a high-front vowel. This could be a result of trying to mimic the sound too closely, resulting in the production of a rounded tone (no rounded high-front vowel exists in English). Three participants perceived and produced Synthesized Vowel 1 in this way, all falling within either the formal or informal musician category. The spectrogram below shows one of these participant's Synthesized Vowel 1 production.

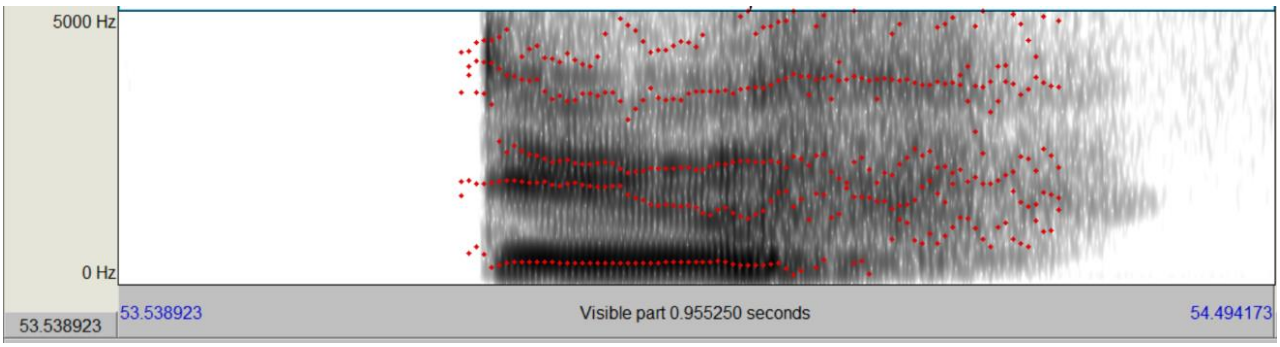


FIGURE 1. Synthesized Vowel 1 production: /u/

Other than the three participants who perceived and/or produced Synthesized Vowel 1 as /u/, the predominant response among both the formal and informal musician categories was the vowel it was intended to emulate, with about sixty percent perceiving and/or producing the sound as /i/. In the non-musician group, exactly fifty percent of the participants identified the sound as /i/, with the remaining responses appearing more varied (i.e. including one identification each for the vowels /ɑ/, /e/, and /ɪ/, in addition to one ambiguous production indicated by "?" in TABLE 1). It may be worth considering how despite this slightly higher variation, several participants in the non-musician category still managed to closely match their productions of the synthesized vowel with their own perception of it, even if their perception didn't quite match what was actually provided.

There was only one case wherein a perception of Synthesized Vowel 1 differed from not only the target vowel that it was designed to represent, but also the participant's own production of the synthesized vowel (see SPECIAL CASE: PERCEPTION → PRODUCTION DISCREPANCIES BETWEEN ALL THREE SYNTH. VOWELS). This has been attributed to a lack of understanding on the part of the participant for what was expected of them, which may ultimately constitute as an experimental error (see respective section, as well as POTENTIAL EXPERIMENTAL ERRORS).

Synthesized Vowel 2: /o/						
Sound	**Perceived vs. Produced**	**Total [Formal Musicians]**	**Total [Informal Musicians]**	**Total [All Musicians]**	**Total [Non-musicians]**	**Total [All Participants]**
/ɪ/	Perceived	1	2	3	1*	4
	Produced	1	2	3	1*	4
/i/	Perceived	2	1	3	---	3
	Produced	2	1	3	---	3
/ɑ/	Perceived	1	---	1	1	2
	Produced	1	---	1	1	2
/u/	Perceived	1	---	1	---	1
	Produced	1	---	1	---	1
/ʌ/	Perceived	---	---	---	1	1
	Produced	---	---	---	1	1
/e/	Perceived	---	---	---	1	1
	Produced	---	---	---	1	1
/æ/*	Perceived	2	---	2	1	3
/ɛ/*	Perceived	---	---	---	1	1
/ẽ/*	Produced	1	---	1	---	1
/ɑ̃/*	Produced	1	---	1	---	1
/æ̃/*	Produced	---	---	---	1	1
/iʔ/*	Produced	---	---	---	1	1
?	Produced	---	---	---	1	1

TABLE 2. Study participants' identification of synthesized vowel /o/. The asterisk * reflects the cases wherein participants' perceptions did not match their own productions: **/æ/ (perceived sound) → /ẽ/, /ɑ̃/, or /iʔ/ (produced sounds)**; **/ɪ/ (perceived sound) → /æ̃/ (produced sound)**; and **/ɛ/ (perceived sound) → /ɪ/ (produced sound)**

The data representative of Synthesized Vowel 2 are highly varied, especially when compared with the data recorded for Synthesized Vowel 1. There is also a higher rate of perception-production discrepancies for Synthesized Vowel 2, with two of the individuals from the formal musician category producing something different – more nasalized, in both cases – from what they initially perceived and three of the individuals from the non-musician category producing a completely different sound from what they reported as having perceived. While no participant accurately perceived nor produced the sound that Synthesized Vowel 2 was designed to emulate (the mid-back rounded vowel /o/), a few came relatively close: a formal musician and a non-musician each perceived and produced it as the low back vowel /ɑ/, while another formal

musician identified it as the high back rounded vowel /u/, all three correctly interpreting the vowel's backness, but not the height nor roundness. When trying to determine a predominant sound selected by the participants for Synthesized Vowel 2, it can be argued that most individuals tended to perceive and/or produce a high front vowel (either /i/ or /ɪ/); however, this assertion might only reasonably apply to the formal/informal musician cohorts, as non-musicians were characterized with a slightly wider range of identified sounds, including the sounds /ʌ/, /ɛ/, and /e/ in addition to the sounds /u/, /ɪ/, /ɑ/, and /æ/, as well as the ambiguous sound "?".

			Synthesized Vowel 3: /æ/			
Sound	**Perceived vs. Produced**	**Total [Formal Musicians]**	**Total [Informal Musicians]**	**Total [All Musicians]**	**Total [Non-musicians]**	**Total [All Participants]**
/i/	Perceived	3*	---	3	2	5
	Produced	2	---	2	1	3
/ɑ/	Perceived	1	---	1	1	2
	Produced	1	1*	2	1	3
/u/	Perceived	---	---	---	1	1
	Produced	---	1*	1	1	2
/æ/	Perceived	---	---	---	1	1
	Produced	---	---	---	1	1
/e/	Perceived	---	---	---	1	1
	Produced	---	---	---	1	1
/ɪ/	Perceived	1	---	1	---	1
	Produced	1	---	1	---	1
/ʌ/*	Perceived	---	1	1	---	1
/ʊ/*	Perceived	---	1	1	---	1
/iʔ/*	Produced	1	---	---	---	1
/eɪ/*	Produced	2	---	2	---	2
/ɪŋ/*	Produced	---	---	---	1	1
Other: "hitting two spoons together"	Perceived	---	---	---	1	1
"Unsure" ?	Perceived	1	---	1	---	1*
	Produced	---	---	---	1	1

TABLE 3. Study participants' identification of synthesized vowel /æ/. The asterisk * reflects cases wherein participants' perceptions did not match their own productions: **/i/ (perceived sound) → /iʔ/ (produced sound)**; **"unsure" perception → /eɪ/ (produced sound)**; **/ʊ/ (perceived sound) → /u/ (produced sound)**; **/ʌ/ (perceived sound) → /ɑ/ (produced sound)**; and **/i/ (perceived sound) → /ɪŋ/ (produced sound)**

The data collected for Synthesized Vowel 3 constitute another case of extensive variation among participants in their perceptions and productions of the synthesized vowel, which in this case was the low front vowel /æ/. Only one individual both perceived and produced the vowel as such (a non-musician); their production of the vowel as compared to the actual formant structure of the synthesized vowel is shown in FIGURE 2 and FIGURE 3. The first two formant values of each have been included in the captions to reiterate the closeness with which this participant's production matched the target sound.

