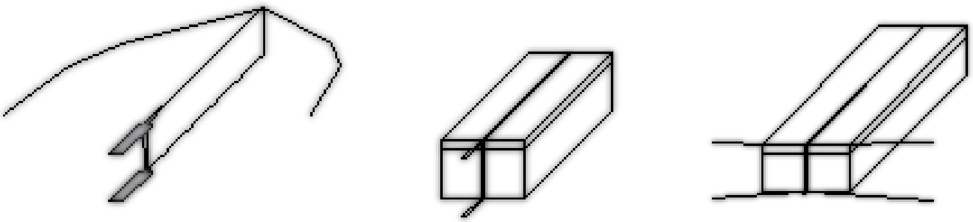


Synthetic and natural voice: An inquiry into sensing and perceiving vocality

Lawrence McGuire

bachelor's thesis — Institute of Sonology, 2026

Abstract

This project tackles the issue of describing, composing, and perceiving vocality in a synthetic context, highlighting an experiential approach to the perception of a vocal signal. The research primarily focuses on the idea of *fusions of sounds*, particularly fusions between synthetic and natural voice, where the resulting quality enriches a vocal experience through the ambiguities and multiplicities it brings forth. Design choices and aesthetical considerations of a computer program for vocal synthesis are then discussed in relation to my own approaches to vocal composition.

Acknowledgements

I would like to express my gratitude to all my mentors. To Peter Pabon, for instigating and fueling my obsession with voice in all its forms. To Ji Youn Kang, for being as supportive and critical as a mentor could be. To Richard Barrett and Johan Van Kreij, for the thoughtful conversations and suggestions. I would like to extend my heartfelt thanks to all the teachers I had the pleasure of being taught by, and to Kees Tazelaar, for leading an inspiring group of people.

Thanks to Jiyoung Wi, Jacob Eckhardt, Cis De Gendt, Lucie Nezri, Amarante Nat, Andrejs Poikāns, Amos Peled, Liza Kuzyakova, Lenny Sprenger, Kerim Kali, Roc Montoriol and Gaia Heichal.

This thesis is dedicated to my dear friend Anton Van Laer.

Contents

List of Figures

List of Tables

Listings

List of Sound Excerpts

Listen to the following sound excerpts through this link:

`https://hogobogobogoo.bandcamp.com/album/vox-sound-excerpts`

1 Paul Panhuysen *Reading*

2 Michael McNabb *Dreamsong* 0m00-1m27

3 Lawrence McGuire *Keelcore* 4m11-5m26

4 Lawrence McGuire *Keelcore* 1m40-2m19

5 Lawrence McGuire *Keelcore* 3m26-3m37

6 Lawrence McGuire *Keelcore* 3m37-3m49

7 Lawrence McGuire *Keelcore* 3m47-4m11

8 Kaija Saariaho *Vers le blanc* 0m00-0m48

9 Kaija Saariaho *Vers le blanc* 9m15-10m15

10 Kaija Saariaho *Vers le blanc* 14m30-15m04

11 Gérard Grisey *Partiels* Opening

12 Tristan Murail *L'esprit des dunes* 5m44-..

13 McGuire&Nat *Enokian Soupe I* part 1 mm. 1-3

14 McGuire&Nat *Enokian Soupe I* part 2 mm. 25-30

15 McGuire&Nat *Enokian Soupe I* part 2 mm. 32-33

16 McGuire&Nat *Enokian Soupe I* part 3, section 8

17 McGuire&Nat *Enokian Soupe I* part 4 Consonant Trajectory [p]

18 McGuire&Nat *Enokian Soupe I* part 4 Vowel Trajectory [i]

19 McGuire&Nat *Enokian Soupe II* section B, m. 4a

Chapter 1

Introduction

Thought is made in the mouth

Tristan Tzara

A pivotal moment in my musical journey occurred when, within a span of two weeks, I encountered both *Dreamsong* (1978) by composer Michael McNabb and watched the blockbuster movie “Farinelli, il castrato.” *Dreamsong* acted to a certain degree as a gateway to synthesized and transformed voice, amplified (hyperreal) or attenuated (acousmatic) vocality and voice identities and its wide range of extra-musical associations. While “Farinelli” brought up questions concerning perception of voice without a physical body. In *Dreamsong* vocal textures and slow timbral changes seem to be imbued with weirdness and otherness that maintained a constant and intentional shadowed presence throughout. This perceptual quality guided this project’s direction towards dealing with artificiality and vocality at the same time. Both terms, in isolation and combination, raised a curiosity leading to inquiries into voice and non-voice (Chapter 2); the link between machines and animism; and signifiers involved in a vocal event (i.e., semiotics), especially in synthetic voice (Chapter 4). How could one describe a transition from voice to non-voice, or between two voices? If so, there must be a way to navigate between these two aural spaces by some means. Therefore, a synthesized voice will serve as a experiential *vehicle* to explore this perceptual space. In Chapter 2, I’ll refer to Bergland’s notion of a maximal and minimal voice continuum in Section 2.1, and how to operate within this by artificial means in Section 2.1.2. I’ll discuss my composition *Keelcore* (2021) to clarify technical decisions and artistic objectives relating to speech perception in what I call *vocal archetypes* or *proto-voices*.[1] Then, in Section 2.2, I distinguish three perceptual processes with regard to listening to singular or multiple voice-like sounds.

Experiencing an ambiguous sound event relating to voice, albeit in a textual, visual or musical format, or a combination between these, has an effect of enriching the material involved. It does this by allowing multiple, simultaneous interpretations in the act of reading, seeing and listening. Synthetic voice is applied in this regard as a catalyst for experiencing voice in an alternative way, where one's presumption of an unaltered voice is put into question by exposing the *simulacrum*.[2] This influences the way we listen and make sense of something vocal by also experiencing its incomplete and imperfect brother. An Aristotelian approach in gathering knowledge about voice, where we define or understand voice by what it is not, or what it is lacking. In Chapter 3, compositional techniques on achieving vocal ambiguities will be discussed through analysis of two works from the same series, namely *Enokian Soupe I* and *Enokian Soupe II*.

I would like to state that this project deals solely with synthesized voice and not processed voice, the *simulated* and not the *transformed*. This is a conscious choice, partly due to a personal interest in working with and enjoying listening to music involving physical models of the human vocal apparatus, and mainly because the voice itself is the crux of the problem. In the context of voice, the tool that generates the sound material should allow for a flexible and efficient control over gestures and timbre. Once a mathematical description for the production of a voice is found, a scope of freely navigable choices presents itself to the composer. This aspect of working within a sound's possible sonic output is exciting and eventually has led me to model the human vocal apparatus in the form of an application called NKOAPP, which is discussed in Chapter 4. The vocalization mechanism can be considered the most complicated to model, that is, when one is aiming for an *exact* simulation: a representation that allows for the production of utterance, song, spoken language, and expression of emotion, self-hood, identity and intention. Rather, the goal in building and using these models is to keep them abstract or basic in terms of parametric complexity. A broader theme for using synthesis is centered around this notion of the ubiquity of perceiving an artificial sound in relation to an unaltered, natural voice and how an experiential approach in listening to and gathering an understanding of a vocal sound can be conducted through listening to permutations of the two.

[1] Proto: the original, the first.

[2] "The incomplete simulation" or "the alike".

Chapter 2

Artificial vocality

Besides dealing with other complex sonorities, considering sounds of a seemingly vocal quality prompted reflections on aspects of identity, allusion, meaning and disposition. One can describe a vocal experience—a vocal quality of a sound event perceived by a human—as an auditory phenomenon distinct from any other worldly signal. This vocal quality can be defined as vocality, or as "a perceptual analogy with the natural voice."(Bossis 2004, 91) From this definition we can deduce that sound can exhibit vocal qualities, or at least carry some traces of a voice, while stemming from non-human sources. In this project, in the activity of musicking[1], sounds are treated and analyzed in a referential mode of listening, as opposed to the reduced mode of listening, which implies treating a sound as an object. The intention of approaching and listening to that *object* without any referential ties to its sonic origin, while the referential mode focuses on an acknowledgement of the sound's perceived origin and what relations it might evoke with the listener (Chion 1983). Based on personal experiences and neurophysiological evidence (Repp 1994), I would argue that it is inherent to human nature to actively monitor for vocal patterns both consciously and subconsciously, and that a specific attention will be diverted to those patterns— decoding the temporal, tonal and timbral aspects of what is heard into units of meaning.— In the reduced mode, we can arguably gain insights once voices are excluded from the possible set of sounds; in the act of consciously omitting the origin of a sound, we can approach the sound in a relatively subjective way, untainted by personal preconceptions and experiences where patterns can be identified through de-contextualisation of the material. When considering vocal sounds, the aforementioned reflections and internalizations they elicit relating to identity and body

[1] As in "to music". To highlight music as a process (verb) not an object (noun) and "To music is to take part, in any capacity, in a musical performance, whether by performing, by listening, by rehearsing or practicing, by providing material for performance (what is called composing), or by dancing."(Small 1998, 9)

will be left out if one uses this approach. I am specifically interested in the richness of perceptually ambiguous sounds, the ones that enforce a questioning of one's ability to distinguish the abstract from the representative, the arcane from the familiar. For that reason, the voice in its sonic manifestation, and more specifically the electronically simulated voice, presents itself as an excellent candidate for working on and between these thresholds.

A vocal perception of sound, not produced by speech organs or traditional vocal utterances, has sparked significant interest in my explorations with voice. This involves exploring the boundaries between familiar and unfamiliar vocal sound entities and their derived forms. The aim is to compose narratives that focus on the relationships between sound and its source. So, a vocal timbre, articulation, gesture, or its envelope can be intentionally produced in the listener's environment or come from a non-human origin, either originating from a natural or electro-acoustic source. The interest lies in finding *how much* of *what* features can one "leave out of" or "add in to" a signal to diverge from or converge on holistic vocal sound image respectively. For example, in the piece *Reading* by Dutch composer Paul Panhuysen, voice recordings are routed via galvanometers through steel wire to the strings of a double bass, resulting in an uncanny humming and buzzing (excerpt 1)(René, 1993).[2] An incomplete translation of the speaking voices occurs, deficient in human sonic attributes, but containing just enough to cast a vocal shadow on the listener, as if something is being uttered somewhere to someone.

Tying artificiality into the context of vocality, we get what Bruno Bossis describes as *artificial vocality*; when framed within electroacoustic music Bossis mentions:

> By its very nature, artificial vocality establishes a new link between the vocal quality of a sound event (its vocality) and technology (its artificiality) within this type of music. (Bossis 2006, 91)

Until now the altered, and possibly artificial voice, has been discussed as a tool to speculate on and investigate the real and not the other way around. Can the speech organ also affect an artificial response? In other words, can a sound produced by a human come close to the artificial, where the performing voice intentionally intersects with the synthetic voice? How does one distinguish the key acoustic and perceptual elements involved in this subversion? This bidirectional notion drives my

[2] Taken from the album *Lost For Words*. Recordings of reading voices by members of the Macunia Ensemble, of which he was also a part. From the liner notes by René van Peer, "For this album the installation comprises five walkman cassette players with nine galvanometers plugged in. The steel strings, all 2 meters long, have been attached to the middle of duochords that each have a different length. Their lengths (and therefore their pitch relationships) have been determined through calculations based on the number 3, resulting in two pentatonic rows: one starting from 81cm upwards, the other from 81cm downwards."

motivation to work and compose for natural and synthesized voice: especially in duet formats where fusion and fission of timbres, dramatis personae, and linguistic components of both human and artificial performers highlight a distinct two-way interplay Figure 2.1.

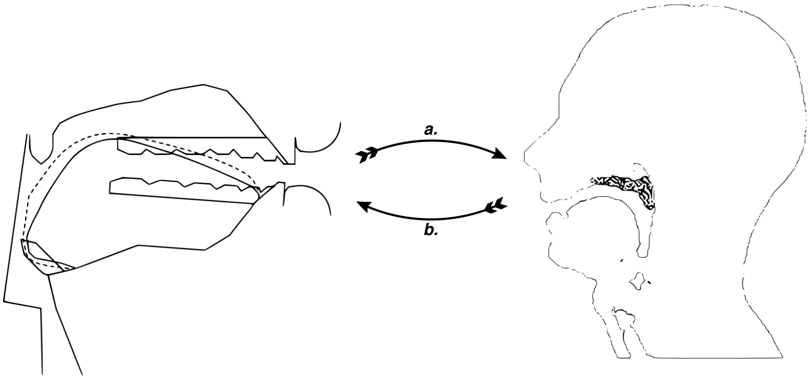


Figure 2.1: How can *a* seem artificial to *b*, and vice versa

2.1 Maximal and minimal voice

McNabb mentions in his article "*Dreamsong*: The Composition" that "the basic intent of the piece was to integrate a set of synthesized sounds with a set of digitally controlled recorded natural sounds to such a degree that they would form a continuum of available sound material."(McNabb 1981, 2) For example, the piece opens with a sound texture similar to chatter in a refectory during lunch hour, which then quickly morphs into a natural, sustained singing voice timbre on a D♭. A critical zone seemingly appears where one's sense of perceptual distinction between voice and non-voice is not discrete, but rather ambiguous and of an unstable nature (excerpt 2). The focus lies not on pitch or other traditional musical parameters, but on perceiving the transgression of a threshold from one sound identity to another. McNabb's aim was to build a traversable and observable continuum between "sonic realities" where realities can become ambiguous, and induce oscillations between multiple perceptual identities. These oscillations might seem to emerge and cease at random, but actually occur due to a set of shared, missing or overpowering acoustic and/or psycho-acoustic features that are characteristic to a particular sound identity. Once one gains familiarity with vocal features that establish the principal building blocks of specific vocal Gestalts (Bregman 1990) and identifies the instances where the sound image splits (e.g., knowing the thresholds), a sonic dance between them can be choreographed. A strong example can be found in the operatic finale of the last scene of *Farinelli, il castrato*, where the castrato singer and Baroque pop-star Carlo Broschi performs a virtuosic rendition of Handel's "Lascia ch'io pianga"[3] and momentarily reaches a very high note with an alien-like, mechanical vibrato[4]: the

[3] `https://youtu.be/645ayYMgTcE?si=VEqkNlWxn_kzQbgL&t=182`

[4] Researchers at IRCAM were given the task to devise a synthesis scheme to allow for a continuous change in registers between a coloratura-soprano and a counter-tenor, to cover the entire range and compensate for technical difficulties. Essentially to "bring back to life a repertoire which could not be sung anymore." (Depalle, Garcia, and Rodet 1994, 1)

digitally synthesized voice evokes a completely different reality, seemingly without the physical operatic force, even though viewers can clearly see Broschi sing. It is as if the voice fluttered out from his chest to his head and away from his body. In the coming writings I'll clarify how the outline of these movements can often be traced back to moments of perceptual fusion or fission, rivalry and masking, by referring to psycholinguistic studies and experiments in my own music.

I will often refer to Bergland's framework for treating the zone between voice and non-voice as *the continuum between maximal* and *minimal voice*[5], between two extremes "as reference points against which the experience of different types of transformed or manipulated voices might be judged and compared."(Bergsland 2010, 3) Finding techniques on how one can enter, operate within, and leave this zone is of particular interest to me. In my compositions, the synthesized voice takes the role of a probing tool and agent to respectively monitor vocal signifiers and operate between these maximal and minimal voices. In my thesis, Bergland's framework becomes useful specifically when encountering the difficulty of analysis of both the communicating speaking voice and the critical boundary between voice and non-voice (Bergsland, 2010).

2.1.1 *Keelcore*

From early on in my musical endeavours, listening to music that involved synthetic voice evoked a variety of imaginary realities. From Paul Lansky's phonemic and cartoonesque *Idle Chatter*(1994), to vocaloid covers of Baroque laments, to Alvin Lucier's experimentations with vocoded synthetic glossolia in *North American Time Capsule* (1967), to Herbert Eimer's haunting *Epitaph für Aikichi Kuboyama* (1962), to Neil Young's disco-voder alter-ego singer in *Transformer Man* (1982), to SOISONG's *T-Hu Ri Toh* (2009), to Ron Kuivil's mutant speech forms in *Linear Predictive Zoo* (1987), to Larry Wendt's love song *Galaxy Love* (1988), and so on.[6] What these musical references have in common: some form of digital mediation of the voice. In this context, the synthetic voice acts as a broken mirror of the familiar voice. Reflecting distorted, yet perceptually potent vocal images onto the minds of the listener. It is up to the listener, then, to decide whether to reconstruct these images into a cohesive whole or

[5] The maximal and minimal voice is a concept Bergland borrows from literary theorists Donald Wesling and Tadeusz Slawek, in which the maximal and minimal voice serve as conceptual tools for the analysis of literary texts. The maximal voice is described as a typical informative and neutral speaking voice (e.g., radio voice), while the minimal voice is usually highly manipulated and often quite abstract, thus defineing the zone between what is voice and what is not voice (Bergsland 2010, 3).

[6] *Idle Chatter* - `https://www.youtube.com/watch?v=Lfq2OZ8lpA8`; two synthetic cover of *Dido's Lament* (1689) by Henry Purcell - `https://sintel.bandcamp.com/track/mikus-lament` and `https://nickhoffman.bandcamp.com/track/dido-s-lament-purcell`; *North American Time Capsule* - `https://www.youtube.com/watch?v=ot_EEK9VRd4&t=155s`; *Epitaph für Aikichi Kuboyama* - `https://www.youtube.com/watch?v=_-m-YKtH43g`; *Transformer Man* - `https://www.youtube.com/watch?v=iHdS1zdhWQs`; *T-Hu Ri Toh* - `https://www.youtube.com/watch?v=FKRR_mMVHzY`; *Linear Predictive Zoo* - `https://www.youtube.com/watch?v=DZ5pUUXqkMc`; *Galaxy Love* - `https://www.youtube.com/watch?v=HVhaWkt2040`

to accept them as variations on the known or familiar. In 2021, *Keelcore* was composed to experiment with synthetic voice with the aim to imbue my music with a vocal quality that balanced on the limit of purely synthetic yet containing traces of natural voice. *Keelcore* challenges the canonical view of what a vocal experience is or could sound like.

I discuss the technical decisions involved in developing an articulatory speech model responsible for most of the sound materials in *Keelcore*, while relating the decisions on their impact on emphasizing the artificial vocality of a sound. This articulatory speech model simulates the human vocal tract as a network of digital filter simulations of acoustic tubes or *tubelets* where vocal articulation is controlled by changing the tubes' radii.

2.1.1.1 Vocal Archetypes

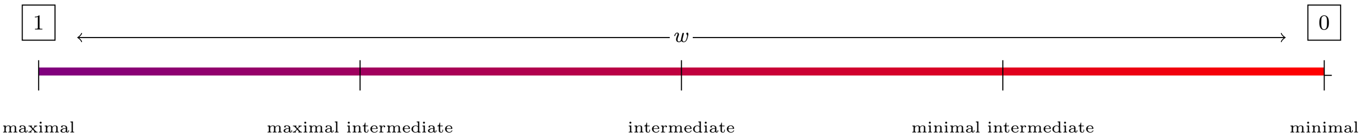
In *Keelcore*, the synthetic voice wanders between moments of clearly recognizable vowel formant structures and unfamiliar timbres. Oscillations transpire between articulator movements harmonious with, for example, sustained singing voices, and mechanical mouth and throat movements. Of these sounds, vocal facets such as articulation, intonation and voice quality can be linked to a vocal archetype or proto-voice. For example, a vocal archetype can be a crying newborn baby, someone yodeling, or a person speaking in a dialect. The proto-voice often portrays a quality of human origin or essence, emphasizing one or more distinctive features. For example, the most prominent feature of the sound of a crying baby is that of a loud vocalization of [æ] or [wɛ]. In the case of the sound of someone yodelling, a rapid transition between a higher and lower register in pitch is of more signifying potency than the actual pitch interval. Connecting this idea to the maximal and minimal voice framework, a couple of things can be adjusted. The maximal voice, exemplified by Bergsland as the "radio voice", formulates this as a set of premises (Bergsland 2010, 142-144) :

1. **Linguistic-semantic focus of attention**: The semantic level within the linguistic domain receives sustained and maximal attention.
2. **Balanced information density**: The information density (i.e., the amount of information that a listener can infer during a certain period of time) of the content should neither be too high nor too low for processing the verbal features of the voice.
3. **Naturalness**: The sound has maximal resemblance with one produced by a human being and his/her/their vocal apparatus.
4. **Presence**: The listener experiences a sense of a shared "here and now" with a vocal persona.
5. **Clarity in forming meaning**: Meaning can be constructed from the voice with a high degree

of clarity – also implying specificity, certainty and coherence.

6. **Feature salience**: Vocal sounds and features “stand out” perceptually – for themselves and relative to other sounds and features.
7. **Stream integration**: The sound of the voice is integrated into one coherent and continuous sound stream (cf. auditory scene analysis).

For a sound to qualify as somewhat experientially similar to its vocal archetype, it must possess at least one prominent premise. In comparison to others, this component becomes heavily weighted, expressed by the value w. Each of the seven premises of the vocal sound is analyzed on a weighted scale where $0 \leq w \leq 1$.