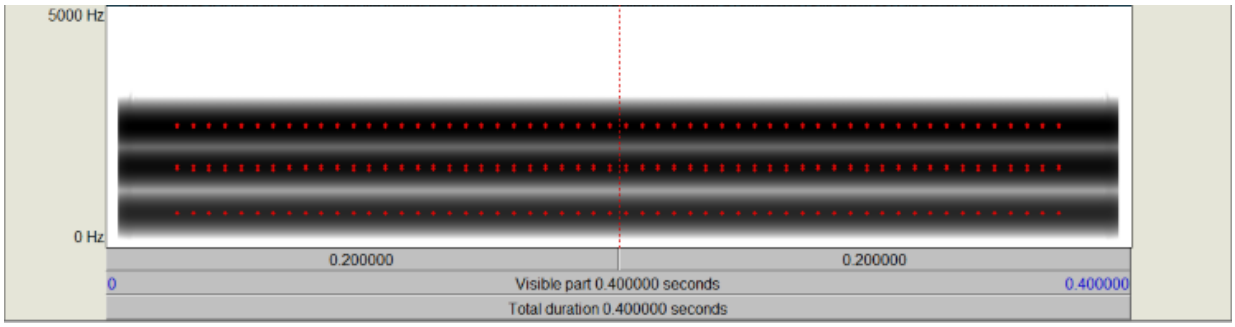


FIGURE 2. Synthesized Vowel 3: /æ/ (F1: 689 Hz; F2: 1582 Hz)

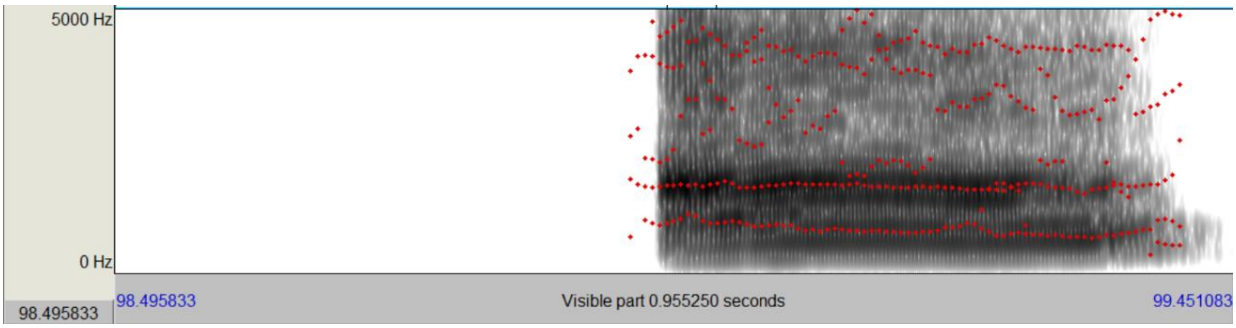


FIGURE 3. Synthesized Vowel 3 production: /æ/ (F1: ~976 Hz; F2: ~1650 Hz)

Notice that the greatest marked difference between the synthesized and participant-produced formants is just below 300 Hz, a difference that is relatively small compared to all other comparisons drawn.

Like Synthesized Vowel 2, Synthesized Vowel 3 resulted in some perceptions/productions that could be considered more accurate in terms of frontness (as in the case of the common perception of the high front vowel /i/ among participants) as well as height (as in the participants who perceived/produced the low back vowel /ɑ/). There were also two cases of diphthongization in the production of the target vowel. Both roughly constituted the diphthong /eɪ/ as in "ate", and both were produced by individuals in the formal musician category who did not have a clear idea of what they perceived the sound to be. A visual comparison between Synthesized Vowel 3 and the /eɪ/ sound produced by one of the participants is shown in FIGURE 4 and FIGURE 5.

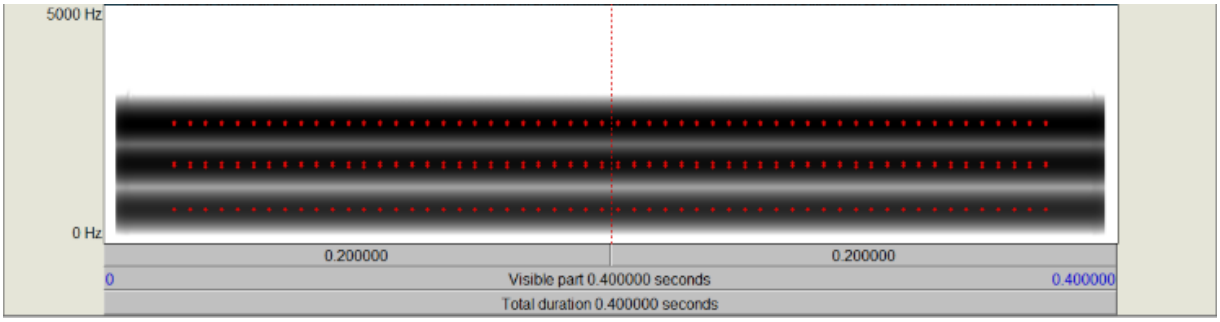


FIGURE 4. Synthesized Vowel 3: /æ/

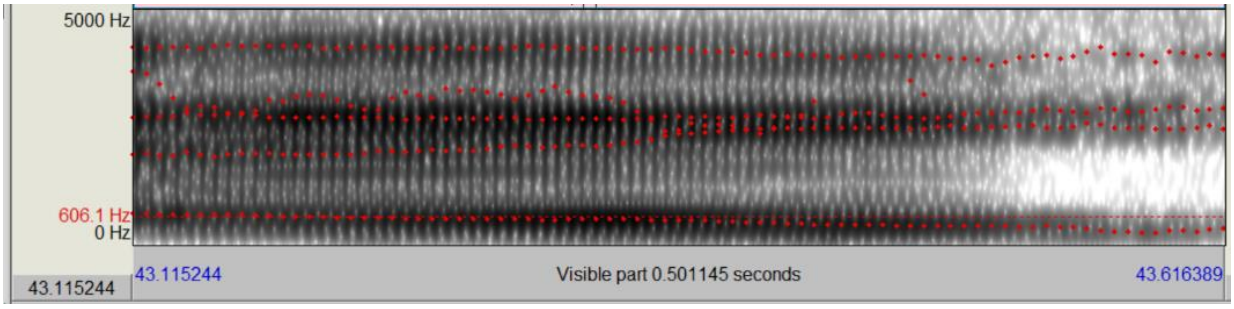


FIGURE 5. Synthesized Vowel 3 production: /eɪ/ (version 1)

This participant's production of Synthesized Vowel 3 clearly represents the diphthong /eɪ/, as shown by formants 1 and 2: 1 lowers slightly, indicating an increase in height, while 2 rises, indicating a slightly fronter articulatory position.

3.2. PERCEPTION → PRODUCTION RELATIONSHIPS

To measure the extent to which participants were able to produce what they thought they heard, comparisons were drawn between the sounds that they indicated to have heard in written form (using the options provided by the study activity document) and the sounds that they actually produced when prompted. Each participant's sound productions were analyzed using the spectrogram feature of Praat and were further compared to spectrograms and formant value tables retrieved from external sources to evaluate each recording of this study within the broader context of human vowel production as a whole – that is, the formant patterns empirically measured for each American English vowel (EdUHK 2021). Note that not all analyses of participant recordings are described here – just those that were deemed the most relevant based on the extent of their discrepancies/similarities with the participants' perceptions and/or the empirical formant analyses referenced above.

PERCEPTION → PRODUCTION DISCREPANCIES FOR SYNTHESIZED VOWEL 2

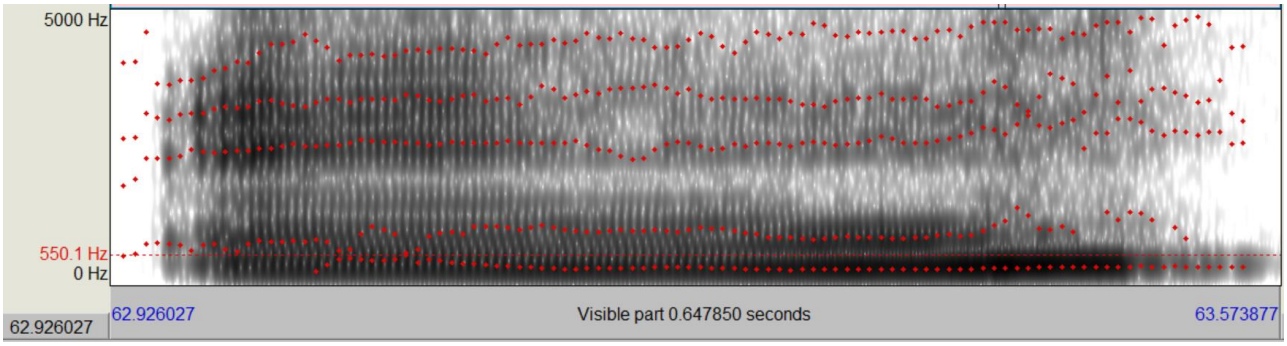


FIGURE 6. Synthesized Vowel 2 production: /ẽ/ (slightly nasalized)

When attempting to reproduce what they heard for Synthesized Vowel 2, this participant (of the formal musician cohort) began by describing the sound as "nasalized" before actually producing the sound. A nasalized quality does appear to be present in the production, based on the higher sporadicity of the shown formants. Otherwise, it resembles formant values empirically recorded for the vowel /e/ (EdUHK 2021) - a bit higher in terms of articulation than the /æ/ vowel that was reported as being perceived.

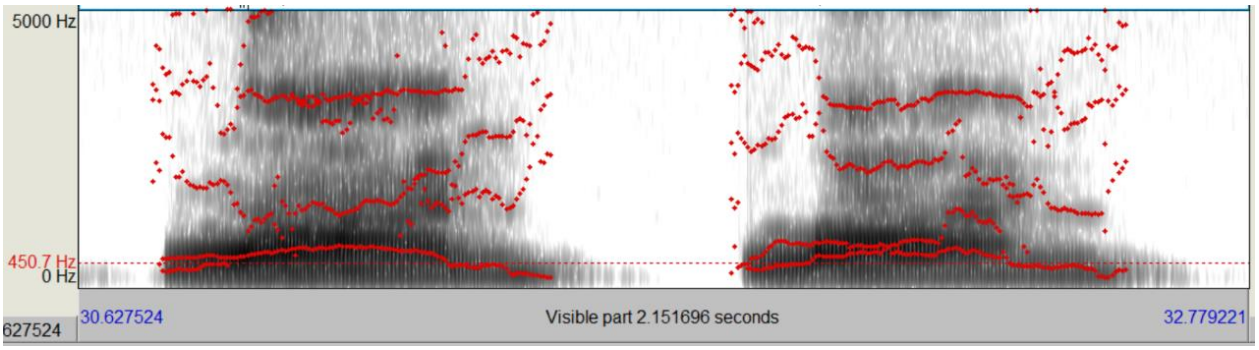


FIGURE 7. Synthesized Vowel 2 production: /ɑ̃/ (slightly nasalized)

Here is another case wherein a nasalized quality is evident in the production of Synthesized Vowel 2, indicated by more sporadically positioned formants. While this could possibly just be a result of a following nasal /n/ (this participant produced the target sound within the context of word "lawn"), it is worth noting that two people ended up producing more nasalized vowels for the second sound. For this participant (also part of the formal musician cohort) specifically, it is also interesting that the production of Synthesized Vowel 2 ended up being a slightly "backer" vowel than what was indicated as having been perceived (the low front vowel /æ/ - referred to EdUHK 2021 for formant comparison).

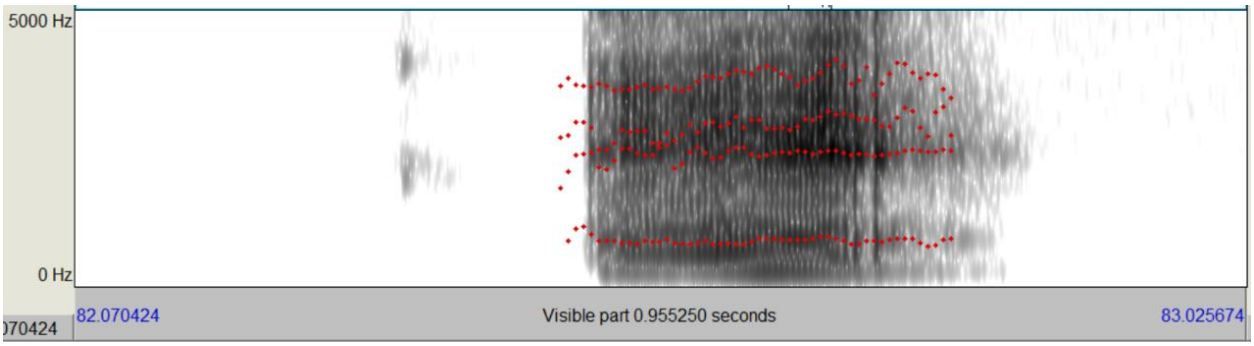


FIGURE 8. Synthesized Vowel 2 production: /æ̃/ (nasalized)

Here is a third case wherein Synthesized Vowel 2 is produced as nasalized, not to mention significantly lower than what was perceived (formant values of produced sound come closer to /æ/ vowel than /ɪ/ vowel perceived - referred to EdUHK 2021 for formant comparison). This participant was part of the non-musician cohort.