A minimal voice is everything that an archetype or proto-voice is not, making it easier to define the maximal voice than the minimal voice. Take yodelling as an example of a vocal archetype. The premises are weighted and labeled categorically,

1. **Linguistic-semantic focus of attention**:*not relevant $w = 0.0$ - *minimal*
2. **Balanced information density**: *not relevant $w = 0.0$ - *minimal*
3. **Naturalness**: *same premise $w = 1.0$ - *maximal*
4. **Presence**: *same premise $w = 0.0$ - *minimal*
5. **Clarity in meaning formation**: in yodeling you have to sing the word /jo/, but if this isn’t that clear it is not that detrimental $w = 0.5$ - *intermediate*
6. **Feature salience**: *same premise $w = 1.0$ - *maximal*
7. **Stream integration**: *same premise $w = 1.0$ - *maximal*

In *Keelcore*, the synthetic voice points towards the archetype of a male voice timbre. The synthesized material largely resides deep in the realm of the minimal voice, occasionally approaching the maximal voice to a faint extent. For the entirety of the piece, the speech potential of the voice remains limited: no clear speech can be made out. However, sounds resembling vowels and occasional glottal clicks can be heard throughout the duration, mimicking an untrained or impaired speaker with no speech motor control, producing uncontrollable babbling utterances. In the case of a healthy male voice timbre as the vocal archetype, two primary premises necessary to experience this archetype are naturalness and presence. The feature salience takes on the role as secondary premise and is weighted lower. Concerning the rates of articulation, it is clearly noticeable that there is a destructive effect to the perceptual construction of the natural and healthy male voice. In the computational model, the physical inertia of masses can be ignored. Tubelet cross-sectional areas can be modulated at rates from 0 Hz upwards to audio rate. This is in violation with the possible spatiotemporal resolution of sustained sounds (i.e., vowels) of a male speaker, which is between 200 and 250 milliseconds, as seen in Figure 2.2. Furthermore, the synthetic agent's vocal identity frequently becomes ambiguous. For instance, in excerpt 4, the length of the larynx is prolonged up to 4 times its original length.[7] Sonically resembling a beefy and spectrally dense drone coming from an analog synthesizer. The synthetic voice also does not breathe or take periodic brakes during its vocalizing periods; in excerpt 5, a sustained breathy-voiced vowel can be heard for another ten seconds and brings forth this image of unnaturally large lungs with a virtually infinite lung capacity. There are also moments of slight convergence to a signifying feature of the male voice (excerpt 6). For instance, in certain parts, a hoarse and grainy voice is perceived and is somewhat similar to the uvular fricative [χ]. Then, in excerpt 7, these grains become louder and attain a gritty, digital ringing form; discarding its previous identity almost completely and making a leap back into the realm of minimal voice. In conclusion, it is essential for the primary premises to be present in order to achieve an auditory experience close to a vocal archetype.

Depending on the design choices made in the synthesis system, one can amplify or attenuate the presence of a certain premise. The program used to generate sound materials for *Keelcore* is designed to operate between super-human movement speeds and normal movement speeds of the articulators. The system is capable of operating between the two poles of the naturalness continuum of the vocal archetype of a male timbre.

2.1.1.2 Synthesis and control system

In *Keelcore*, the aim was to use one type of physical model from which all material could be generated. The system consists of a synthesis and control part built on top of an existing MaxMSP patch by Peter Pabon, a patch using SuperCollider's NTube Ugen[8], and an accompanying custom eight-channel spatialization system. Due to the relatively small parameter size of the models when compared to a fully fledged articulatory synthesis model as in VocalTractLab (Birkholz 2013) or Praat (Boersma and Weenink 2001), achieving speech-like outputs such as intelligible phonemes or larger linguistic

[7] This is done by running the DSP at 4 times the initial sampling rate: $48.000Hz * 4 = 196.000Hz$. Essentially making the larynx go from 16.9 cm to 67.6cm in length.

[8] A physical model of an amount of cylindrical tube segments that describes the vocal tract to some extent. The length and diameters of these segments can be modified on the fly.

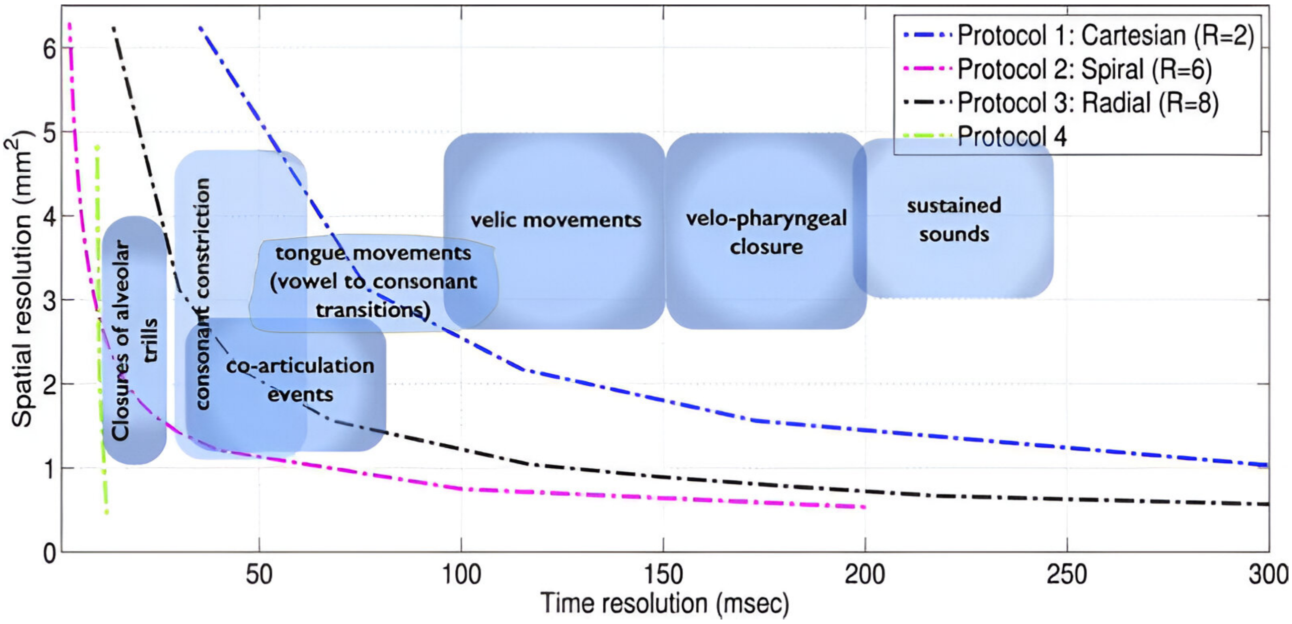


Figure 2.2: Spatiotemporal resolution of speech tasks. The highest rate at which a vocal sound can be made is around 90 Hz.

structures (e.g., morphemes or words) was never its intended use, nor did it influence the control system's architectural design choices. Its primary use was to generate materials on the cusp of humanly possible articulations and timbres.

The system employs a source-filter approach where the source is a physical model of the vocal folds (Tokuda and Herzel 2009) mixed with white noise. This signal serves as the excitation of the filter. The filter is a computational physical model of the human vocal tract, known as a "Digital Waveguide".[9] The larynx, mouth, and lip openings are discretized into a set amount, N, of cylindrical sections or *tubelets*. Each tubelet has a cross-sectional area A_i, which, together with its precedent and antecedent areas A_{i-1} and A_{i+1}, describe the reflection coefficients k_j and k_{j+1}, with $i \in \{0, N-1\}$ and $j \in \{0, N-2\}$. The coefficient k describes the extent to which the acoustic pressure wave is reflected downstream, towards the glottis and subglottal area (e.g., lungs), and how much of it is transmitted upstream towards the mouth. Chained together, these waveguide sections make up a Kelly-Lochbaum model (Figure 2.3). Its full implementation is shown in appendix B.

The tubelet areas can thus be changed over time, allowing for different shapes of the vocal tract; resulting in various filtering of the excitation signal. Depending on the level of constriction, whether full ($A \approx 0$) or partial ($A > 0$), different phonations are produced. These include pure vowels, as well as

[9] Digital waveguide synthesis models are computational physical models for certain classes of musical instruments (string, winds, brasses, among other) which are made up of delay lines, digital filters, and often nonlinear elements (Smith 2006, 1).

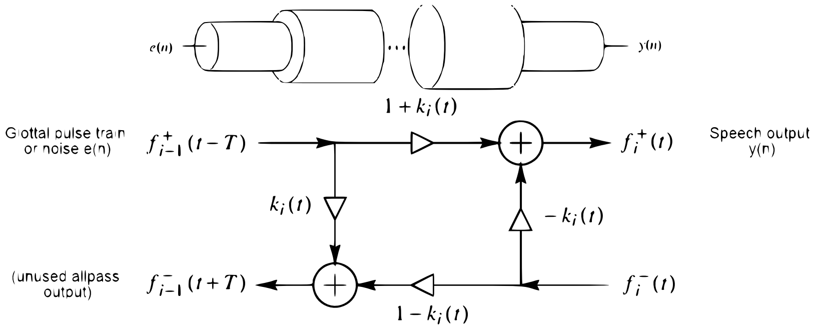


Figure 2.3: Kelly-Lochbaum piecewise-cylindrical acoustic-tube model for the human vocal tract

vowels with friction, and guttural clicks. This brings us to the control interfaces as seen in Figure 2.4, which implement a method for accurately synthesizing vowels through vocal tract area functions (Story 2001):

> Story and Titze (1998) showed that a speaker-specific set of vocal tract area functions (acquired using MRI, Story et al., 1996) corresponding to ten vowels can be decomposed with a Principal Components Analysis (PCA) into a set of **two orthogonal components and a mean diameter function**. These components, which define the spatial shaping patterns of the area function, were referred to as "empirical orthogonal modes" or simply "**modes**." A reasonably accurate reconstruction of the original vowels can be realized by summing the mean diameter function with a weighted combination of the modes (along with a final step of converting the diameter functions to an area function). (Story 2001, 1)