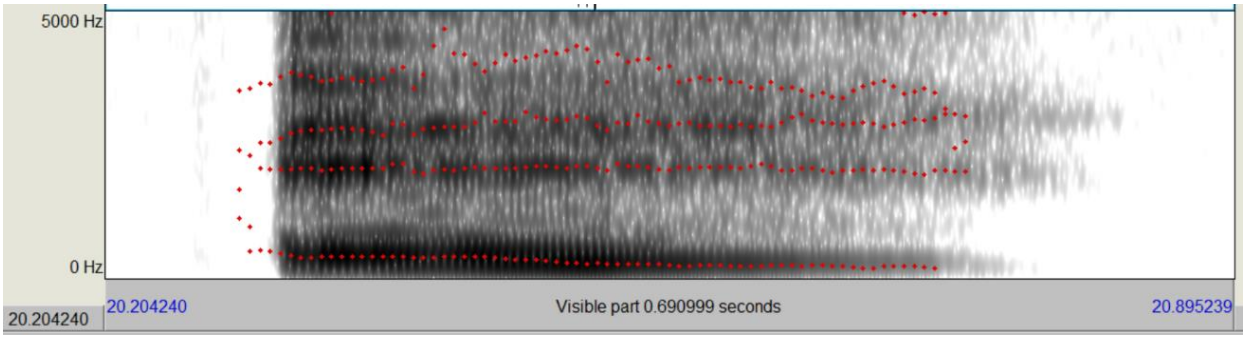


FIGURE 9. Synthesized Vowel 2 production: /i/

This participant's production and perception of Synthesized Vowel 2 differed mainly in terms of articulatory height (production more closely resembled the high front vowel /ɪ/ than the perceived mid-front vowel /ɛ/ - referred to EdUHK 2021 for formant comparison). Participant was also part of the non-musician cohort.

PERCEPTION → PRODUCTION DISCREPANCIES FOR SYNTHESIZED VOWEL 3

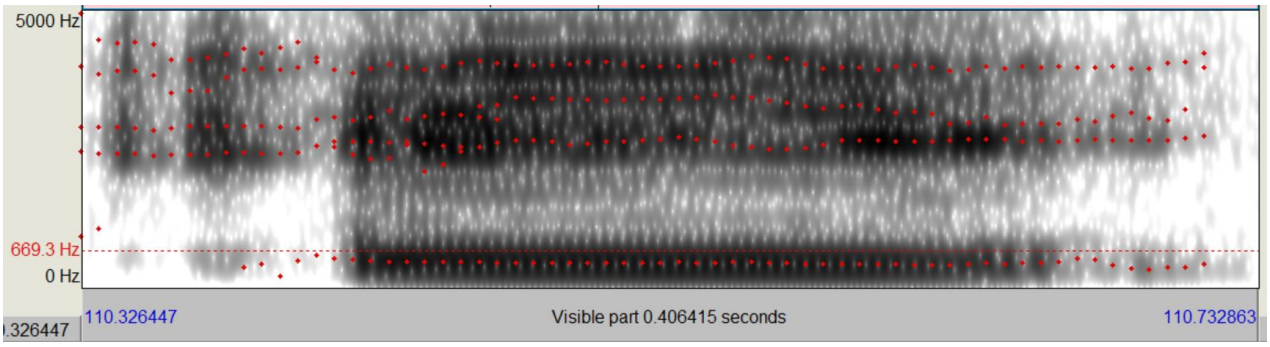


FIGURE 10. Synthesized Vowel 3 production: /eɪ/ (version 2)

While this participant's reported perception was unclear (chose "unsure" option), their production seemed to resemble the diphthong /eɪ/. While the formant structures predominantly resemble those typically measured of the vowel /ɪ/ (EdUHK 2021), the first formant seems absent at the onset of the vowel (interpreted as about the 110.326 sec. mark), leading to the

possibility of an articulatory position that produced a slightly lower vowel (such as /e/). Participant was part of the formal musician cohort.

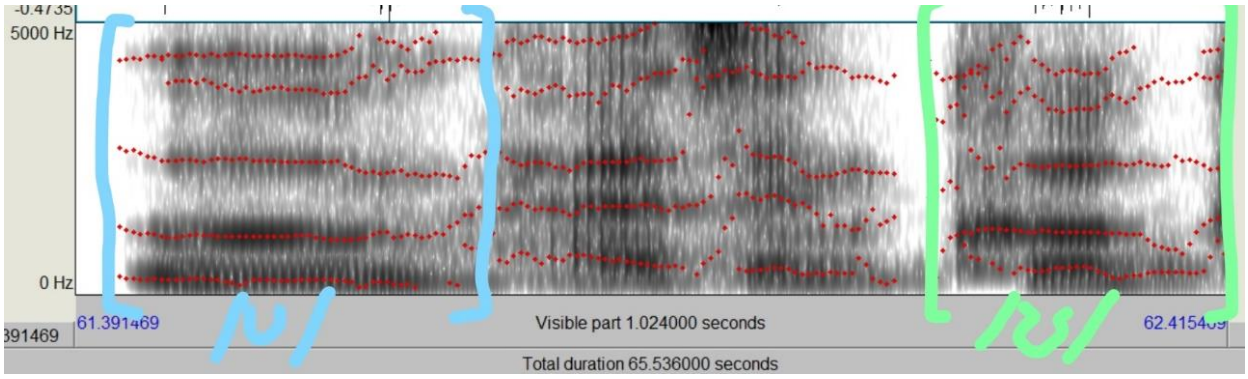


FIGURE 11. Synthesized Vowel 3 productions: /u/ and /ʊ/

While this participant (an informal musician) initially produced Synthesized Vowel 3 as /u/ (slightly differing from the /ʊ/ sound indicated as being perceived through its higher frontness and lower height), they then repeated it within the context of the word "cook", resulting in the actual production of the vowel /ʊ/. Despite this occurrence, the participant seems to have conceptualized both sounds as the same in both perception and production.

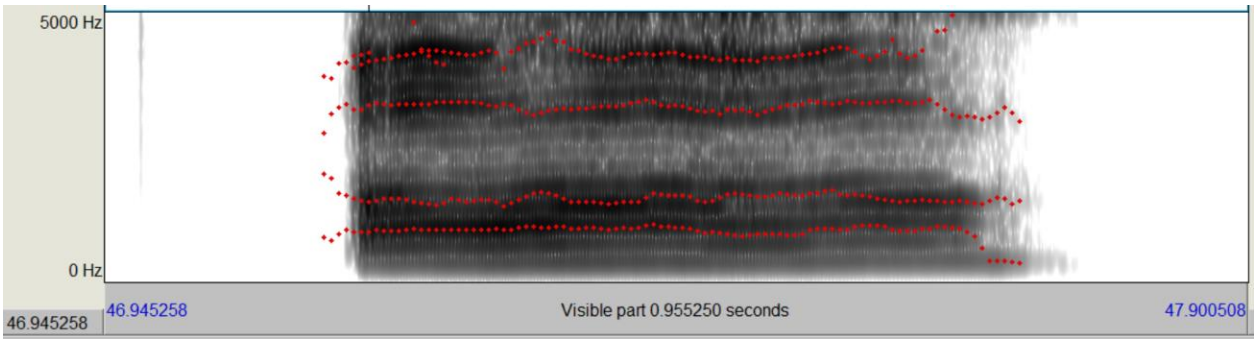


FIGURE 12. Synthesized Vowel 3 production: /ɑ/

This participant's production of Synthesized Vowel 3 appears slightly backer and lower than perceived (more closely resembles the sound /ɑ/ rather than the /ʌ/ sound perceived, as compared to formant values prescribed by EdUHK 2021). It could be argued that this is a result

of the participant using their singing rather than speaking voice when producing the sound, leading to the question of whether pitch and/or tone quality can affect perception/production discrepancies. Participant was part of the informal musician cohort.

SPECIAL CASE: PERCEPTION → PRODUCTION DISCREPANCIES BETWEEN ALL THREE SYNTHESIZED VOWELS

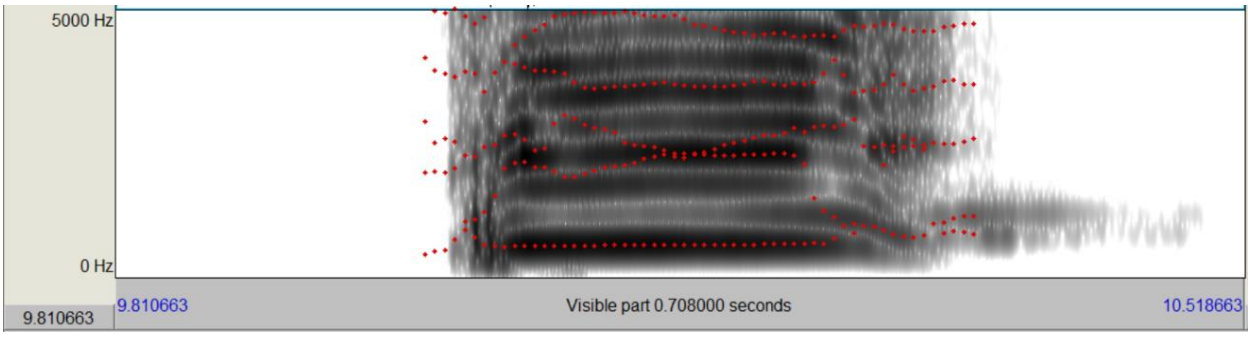


FIGURE 13. Synthesized Vowel 1 production: /i/ (no glottal stop)

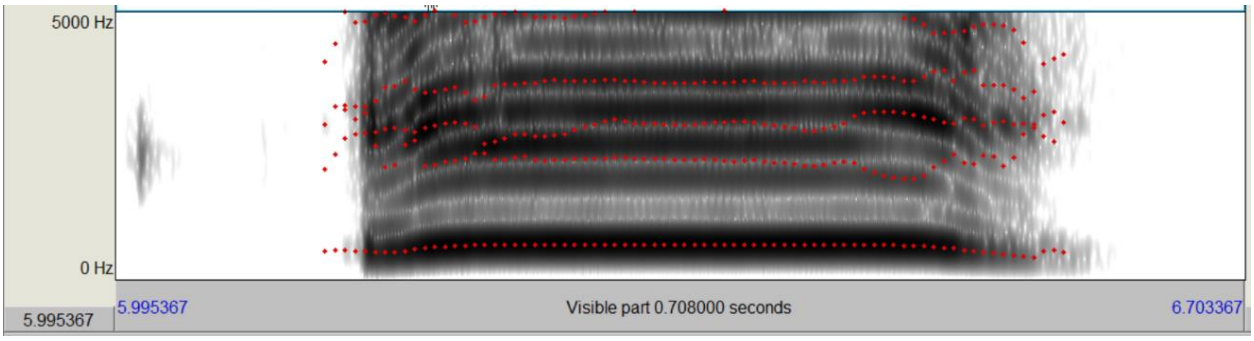


FIGURE 14. Synthesized Vowel 2 production: /iʔ/ (with glottal stop)

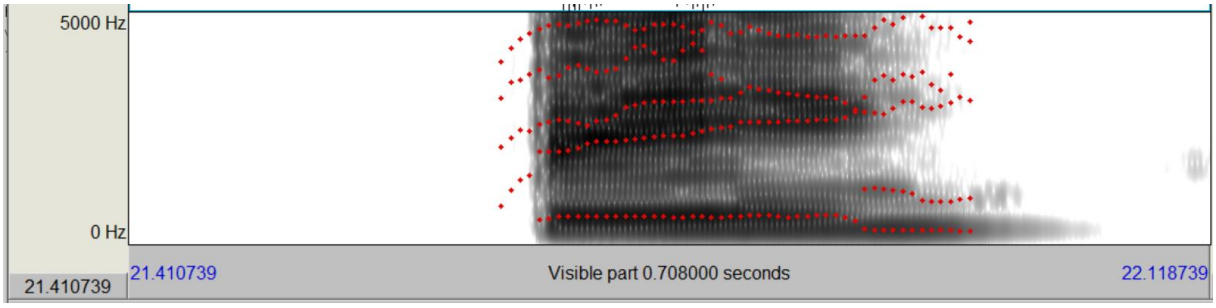


FIGURE 15. Synthesized Vowel 3 production: /ɪŋ/

Here is an example of a case characterized with extreme differences between what was heard and what was actually produced for all three synthesized vowels, the only case of its kind in this study. The participant (a non-musician) produced Synthesized Vowels 1 and 2 as significantly higher/fronter than what was perceived (the vowels /e/ and /æ/, respectively), which is suspected to have most likely been a result of misunderstanding/miscommunication of the study document's sound list and how its options may relate to the sounds played. While this participant's production of Synthesized Vowel 3 came closer to what was perceived (being the high front vowel /i/), there was a more explicit nasal feature present that was not included in the reported perception. This nasal quality differs from that apparent in other participants' productions in the way that the vowel formants are mostly steady (rather than sporadic) until the back end of the production, indicating a clearer phonemic distinction (i.e. presence of a full-fledged nasal consonant).