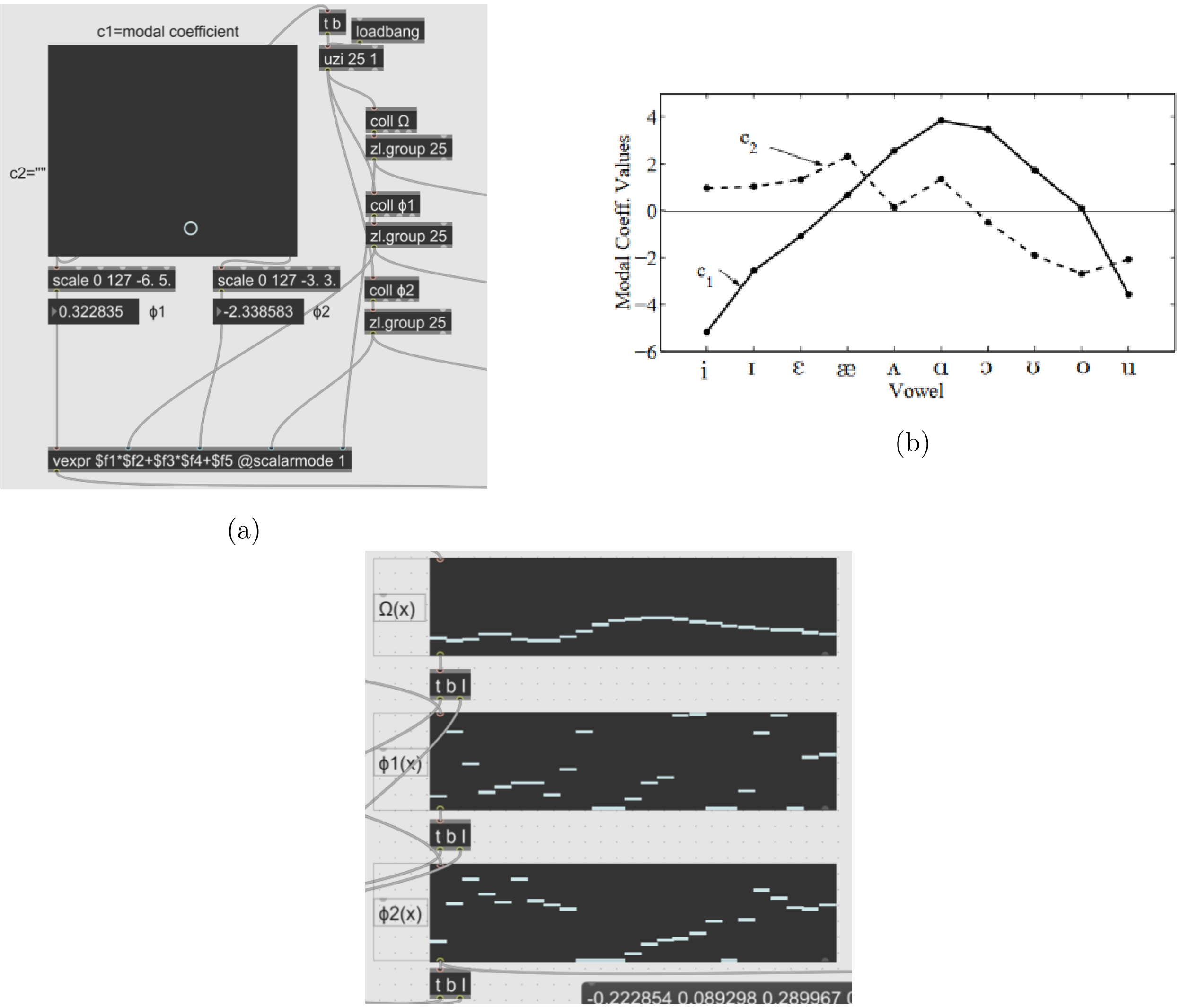
This area function can then be expressed by,

$$V(x) = \frac{\pi}{4}[\Omega(x) + c_1\phi_1(x) + c_2\phi_2(x)]^2 \tag{2.1}$$

where $\Omega(x)$ is the mean diameter function, $\phi_1(x)$ and $\phi_2(x)$ are the two modes, and c_1 and c_2 are weighting coefficients that reconstruct an area function for a given vowel (Ibid.). This method allows for a smooth control of the vocal tract shape and also prevents the filter from blowing, which results in extreme loud clicks. In *Keelcore*, recordings were made during modulation of these tubelet diameters, where two approaches with regard to the modal coefficients c_1 and c_2 are distinguished as,

a. A manual control of the modal coefficients (e.g., with mouse movements; position randomization of the XY-pad)

b. An automated control of the coefficients (e.g., inputting oscillators for moving across X & Y axes)



(a)

(b)

(c)

Figure 2.4: Control interfaces for F1 and F2 changes; (a) XY-control space for c_1 and c_2; (b) Modal coefficients for reconstructing the original ten vowels with Equation (2.1); (c) Control space for nominal diameter function $\Omega(x)$ and modal shapes $\phi_1(x)$ and $\phi_2(x)$. *Keelcore*

The interface used in approach *a*, shown in Figure 2.4(a), presents an XY-interface that can independently and continuously control the first two formants, F1 and F2. In Figure 2.4(b), ten value pairs are given for c_1 and c_2 to reconstruct ten vowels, which respectively determine the ranges for the X and Y axes.

Approach *b* accesses a wide range of modulation rates, enabling a seamless transition between control and audio rate articulatory movements (i.e., tongue, lips, larynx shape). Furthermore, plugging two periodic signals into the X and Y inputs results in more interesting behaviours compared to manually controlling the input with a mouse. Depending on the degree of correlation between the signals in their phase and frequency relationships, Lissajous patterns could emerge where the formant space is navigated in a periodic fashion.[10]

Another interface, shown in Figure 2.4(c), allows for modifying $\Omega(x)$, $\phi_1(x)$, and $\phi_2(x)$. By manually drawing the values for the tract areas through a mouse input, sudden constrictions and expansions occur. When large, discrete jumps take place between two cross-sectional areas of the tubelets, the signal value is significantly amplified and produces loud clicks at that waveguide section in the delay line. This means that these clicks still need to pass through the rest of the delay line, which filters this signal in the same way the mouth filters a sound. These comb- and formant-filtered clicks and impulses are seen as artefacts of the machine and blur the boundaries even further between natural and synthetic vocal sounds.

The vocal archetype of a person slurring their speech was used as an inspiration to compose the spatialization of the synthetic voice. When speech is slurred, there is typically a deviation from the typical articulation patterns, resulting in a less precise and more relaxed or distorted vocal output. This element aligns with the premise of naturalness. Each spatial trajectory adheres to a movement around an octophonic loudspeaker ring. When one sound has to travel from the front to the rear center speaker, there are two choices to arrive at that destination; clockwise or anti-clockwise. This is done to instill a sense of continuity and speech-like behaviour in the dynamic unfolding of the synthetic voice, and not let it suddenly “teleport” around the room. Furthermore, depending on the velocity and distance at which the sound travels from one loudspeaker to the other, the delay time of a cubic delay line is modulated. For the same reason as above, a lagging speech-like quality is added to the otherwise constant vocalizations of the synthetic model (excerpt 3).

[10] Depending on the amplitude difference of the two periodic signals, the curve is *stretched* along the X and/or Y axes.

2.1.2 An Operable Continuum

To thoroughly explore the continuum between maximal and minimal voice across all seven premises, I devised a synthesis scheme that allowed for flexible control over vocal timbres and articulation. This approach enabled me to efficiently navigate through this continuum.[11] Additional requirements for the synthesis model included expressing a wide sounding scale from abstraction to realistic image; describing the border between an abstract and concrete form of sound perception; and eventually prompting a communication between the virtual and the real (Chafe 2004, 4). Composer and free-improving vocal performer Trevor Wishart mentions in chapter 16 of his book *On Sonic Art* (Wishart and Emmerson 1996, 325-329), that a *sound-model*, or in this case, a vocal archetype, should maintain its perceptual form when articulated upon. Wishart points out,

> Thus, for example, the spectrum (and its evolution) of a bass piano note is noticeably different from that of a high piano note - as sound-objects they are remarkably different yet we relate both of them to the sound-model "piano" quite directly. (Ibid., 326)

Furthermore, Wishart mentions that a sound-model consists of a set of (near-) invariants. If one changes the invariants of a sound model, "the sound model's intrinsic morphology changes and also the set of rules that govern the behaviour of that model change when articulated by some input device." Relating this concept to the position of the computer and the possibilities it brings with:

> . . . We are not confined to basing our sound-models on existing physical objects or systems. We may build a model of a technologically (or even physically) impossible object. We might specify the characteristics of the voice of an imaginary creature. Once, however, the sound-model is specified, we are free to change the invariants of its behaviour. We may transform it into an entirely different sound-model. (Ibid., 327)

I see the vocal archetype or proto-voice as a sound-model which one can choose to diverge from or converge on by changing its invariants. In the physical modelling paradigm, these invariants combine to construct a parameter space in which the model can be articulated upon while sustaining its image. To navigate this parameter space, whether with the aim to break or maintain a mental image, requires specialized knowledge of the spatio-temporal, timbral, idiolectic, and linguistic cues involved in the vocalization process of the vocal archetype. For example, let us take the female vocal apparatus as the sound-model. There are physiological constants that distinguish biological genders such as length

[11] In Chapter 4, the synthesis and control system of the NKOAPP program is discussed.

of the vocal folds and tract, size of the voice box and pharynx. These amount to different timbres, vocal pitches, and resonance properties (Fitch and Giedd, 1999). When these constants are altered to the extent that the perceptual form disintegrates, the sound model changes, retreating further back towards the minimal voice end of the continuum. A more specific *sub*-sound model of a female voice could be that of an adult using “motherese” or “baby talk” while speaking to her newborn. This is an archetype of speech that is characterised by “a high tone of voice and exaggerated contours, simplified language, and a high incidence of interrogatives, imperatives and repetitions.”(Taverna 2021, 304) Then, switching to speaking in a dead-panned voice will sound and possibly be labeled as “frightening” or “weird,” breaking the impression of the soft, caring personality of a mother. These sub-archetypes can be seen as *nested* within a more general sound model or super-model. As one branches further down the vocal archetype tree, the sound models become less stable, but more defined, and the quasi-invariants become more *invariant.*

Taking the “radio voice” (Bergsland 2010, 139) as the primary or root archetype, we can then branch out into two arbitrary subsidiary archetypes or nodes: male and female voices. These higher order archetypes still contain the essential components necessary to depict the “radio voice,” but ask for a more defined description. From the male voice, one can delineate, for instance, the archetype of a male teenager’s voice during his coming-of-age period. This vocal archetype is usually characterized by its instability and inconsistency in pitch, tone, and resonance, often marked by voice cracks in their speech.[12]

In *Keelcore* the sound-model is the human vocal apparatus of a middle-aged male. It is abstracted into discrete parts, forming a simplified representation of the vocal tract and its excitation. The male voice behaves according to the same rules that govern invariants belonging to female voice production. Here, three invariants are deliberately altered to traverse the continuum through various paths. In Table 2.1 their respective ranges or types are given.

Table 2.1: Invariants and respective ranges or types for maintaining a stable male vocal archetype

Invariants$_{male\, voice}$	**Ranges or Types**
vocal register & range	$90 - 155Hz$ & $80 - 330Hz$
spatio-temporal resolution	$1 - 5mm^2$ & $10 - 250ms$
excitation signal	glottal pulse & turbulent noise

Depending on the phonatory mode (i.e modal, breathy, pressed, among other), the excitation

[12] Under my pseudonym *hogobogobogo*, the EP *pepsi/coke: An Electronic Sound Poetry Transmission* sprung to life based on the idea of hearing your own voice change throughout puberty. In the tracks, adolescent voice cracks and unstable tones are used as signifiers of this psychological and physiological transformative period. `https://hogobogobogoo.bandcamp.com/album/pepsi-coke-an-electronic-sound-poetry-transmission`

resembles resemble an impulse train coloured to some extent by turbulent noise. Where pitch and intensity movements regularly exhibit familiar prosody and intonation patterns. The oscillator's pitch exceeds the minimal and maximal boundaries often, and articulations go far into audio rate territory.