PERCEPTION → PRODUCTION CORRELATION

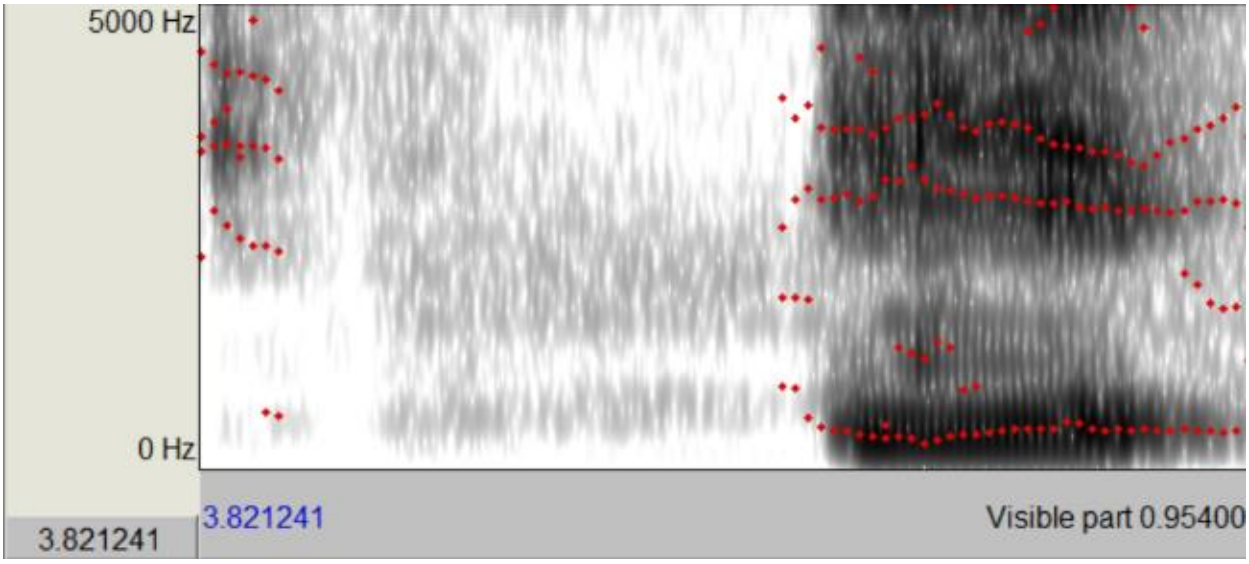


FIGURE 16. Synthesized Vowel 1 production: /i/

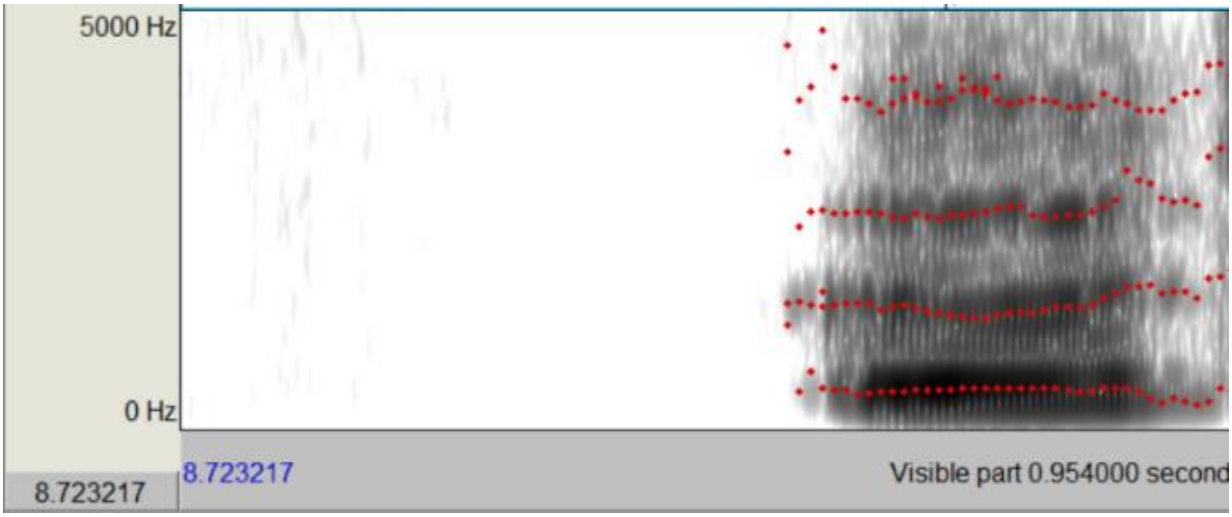


FIGURE 17. Synthesized Vowel 2 production: /u/

FIGURE 18. Synthesized Vowel 3 production: /ɪ/

Out of all the formal/informal musicians that participated and whose data could be used (i.e. those unimpeded by experimental error), only one produced all three synthesized vowels as closely matching those that they indicated as having perceived, as indicated through the comparison made between the produced formant values with formant values prescribed for the perceived vowels (EdUHK 2021). The other three cases wherein all perceptions matched with productions involved participants in the non-musician category.

4. DISCUSSION & CONCLUSION

4.1. SUMMARY OF DATA/OBSERVATIONS WITHIN CONTEXT OF HYPOTHESES

Ultimately, there seem to have been more instances of close correlation between sounds as they were perceived and produced among the non-musicians as opposed to those with some degree of musical background. While this refutes my hypothesis regarding the nature of this relationship, it does make sense considering the extent to which almost every participant in the formal and informal musician categories seemed to have made more of an effort to mimic the sound that they heard, regardless of the vowel that they may have initially perceived. Due to the limit that both English and IPA orthography impose on what can be transcribed within the context of natural human speech, it is understandable if any additional details an individual may try to implement when actually producing what they hear differentiate from what they claim to have perceived using the limited letters/symbols they have been provided. What is perhaps even more fascinating to consider is how in many cases, the participants who did appear to add more acoustic features to their sound productions than they described as hearing appeared to have done so unconsciously, perhaps further endorsing their affinity to decode sounds from a perspective centered around musicality as opposed to linguistic meaning. Another observation worth mentioning is the fact that any discrepancies in sound production as compared to sound perceptions among participants in either of the two musician categories appear to be a result of such attempts to add to the perceived sound, as opposed to resulting from a complete misinterpretation of the sound, which seemed more likely to occur among the non-musicians.

When reviewing the rate at which participants accurately identified the synthesized vowels as the vowels they were designed to emulate, one notable observation that can be made is the fact that out of the three synthesized vowels, Synthesized Vowel 1 (/i/) ended up with the highest rate of accuracy, as well as having the fewest occasions of perceptions that did not align with productions. This may indicate that there is a direct relationship between the number of accurate identifications of a synthesized vowel and the frequency at which it is produced as it is perceived.

Because of the extent of variation among participants from all three cohorts in their perceptions and productions of the synthesized vowels, it is difficult to come to a defendable conclusion regarding the effect of one's musical affinity on their ability to accurately identify the synthesized vowels as the spoken vowels they were designed to emulate. However, when considering the proposed negative relationship between musical affinity and perception-production correlation among individuals, as well as the apparent tendency for individuals to more accurately perceive and produce the target sound when their perception aligns with their production, it does appear more likely that a person with a higher degree of musical background may experience more difficulty in attempting to identify a vowel from a set of raw, pure tone frequencies alone. This supports the theory that I used to formulate my initial hypothesis regarding the effect of musical affinity on sound identification accuracy – that is, due to the extensive audiation training that many formal (and perhaps some informal) musicians have, their ability to simplify composite pure tone frequencies within a context other than music becomes compromised as too often they become stuck on the acoustic properties and therefore oblivious to the linguistic implications. Such conclusion may also be supported by previously conducted psychoacoustic research mentioned in the introduction of this paper – in essence, the theory of Absolute Spectral Tonal Color. While asserting the notion that individual vowels can be assigned to specific pitches (i.e. frequencies), the theory also illustrates that despite each frequency having a vowel-like color, these vowel-like colors aren't perceived as the same as the composite vowel actually being produced (Irene & Harris 2022). Thus creates an auditory illusion directly involving the mismatching of vowel production and vowel perception, a phenomenon that appears to be central in the results of the present study.

4.2. POTENTIAL EXPERIMENTAL ERRORS

STUDY TASK DESIGN: INSTRUCTION DELIVERY

As indicated in the case wherein all three synthesized vowels were perceived and produced with discrepancies, there appears to have been a mis- and/or lack of communication between me and at least one of the study participants – primarily regarding how/what to produce upon being prompted to reproduce what they thought they heard. As the study progressed and I gained a better understanding of how people would typically respond to the task at hand, I attempted to

improve/clarify the instructions – specifically to the effect of ensuring that when trying to produce the sounds, participants were aware that they could produce vowels as they would usually speak (as opposed to mimicking the synthesized sounds exactly as heard).

STUDY TASK DESIGN: LIST OF POSSIBLE SOUNDS

It was not until I had already begun facilitating the study task to participants when I realized that I had forgotten to include diphthongs in the list of sounds I provided for participants' reference on the study handout. Similarly, I realized that I could have also included other voiced sounds (sonorants, in particular). I am unsure as to whether this would have affected the overall patterns of perception-production correlations among participants or whether it would have helped them correctly identify the target vowels.

CONDITION OF AUDIO

Some participants ended up hearing at least one of the three sounds at a higher amplitude than the original amplitude of 70db due to reported volume issues. I am uncertain as to the exact source of the problem, especially since not everyone experienced the issue; with this in mind, I expect it might have been a result of variation in tech quality across participants. Any data collected from a situation involving an increase in amplitude/intensity were excluded from the final analysis.

4.3. STUDY IMPLICATIONS AND QUESTIONS FOR FURTHER RESEARCH

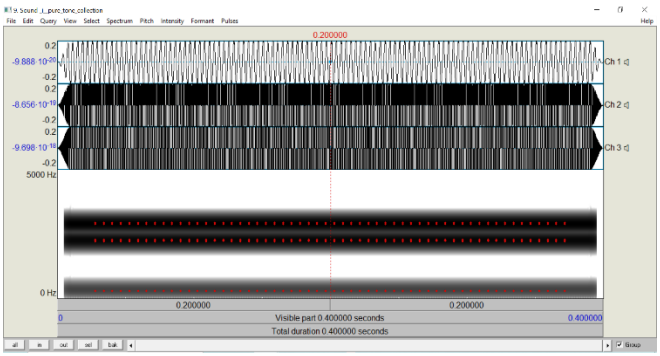
I am satisfied to acknowledge that the results of this study – while leaving some conditions of the measured relationships still up for interpretation – ultimately support that there is indeed a relationship between the way an individual processes the sounds of music and how they process the sounds of language. The implications for this assertion are evident not only within the field of linguistics as a whole, but perhaps specifically within the area of language acquisition, as it brings into question how to best treat language learners who may find it especially difficult to learn a language's sound system due to their stronger inclination towards processing individual speech sounds in terms of perceived "musical" attributes rather than phonemic meaning. The

results of this study have also left me with an assortment of further questions that might be worth addressing in the future, including:

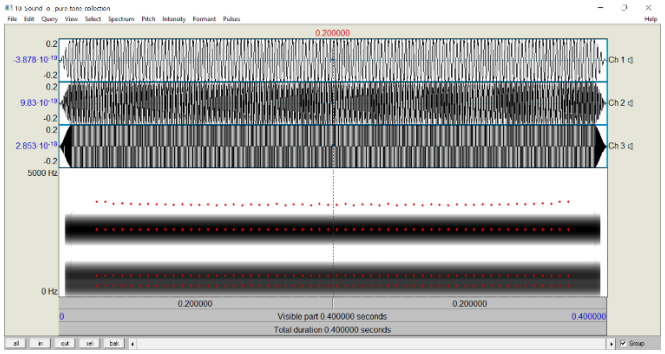
- Is there any relationship between an individual's ability to decipher vowel-like sounds from pure tone collections and their ability to identify musical pitches and/or chords from the same collections (designed like the ones used in this study)?
- To what extent might other acoustic attributes of the composite pure tones of a sound affect perception/production (e.g. amplitude/intensity, duration, sequence, etc.)?
- Does dialect play a role in a person's ability to process synthesized speech? What about fluency in other languages (including cases wherein English is not an individual's first language)?
- Are there similar relationships evident between individuals not only of languages other than English, but also who come from non-Western music backgrounds?

APPENDIX

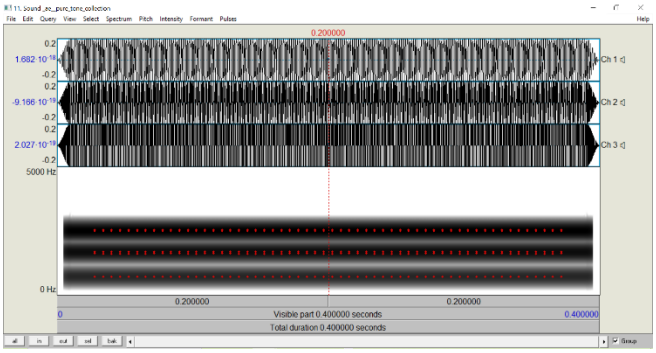
1. SYNTHESIZED, PRAAT-GENERATED VOWELS USED IN STUDY TASK



Synthesized Vowel 1: /i/ pure tone collection



Synthesized Vowel 2: /o/ pure tone collection



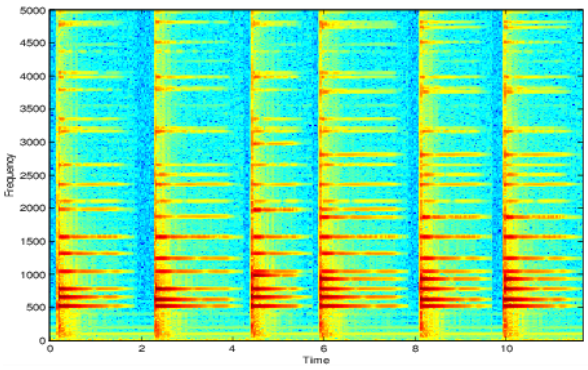
Synthesized Vowel 3: /æ/ pure tone collection

FORMANT VALUES USED TO CREATE SYNTHESIZED VOWELS, ADAPTED FROM EDUHK (2021), PICHÉ (1994-97), AND SOUNDBRIDGE (2019)

Synthesized Vowel	F1 (Hz)	F2 (Hz)	F3 (Hz)
/i/	280	2207	2254
/o/	408	803	2579
/æ/	689	1582	1660

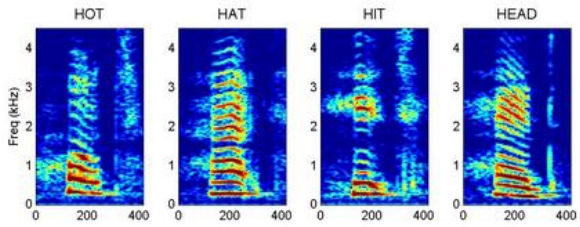
2. SPECTROGRAM COMPARISONS

The below spectrograms illustrate the similarities between sounds prescribed as musical chords (see figure 1) and sounds prescribed as vowels (see figure 2), as well as the difference between vowels and voiced consonants (see figure 3), all from a harmonic frequency perspective. Notice the common presence of layered frequencies in the productions of the chords and the vowels, with the dark red marks indicating the most resonant frequencies (i.e. formants). Also notice how in the third spectrogram, there is an absence of multiple resonant frequencies during the articulation of the two consonants /b/ and /j/, even though they are still considered voiced phonemes. Such emphasizes the importance of vowels' harmonic properties when distinguishing them from other voiced phonetic sounds.