2.2 Between voices

2.2.1 The singular and poly-voice

Up until now only a singular synthetic voice has been discussed, isolated from other voice-like sounds. In this scenario, tensions arise due to the presence of some shared prototypical features with the vocal archetype. Information in the signal can appear inconsistent or in conflict with the mental representation it is trying to propel. What does it mean for the natural voice trying to resemble or approximate the artificial? Can the synthetic fuse with the natural voice? How can this be achieved? On the other hand, the natural voice can also converge to the synthetic; implying that the synthetic takes on the role of the vocal archetype. Informed by experiential, psycho-acoustic, and linguistic factors, this becomes an investigation into the vocal agent's presence and perceptual stability in relation to the other heard voice. A sound form is considered stable or whole when a listener can distinguish it from its sonic surroundings for an observable time frame, and discard the components not relevant to the object. Essentially, this is a process of grouping, where separate process happens on a neurological level unique to speech-like sounds.[13] This mechanism is extremely good at decoding speech-like sounds into its lexical, semantic and musical information. An utterance can be stripped from most its spectral information (e.g., F1 and F2, or spectral envelope) yet still be perceived as vocal from the presence of the stressed and unstressed syllables, pauses, and loudness and pitch contours.

Consider two synthetic sounds with vocal traces, positioned somewhere along their minimal-archetypal voice continua. Both can be said to manifest an abstract, perceptual *distance of likeness* to their archetypes. Additionally, one can arguably also perceive distances between the synthetic entities themselves, and between the archetypes that they're mentally imposing on the listener; each acting as a node in a network that is perceptually interconnected to varying extents. These connections elicit tensions depending on the strength of likeness between synthetic sounds and their *proto-voices* (Figure 2.5).

What if we expose the archetypal voice simultaneously with the synthetic? Depending on the clarity of the archetypal voice and how close the synthetic approximates it, perceptual phenomena

[13] There is evidence that show distinct brainstem responses to speech sounds compared to non-speech sounds, indicating separate processing (Russo et al. 2004, 1).

described as fusion, fission or rivalry can take place. These phenomena are discussed in the following paragraph.

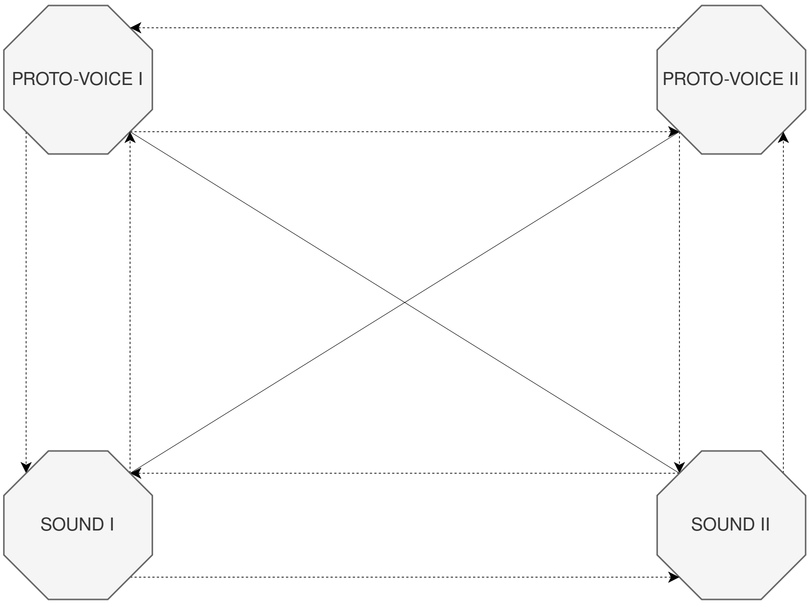


Figure 2.5: Tensions that two simultaneously exposed vocal sounds can undergo

2.2.2 Vocal Phenomena

During the audition of simultaneous vocal sounds, one can witness a variety of interactions related to their origins and perceptual stabilities. These auditory phenomena happen due to the presence or lack of shared components encoded in both signals. Consider the situation of a synthetic voice and natural voice both sounding and being heard. First of all, each sound holds a certain perceptual distance to its vocal archetype. This distance can be expressed by the premises discussed before, where smaller distances contribute to the rigidity of the perceptual sonic form. Second, measuring similarities in acoustic, psycho-acoustic and psycho-linguistic features between both sounds is useful in order to describe, document, categorize and explain these phenomena. In the process of auditory perception of vocal sounds, I distinguish three vocal phenomena based on research done by psychologist Albert Bregman in “Auditory Scene Analysis” (Bregman):

1. Masking: Masking occurs when the perception of one sound is affected by the presence of another sound. The masking sound can either be simultaneous (occurring at the same time) or forward (preceding) in time. Masking can make it difficult to hear quieter sounds in the presence of louder ones, or to distinguish between sounds with overlapping frequency ranges.
2. Fusion: Auditory fusion refers to the perceptual integration of multiple sound sources into a single auditory object or stream. It occurs when the brain combines auditory signals that are

temporally or spectrally similar, treating them as a single perceptual experience.

3. Rivalry: The brain alternates its focus between two or more competing auditory stimuli. When presented with conflicting auditory inputs, the brain switches attention between them, resulting in fluctuations in perception.

The masking of one sound with another implies a loss of information, caused by excessive information density in one of the sounds. Rivalry exhibits an oscillatory behaviour between the vocal archetypes. Fusion, or perceptual integration, is the experience of the two vocal sounds into a single auditory percept. It is important to note that rivalry and fusion are not exclusive phenomena, but interact in a probabilistic way (Cutting 1976, 123). And, if fusion is possible, so must perceptual fission be; implying the disintegration of the previously fused percept. A fluctuation can take place between a singular perception and perception of several vocal identities.

2.3 Conclusion

In this chapter, the importance of listening to a vocal sound in a relational mode has been discussed, where information on identity, allusion, meaning and disposition are not left aside. The chapter shows methods to describe the voice through the use of synthesized voice, and conversely: to describe the synthetic through the natural. Both are situated on the perceptual spectrum between minimal and maximal voice, where the maximal voice is interchangeable with a vocal archetype or proto-voice. The maximal voice is formulated as a set of premises. If these premises are not met, we venture into the ambiguous realm of minimal voice. Operating between the ends of this continuum brought up the notion of Trevor Wisharts's *sound-models* and its set of invariants, which, in turn imply boundaries and thresholds that the musician is free to explore. Auditory phenomena such as fusion, masking, and rivalry between two vocal sounds are introduced and described as the result of a combination of two factors: the perceptual stability of each sound's vocal archetypal and the shared or lacking features between the two vocal sounds.

Chapter 3

Achieving Vocal Ambiguity

In this chapter, I expand on ideas of speech and verbal sounds in my own work and introduce the terms “vocal” and “timbral ambiguity,” which could arise from the simultaneous combination of a synthetic and natural vocal sound. A well known approach in Spectralist music is the idea of *spectral fusion*, which is just one of several approaches in achieving a fused sound percept. This approach, along with others, is discussed and analyzed by referring to works closely related to the ethos of the Spectralist movement, as well as referring to my own compositions. For instance, in Tristan Murail’s *L’Ésprit des dunes* (1994), spectra from the sounds of Mongolian throat singers and Tibetan trumpets are projected onto the orchestra and live electronics. A fusion of spectra between the two actors ultimately leads to a hybrid sound. This combined sound is often perceptually unstable and correlates with Deleuze’s notion of a “phantasm”; a mental image that lacks stability and permanence. The phantasm operates as a representation that serves as a bridge between different senses and allows for the creation of a variety of meanings in the mind of the listener (Deleuze 1990). Diana Deutsch categorizes this as an “auditory illusion” (Deutsch 1974), which I interpret to be perceptually located somewhere on the continuum between the vocal archetypes and their minimal voices.

In Kaija Saariaho’s piece *Vers le blanc* (1982), an extremely slow interpolation between four spectra manifests a particular attention to the formant structures of voice, where the central frequencies of the formant regions of the synthetic voices change over time and sometimes combine into a vowel-like timbre.[1] As is also the case in McNabb’s *Dreamsong* (1978), the focus lies not on the parameter of pitch, but on moments of clear or imperceptible transitions between vocal and non-vocal timbres. I

[1] There are no orchestral recordings available, but some of the synthesized parts can be listened to in sound excerpts 8, 9, and 10. Taken from (Morrison 2021).

describe these moments of fusions or interpolations as occurences of “vocal ambiguity”.

Parallel with these artistic and analytic considerations of other composers, I reintroduce and reinterpret approaches found within them, and contribute new ones regarding timbral and linguistic ambiguity through the lens of the physical modelling synthesis paradigm. This differs from the spectral or acoustic model where one starts from a known harmonic or inharmonic context and extrapolates from there. While in physical modeling, one starts with an abstracted, but physically correct representation of the production mechanism of a sound source and where its harmonic content and sound identity emerge from the excitation of the model.

Two compositional techniques that are discussed are hybridization and interpolation. The first is a superposition or morphing of sounds alike in timbre or linguistic components and is elaborated upon by referring to my compositions *Enokian Soupe I* and *Enokian Soupe II.* In *Keelcore*, interpolation is discussed and covers sudden or seamless transitions from one vocal archetype to another.

3.1 The Perceptual Organization of Complex Sound Objects

The following text is a paraphrased and translated version of the article “L’organisation perceptive de l’environnement sonore” by music cognition expert Stephen McAdams. McAdams expands on the relevant aspects concerning auditory perception that I adress in my approach to the perception of voice.

> When a mixture of sounds arrives at the tympanic membrane, the acoustic signal is decomposed and felt by frequency selective auditory fibers. Each fiber can carry information from **multiple sound sources**. Our central auditory system analyzes this distributed activity to recover the original constituents, that is, **to form appropriate mental representations of sound sources in the environment**. The central auditory processing system is categorical by nature and aims to carry out a **perceptual grouping in order to link the components of the sensory representation coming from the same sound source and separating those coming from distinct sources**. These grouping processes apply to components that overlap temporally and to those that follow one another over time (Bregman, 1990). To carry out such processing, there must be a phase of peripheral analysis of the sensory data followed by a series of simultaneous and sequential grouping processes which attempt to assign ’descriptors’ or ’primitives’ of the neurophysiological analysis to the descriptions of the sources on the basis of coherent behavior of subgroups. (McAdams

> 1997)

And also mentions,

> The creation of virtual sources by electroacoustic means nevertheless demonstrates that the analysis into distinct sound objects is not necessarily impossible when faced with a single physical source. For example, if one would listen to a monophonic recording of a symphonic orchestra, one hears and understands the presence of multiple objects. (McAdams 1997)

The Spectralists would create situations where a single "musical object" or "sound identity" is heard by combining several physical sound sources[2], or where two sound identities are used to form a timbrally ambiguous result by cross-synthesis of their respective sound spectra. In these ambiguous situations are where, I, as a composer, see interesting ways to enrich both the hearing and listening of sound identity, more specifically, a vocal identity. This identity refers back to the grouping element in perceiving our auditory habitat. Sound identities can be demarcated both on a perceptual level through a grouping of shared acoustic features, and on a sensory level through past experiences. This motivation for using voice is further supported by evidence found in experimental psychology, where vocal and non-vocal sounds are shown to be processed on different perceptual levels. For example, in dichotic listening experiments[3], the phenomenon of duplex perception serves as possible evidence in categorizing speech as a distinct signal (Cutting 1976). A phoneme is divided both spectrally and temporally into two portions, which are then presented separately to each ear. One ear receives a synthetic 'chirp' and the other the rest of the spectral and temporal information (Figure 3.1).[4] The resulting percept is split, as cognitive scientists James E. Cutting mentions:

> The listener hears more than one auditory event. He does not hear two speech sounds. Instead, he hears one speech sound, [dɒ], and a non-speech sound, a chirp (Cutting 1976, 6).