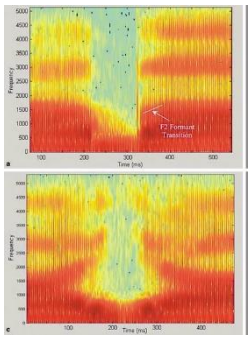
A spectrogram showing six musical chords (root C) played in succession on a piano. Chords from left to right are C, C minor, C major-seven, C seven, C minor-seven, C minor-seven.

https://www.researchgate.net/figure/Spectrogram-of-the-piano-chords-set-C-Cm-CM7-C7-Cm7-Cm7_fig2_228525894



Part of a spectrogram showing the production of four English words, each with a slightly different vowel at its core.

https://soundbridge.io/formants-vowel-sounds/

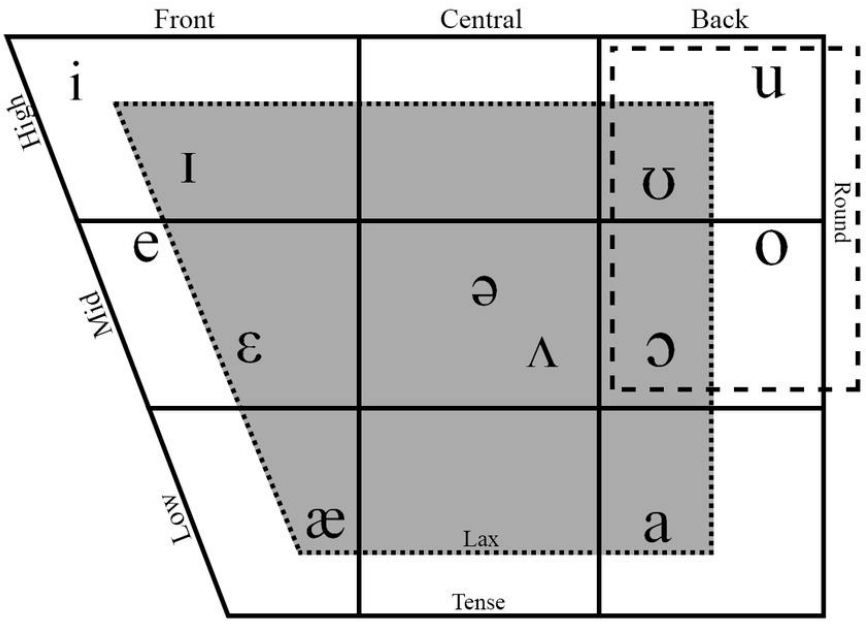


A spectrogram showcasing the presence of two different consonants in between the vowel /a/; the top contains the voiced plosive /b/, and the bottom contains the voiced palatal approximant /j/.

https://www.researchgate.net/figure/Two-examples-of-the-spectrograms-and-electrodograms-used-in-this-study-Panels-a-and-c_fig1_10624516

3. AMERICAN ENGLISH VOWEL CHART

Below is a diagram of the monophthongs of American English, organized relative to each vowel's articulatory position in the human vocal tract. These articulatory positions were considered when determining which vowels to simulate for the study (one vowel from each depicted height as well as one rounded back vowel were included in the effort to represent the most contrast).



https://www.researchgate.net/profile/Benjamin_Bay2/publication/317571199/figure/fig1/AS:504926465781760@1497395258567/For-rhyme-scoring-purposes-we-estimate-vowel-similarity-by-finding-the-distance-between.jpg

4. COPY OF VOWEL PERCEPTION STUDY HAND-OUT

Using the list below, please write what you hear for each sound that is played.

(a) /i/ as in ‘b<u>ee</u>t’ (b) /u/ as in ‘sh<u>oo</u>t’

(c) /æ/ as in ‘h<u>a</u>t’ (d) /o/ as in ‘<u>o</u>rchard’

(e) /ɪ/ as in ‘p<u>i</u>t’ (f) /e/ as in ‘c<u>a</u>ke’

(g) /ɛ/ as in ‘b<u>e</u>d’ (h) /ʌ/ as in ‘c<u>u</u>p’

(i) /ɑ/ as in ‘l<u>a</u>wn’ (j) /ʊ/ as in ‘c<u>oo</u>k’

(k) Other: _____ (l) Unsure

Sound 1: _____

Sound 2: _____

Sound 3: _____

REFERENCES

Altissia. 2022. Music as an effective tool for learning languages. *Altissia* online language-learning platform: https://altissia.org/music-as-an-effective-tool-for-learning-languages/.

EdUHK. 2021. 2.2 Formants of vowels. Online corpus-aided English pronunciation teaching and learning system: https://corpus.eduhk.hk/english_pronunciation/index.php/2-2-formants-of-vowels/.

Engh, Dwayne. 2013. Why use music in English Language Learning? A survey of the literature. *English Language Teaching* 6.113-27. Online: https://files.eric.ed.gov/fulltext/EJ1076582.pdf

Irene, Laurel, and Harris, David. 2022. Filtered listening of vocal regions. *Voice Science Works*. Online: https://www.voicescienceworks.org/filtered-listening-and-vocal-regions.html.

Nettl, Bruno. 2005. *The Study of ethnomusicology: Thirty-one issues and concepts* 2.51. Champaign: University of Illinois Press.

Otieno, Mark Owuor. 2017. What languages are spoken in Hong Kong? *World Atlas*. Online: https://www.worldatlas.com/articles/what-languages-are-spoken-in-hong-kong.html.

Piché, Jean (ed.) 1994-97. Table III: Formant values. *The csound manual (version 3.48): A manual for the audio processing system and supporting programs with tutorials*. Massachusetts Institute of Technology. Online: https://www.classes.cs.uchicago.edu/archive/1999/spring/CS295/Computing_Resources/Csound/CsManual3.48b1.HTML/Appendices/table3.html.

SoundBridge. 2019. Using formants to synthesize vowel sounds. *SoundBridge* blog, 7 July 2019. Online: https://soundbridge.io/formants-vowel-sounds/.

ENDNOTES

[1] Version 6.1.16 of Praat software, developed by Paul Boersma and David Weenink, was used for this study.

[2] One of the sources used in the creation of this study's digitalized sounds, EdUHK (2021), hails from an academic establishment in Hong Kong, where according to World Atlas (2017), the most spoken language is Cantonese. Despite this statistic, all the data reported on the establishment's website appears to be taken from American English speakers (EdUHK 2021). As this is the dialect around which the present study focuses, the author determined EdUHK to still be a credible source for retrieving phonetic information.

[3] All the participants except for one completed the study activity over Zoom; differences in modality and any impacts they might have had on aural perception were not accounted for in this study.