This brings us to a necessary observation, which is that voice can be seen as a complex sound event which behaves differently on a perceptual level and offers the composer a toolbox that allows for

[2] For instance in Grisey's *Partiels* or in Tristan Murail's *L'Ésprit des dunes.*

[3] Two different or identical auditory stimuli are presented to each ear independently so that the stimulus in one ear can't be perceived in the other. This is not to be confused with *binaural* listening, where sounds from multiple sources are present in each ear.

[4] Which does not siginificantly resemble the phoneme [dɒ]. Implying that a perceptual construction of phonemes can happen from speech and non-speech stimuli. Notably, these perceptual constructions occured only under specific favourable circumstances regarding relative values between stimuli, such as fundamental frequency, intensity and onset timing asynchrony.

creative approaches not possible in non-voice settings. Simultaneous combinatory processes of voice can, for instance, result in incidental or accidental linguistic effects, such as hearing [dɒ], pronounced as an English *dot* when hearing both [ga] and [ba] simultaneously (Cutting 1976). This is linguistically incidental because of the fusion being a psycho-acoustic effect. The second formant transitions of [ba] and [ga] lie between [dɒ]'s transition, so a perceptual averaging occurs to some acoustic sounds with regard to voice. The combination of two phonemes yields a new phoneme. On the other hand, given the dichotic pair "banket" and "lanket", which are nonexistent words, one often hears the fused word "blanket". This fusion is only possible when the receiver is a proficient speaker of the English language. Here, a phonemic reconstruction occurs; even in the condition when *lanket* precedes *banket*. Both fusions are idiosyncratic to vocal sounds and are ambiguous in their perceived qualities. Here, the interest lies in using these sensory tools to probe our speech processing system, aiming to increase the probability of occurrence of vocal ambiguities.

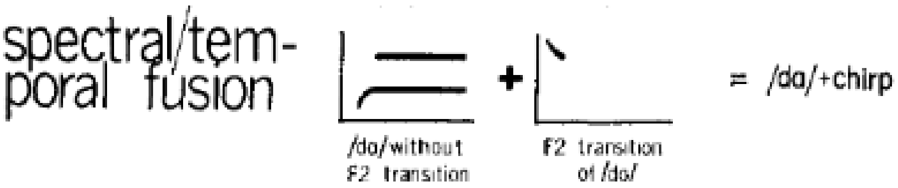


Figure 3.1: Spectral/temporal fusion of the phoneme [dɒ]

3.2 Sound Fusions

Fusions are regarded as perceptual phenomena that involve the integration of two similar or dissimilar vocal sounds into one auditory percept. Two types of fusions are introduced: spectral or timbral fusion, and linguistic fusion. Timbral fusion casts a focus on the perception of vocal timbre, while linguistic fusion addresses phenomena related to language and perception of verbal sounds. Timbral fusion is addressed in the compositions *Enokian Soupe I* and *Enokian Soupe II*, whereas linguistic fusion is present only in *Enokian Soupe II*. These compositions deal with questions and challenges related to vocal arrangement and perceptual hierarchies among the synthetic and natural: the electronic and acoustic.

3.2.1 Spectral Fusion

3.2.1.1 Spectralism: A Timbral Story

When Gérard Grisey composed *Partiels* (1975), he created a music whose form and sound materials are derived from a single acoustic entity, a dynamic spectral analysis of a low E trombone note, heard at the opening (excerpt 11). The relative amplitudes of its partials were then subsequently projected onto the orchestra. In order for the musicians to play this, a transcription of the model into a notational script of the pitches and dynamics of the sound was necessary. This process is also called instrumental synthesis (Hirs 2007, 7). Although exact frequencies are not notated, Spectralist composers often use quarter-tone approximations of the partial frequencies. It can be said that these approximations can result in an unfaithful execution of the individual components. I argue that quantizing the pitch materials would not have changed the eventual outcome, because the Spectralists were not seeking for the absolute re-construction of a sound. They were rather concerned with discerning imperfections and deviations as important factors in achieving ambiguous timbral representation of an initial sound identity. I would argue that the act of imagining, as in predicting the eventual sounding result, plays a vital role in their and my musical process, and allows for certain liberties to be taken on how closely the imagined sound should be perceived.

In addition, each instrument brings its own spectral *baggage* to the resulting sound because of its inherent different physical dimensions, materialities and other complexities. Of significant note is the consistent and deliberate integration of sonic distortion and reconfiguration of the main thematic materials in Spectralist music by reintroducing important themes by altering their partials in a various ways, such as adding, deleting, amplifying or attenuating, and recombining partials. The Spectralists manifested a strong tendency to be inspired by the natural, by the known and heard sounds.

The technological affordance of being able to visualize the harmonic skeleton of any sound, combined with the rise of algorithms for Music Information Retrieval (MIR), has inspired composers to go beyond merely reconstructing a mega-timbre[5] through what Rozalie Hirs calls techniques of "instrumental synthesis"(Hirs and Gilmore 2009). Clarence Barlow called this "Syntrumentation". In his piece *Orchideæ Ordinariæ* (1989) strings are used to resynthesize phrases "why me, no money, my way". Peter Ablinger calls this "Phonographic Realism", where in *A Letter from Schoenberg* (1996), speech is transcribed and played onto a computerized player piano. British composer Jonathan Harvey called this "shape-vocoding" of the instrumental sounds by means of another processed sound (Harvey 2008). The musical element they all share is to push the listeners perceptive ability to familiarize with a sound entity by applying transformations such as harmonic addition and subtraction of partials, ring modulation and resynthesis techniques to recombine certain acoustic features into hybrid sound forms. For instance, in Murail's *L'Esprit des dunes* for eleven instruments and electronics, one's perception of certain *objets sonores complexes* (Garant 2011) can be labeled as ambiguous with regard to the object's origin and production because of the fusion of two spectra. Through spectral cross-synthesis

[5] Livia Teodorescu-Ciocanea mentions in her article "Timbre Versus Spectralism", "Spectral composition of timbre serves two purposes: one is to create a new pitch system based on the overtone series; the other is to reproduce, by means of acoustic instruments, *the structure of certain timbres*; in other words, *to obtain, on a higher level, a megatimbre that evolves in time*" (Teodorescu-Ciocanea 2003, 88).

techniques, a hybrid form of the two emerges. The relative magnitudes of the frequencies of ripping paper are multiplied by the same indexes of a recording of a Tibetan 'dung chen' trumpet (Ibid., 48). This is referred to from now on as a form of spectral fusion. Murail mentions three factors (Ibid., 49-50) that play a role in the spectral fusion of complex sounds, which are, or have a:

a) similar harmonicity of the frequency content
b) coordinated modulation of spectral components in time
c) one's familiarity of the spectral envelope

It is important to note that some of these elements could be in conflict with each other and create perceptually ambiguous situations, which one could label as an auditory mirage (Deutsch 1974). These situations are perceptually attributed to the aforementioned phenomena of masking, rivalry and fusion. Aware of this, Murail frequently reintroduces different resynthesized versions of an important sound identity or theme, with these three factors present to varying extents. In mm. 122-127 (excerpt 12) of *L'Esprit des dunes*, the partially fused sound from the tearing of a paper with the sustained "dung chen" is untraceable to its origin (Figure 3.2). The inharmonic structure of the paper tearing does not correlate with the harmonics of the trumpet sound, and decreases the degree of spectral fusion. However, the presence of the dung chen in the resynthesis is clearer due to being able to perceptually trace back its origin compared with the origin of the torn paper. In contrast, the dynamic articulation of the paper, caused by the ripping and creasing, is more dominant than the sustained dynamics of the trumpet. In the fused sound, some spectral traces of the dung chen trumpet can be observed in its spectral envelope, while also noticing a bright, brassy timbre similar to the overtone structure of the dung chen.

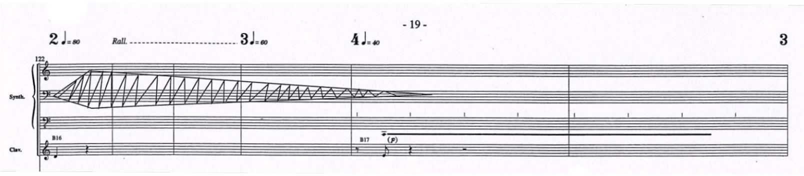


Figure 3.2: M. 122 shows the start of the cross-synthesis between the tearing of paper and a 'dung chen' note in *L'Esprit des dunes*

For many Spectralist and frequency-based composers (Hirs and Gilmore 2009), the voice played an influential role in their musical development, often serving as a central focus for exploring its sonority and the physical realities it produces. For example, in Jonathan Harvey's trilogy *Speakings* (2008) for orchestra and live electronics, the orchestra simulates the learning of speech. In the liner notes of the *æon* CD release of *Speakings*, musicologist Bruno Bossis mentions,

> Speech can be analysed from two points of view: as the communication of a connected language whose words and their organisation are meaningful, *but also as a highly expressive sound texture*(*Jonathan Harvey - BBC Scottish Symphony Orchestra, Ilan Volkov - Speakings* 2010).

This last part resonates with my own approach in listening to and composing with vocal timbres as a way of conveying intention. I use timbre as a communicative effort, mostly detached from language and tied to a more primitive origin of expression. My music should be listened to as to a language whose meaning is mysterious. Where meaning overcomes the listener as being felt, rather than perceived. Harvey's approach to music is of a spiritualistic nature, closely related to Buddhist oral traditions and mantra's. In his work, electronics play an accentuating and contrasting role in relation to the acoustic instruments of the orchestra. This is done by "shape-vocoding" the orchestra, where the spectral envelope of spoken words is applied to the sonic texture of the orchestra.

Peter Ablinger is another composer with an interest in presenting speech in an abstracted form. In *A Letter from Schoenberg*, one can only understand what is being said after the accompanying text is visually shown. Speech is resynthesized in a simplified way, in which most of the acoustic parameters of speech are consciously omitted, leaving the listener to decrypt the text in the "letter". With this piece, Ablinger relies on a perceptual shift to occur, that requires the eye and ear to work together to wash away the ambiguity induced by the seemingly random piano clusters. A technique he refers to as "Phonographic Realism," which focuses on the conceptual and perceptual fusing of speech and music, while Harvey's "shape vocoding" is intended to emphasize the spiritual and expressive potential of the voice within the orchestral context. Both techniques share a similar intention in that the boundary between speech and instrument is blurred in order to tap into a more primitive expressive quality of voice. In my own compositions, I'll clarify my perspective on instrumental synthesis by letting physical models of the human vocal tract share the stage with a human vocalist, and explain how I arrange the synthetic and natural voice material.

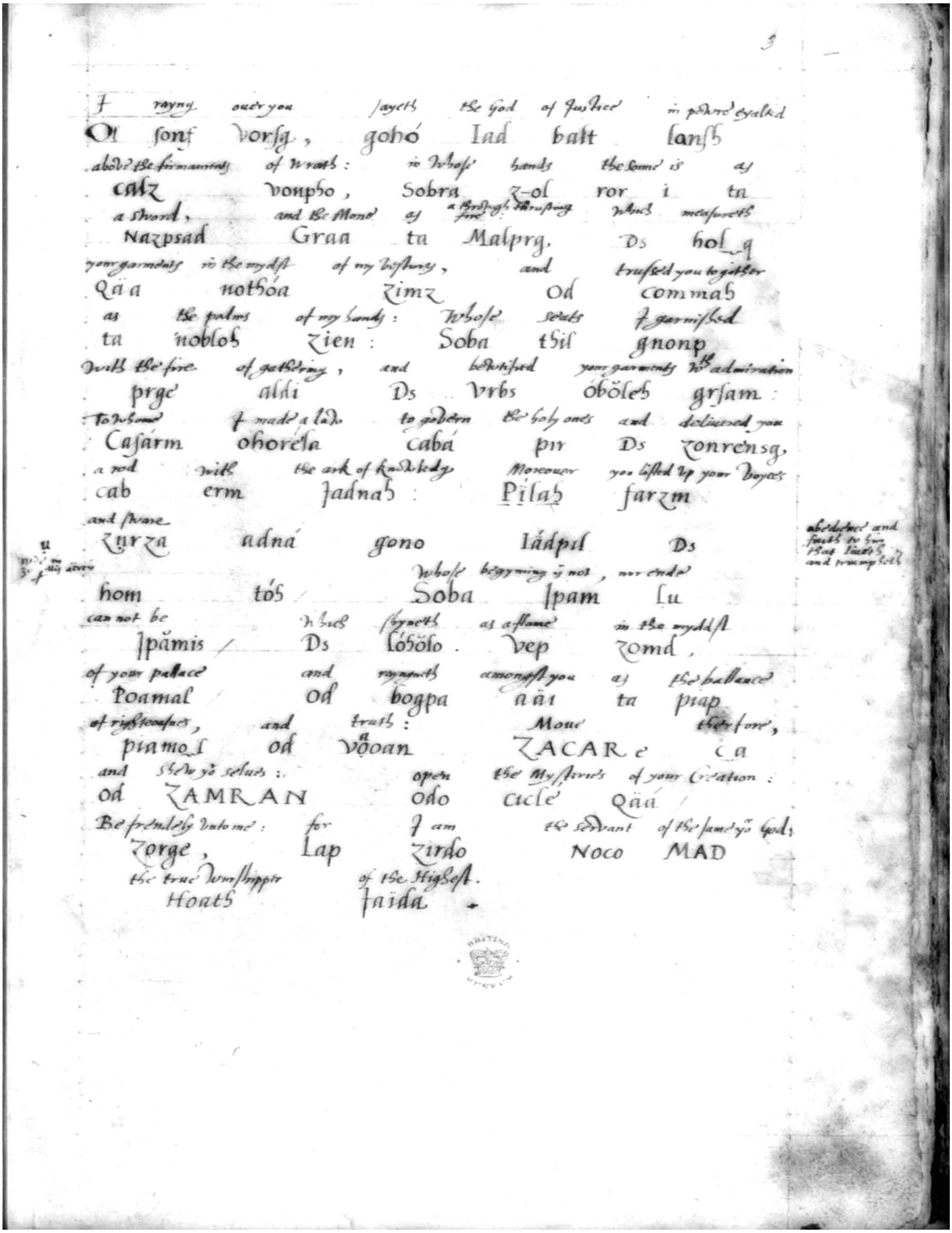
3.2.1.2 *Enokian Soupe I*

Enokian Soupe I is the first iteration in the *Enokian Soupe* series and follows a duet format, consisting of a natural and synthetic voice. It is based partly on text materials from a séance called "The Enochian Keys" written in the Enochian language. The Enochian language is an angelic language that allegedly originated from communication with angels by the Elizabethan scholar John Dee and his associate Edward Kelley in the late 16th century. A sacred language used in occult practices such as Enochian magic. The purpose of the Enochian calls differ depending on the context in which they are used and what the intentions of the practitioner are. The calls are used by some practitioners as a means to communicate with the 30 *Æthyrs* or *Spiritual realms*, or as a means of divination.[6] My artistic interest in this angelic language and the specific scripture of "The Enokian Keys" has two points of origin.

The first emerges from hearing wax cylinder recordings of "The Call Of The First Æthyr" (Figure 3.3) by occultist and spiritualist Aleister Crowley, who, other than being very influential on musicians I'm inspired by, such as Coil, Ossian Brown, David Tibet and writer William S. Burroughs, presents this text with a convincing and hypnotic rhetoric.[7] His tone of voice has a playful yet austere character, which expresses a certain intention behind the invocation. As occultist and ceremonial magician Israel Regardie warns "It is a very powerful system, and if used carelessly or indiscriminately, will bring about disaster and spiritual disintegration."(Griffin 2008, 7) It is also suggested by occultist Donald Tyson as a "means of setting in motion the destructive forces of the apocalypse, as described in the book of Revelation in the New Testament."(Ibid.) The second point of origin derives from listening to the works of the Texas composer Jerry Hunt. Hunt's application of John Dee's symbolism and the Enochian language seems cryptic, and simultaneously, systematic. On his personal web page about the work Jerry Hunt mentions "Since 1974 I have used the angelic tables produced by John Dee through the skrying activity of Edward Kelley as a ground compositional determinant system." He says he uses "sectors of the Angelic Tables from the *Liber Loagaeth*", which is a collection of Dee's writings and instructions related to the practice of Enochian magic (Hunt 2001). In the article "Gesture Modulation of Templates", Hunt applies these angelic tables to generate sound materials in a "sort of integral serialist approach to composition" (Phil Legard, 2013). The way Hunt transformed this hermetically dense text into an instructional tool for composition and performance, exposed me to the musical richness of these scriptures. Phonetically, the language contains dense consonant clusters, such as "malprg" or "znrza". There is also no general consensus amongst specialists on the pronunciation of the language.

[6] These are spirits which are considered to be realms or planes of existence that correspond to different levels of spiritual consciousness.

[7] Bandcamp link "The Call Of The First Æthyr" `https://coldspring.bandcamp.com/track/the-call-of-the-first-thyr-enochian`.



3

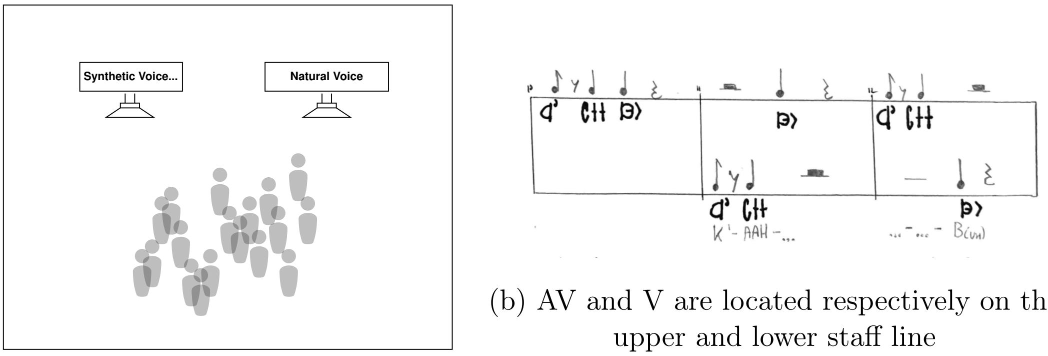
I raygne over you ſayeth the God of Juſtice in power exalted
Ol ſonf Vorſg, gohó Iad balt lanſh
above the firmaments of wrath: in whoſe hands the Sonne is as
calz Vonpho, Sobra z-ol ror i ta
a sword, and the Mone as a through thruſting fire: which meaſureth
Nazpsad Graa ta Malprg. Ds hol-q
your garments in the mydſt of my veſtures, and truſsed you together
Qäa nothóa zimz Od commah
as the palms of my hands: whoſe seats I garniſhed
ta nobloh zien: Soba thil gnonp
with the fire of gathering, and beautified your garments wth admiration
prge aldi Ds Vrbs óbóleh grſam:
To whome I made a law to govern the holy ones and delivered you
Caſárm ohorela cabá pir Ds zonrensg,
a rod with the ark of knowledge. Moreover you lifted up your voyces
cab erm Jadnah: Pilah farzm
and sware
znrza adná gono Iádpil Ds
whoſe beginning is not, nor ende
hom tóh Soba Ipam lu
can not be which ſhyneth as a flame in the myddſt
Ipämis Ds lóhólo Vep zomd,
of your pallace and raigneth amongſt you as the ballance
Poamal od bogpa aäi ta piap
of righteouſnes, and truth: Move therfore,
piamo-l od Vooan ZACAR e ca
and ſhew yo selves: open the Myſteries of your Creation:
od ZAMRAN odo cicle Qäa
Be frendely unto me: for I am the servant of the ſame yor God:
zorge, Lap zirdo Noco MAD
the true worſhipper of the Higheſt.
Hoath Jaida.

obedience and faith to him that liveth and triumpheth

Figure 3.3: The Call/Key of The First Æthyr. Image taken from the Sloane Manuscript 3191 (Dee, Sloane MS 3191, 1).

This, combined with the esoteric nature of the text, created a curiosity to explore this textual and phonic material as a compositional base.

In *Enokian Soupe I*, the natural voice fuses with and masks the synthetic voice, and the synthetic voice does the same to the natural voice. At times, both are clearly distinguishable as two characters, or in the context of a séance, as two individual "mediums", both firing off rhythmic sequences of invocations. A call-and-response behaviour appears by not letting both voices vocalize simultaneously, where the contrast in vocal grain and spoken material is further emphasized by the spatial separation of the voices (Figure 3.4). The piece consists of four parts, each of which articulates the interaction between the two voices differently. The main parameters used to distort or maintain the stability of the fused structure are;



(b) AV and V are located respectively on the upper and lower staff line

(a)

Figure 3.4: (a) shows the speaker setup for *Enokian Soupe I*, and (b) in part two (mm. 10-12) shows an antiphonic relationship between synthetic (AV) and natural voice (V) accentuating the spatial call and response dialogue

a) **Phonation types**: vocal qualities such as modal, pressed, hoarse, breathy, open, which describe the various ways in which the vocal cords can produce sound.

b) **Vocal articulation**: formant positions, timbre, which determine the resonant frequencies of the vocal tract and the color or quality of the voice.

c) **Pitch height** and **dynamic level**: referring to the perceived highness or lowness of the sound and the volume or intensity of the voice.

d) **Fundamental frequency (f0)** and **level contours**: intonation, inflection, which involve the base rate of vibration of the vocal cords and the variation in pitch and loudness over time.

e) **Temporal synchronicity of vocalizations**: timing of events, which ensures that vocal sounds occur in a coordinated and synchronized manner.

f) **Rhythm of vocalizations**: prosody, which includes the patterns of stress and intonation in speech, contributing to the flow and expressiveness of vocal delivery.

The first part begins with no change in these control parameters. The synthetic and natural phonations are vocalized in a monotone, speech-like manner, while avoiding any type of ornamentations in terms of intonation, inflection, and rhythm (Figure 3.5). Both voices seem to be assigned the same character, and imitate the vocal archetype of a "scryer" or "seer", similarly to Crowley's recitation of the Enochian Keys (excerpt 13). All of the events contribute to a repetitive and static flow of time due to its temporal quantization. Fusion hardly occurs because of the two voices being sonically distinct in timbre. There is, however, an incomplete fusion that transpires from a narrative origin; both voices are reading the same text in the vocal style of Crowley and the vocal persona only partially fuse. Their distinct vocal quality and gender inhibit the fusion to occur. All in all, the opening phrases show both voices adhering to a unison-like or fused form.

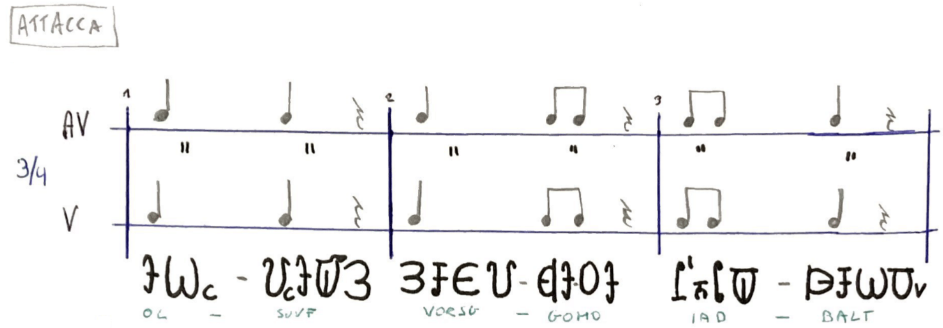


Figure 3.5: The exposition (mm. 1-3) of *Enokian Soupe I* shows the synthetic (AV) and natural voice (V) vocalizing identical material in a fixed 3/4 rhythm. *Enokian Soupe I*

In part two, m. 25, the voices break free from the static speech rhythm (excerpt 14).(Figure 3.6) Words from "The Call Of The First Æthyr" are atomized into their phonetic sounds and recombined or repeated. The phones [s] and [z] of the word "zorensg" are repeated, sustained and phonetically interpolated between. These gestures instigate a listening attention to the vocal qualities of the synthetic and natural. For this timbral fusion to take place, their intensities need to be relatively small. Only minute details such as dynamic variations (e.g., decreasing lung pressure) or slightly different mouth resonances make the fusions less stable. These variations cause short, yet perceivable oscillations between the synthetic and naturally produced sounds. In m. 32, a singing-voice quality commences, signaled by the strong articulation of [z]. First, timbral fusion occurs due to both voices vocalizing the same starting note, and in the ensuing phrases, glissandi in contrary motion occur, where one voice glides upwards and one slightly downwards. This divergence in frequency causes the fused sound-image to split, and emphasizes the separation of the synthetic and natural vocal persona. Second, phonetic variations in

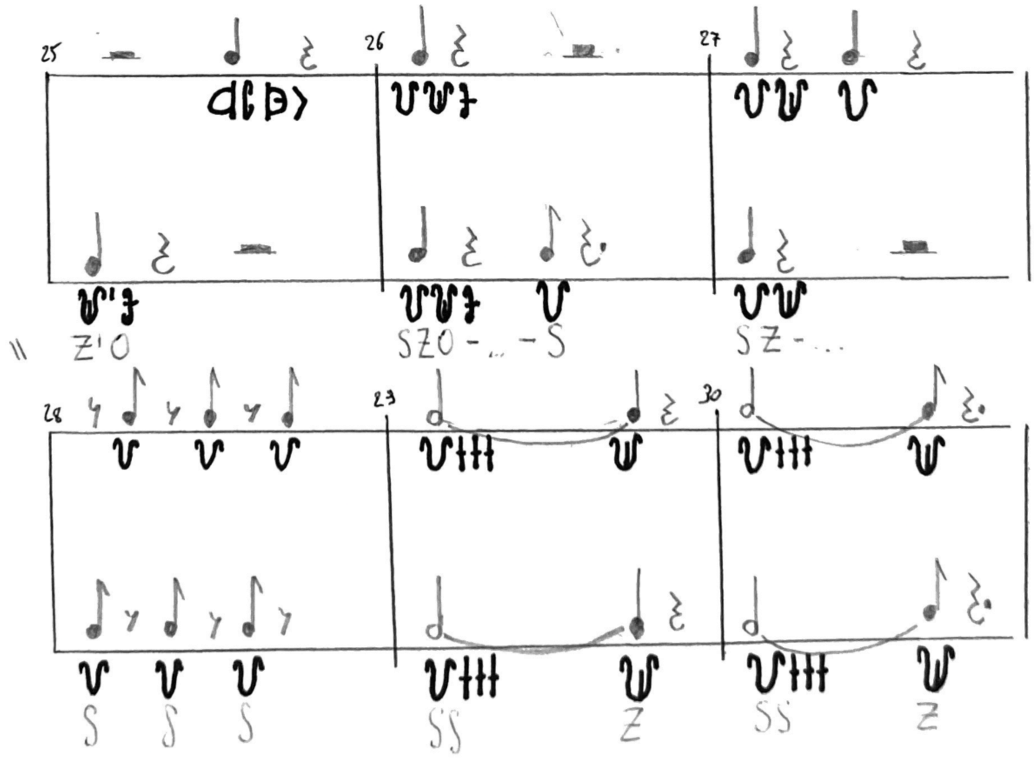


Figure 3.6: In part two (mm. 25-30) rivalry and fusion between synthetic and natural voices of voiced and unvoiced uvular fricatives [s] and [z]. *Enokian Soupe I*

mouth configurations are introduced. For example in mm. 32-33, the synthesized stressed vowel [ɐ̰ˈ] is contrasted by the vowel [o̞ˈ] of the natural voice. The synthetic and natural voices then transitions towards [o̞] and [ɐ̰] in excerpt 15 (Figure 3.7). These phonetic variations act as destabilizing influences on the fused percept. The composition of an event involving timbral fusion between a synthetic and a natural voice in a musical context necessitates the presence of a *bridging language*; one that enables both the synthesist and the vocalist to effectively produce and coordinate nonverbal sounds. Small timbre changes affect the stability of timbral fusions, and relate directly relate to small physiological movements of the involved articulators.

In the third part, sections are timed by a *Free-Roaming Pulse* (FRP), where both voices start vocalizing after getting a cue from a clicktrack. *Free-roaming* means that the section is not fixed to the same temporal grid as is the case in previous parts one and two. The sections are separated arbitrarily, but each section has a specified fixed tempo and time signature that still slightly expresses the repetitive and hypnotic rhythm of the séance. The phonetic content does not resemble the Enochian text material anymore, and the previous phonetic divergences between the voices become more extreme, see excerpt 16 (Figure 3.8).

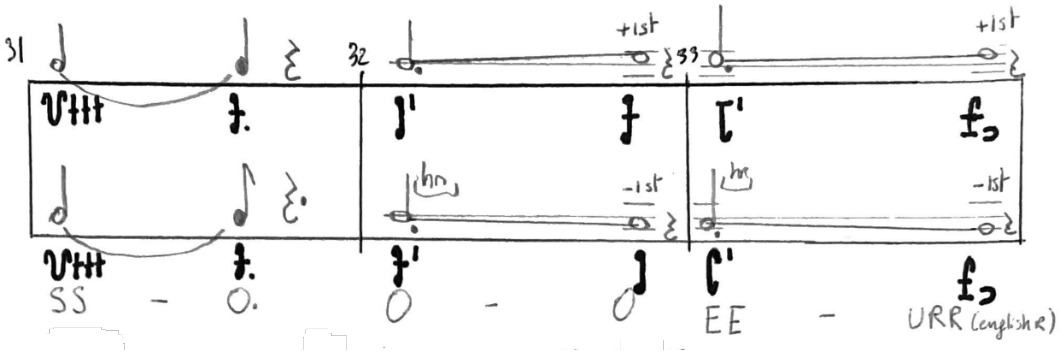
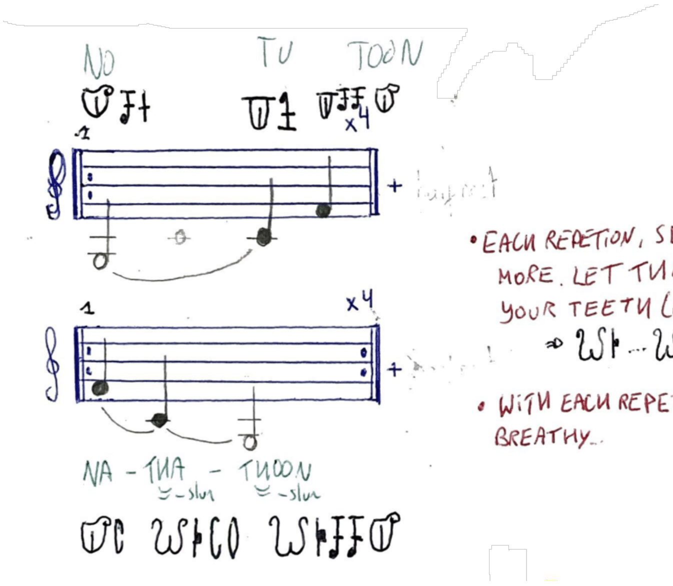


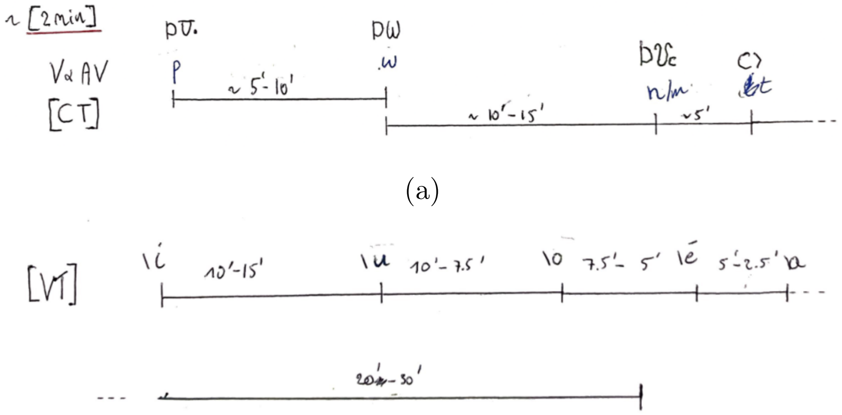
Figure 3.7: In part two (m. 32), glissandi in contrary motion in the synthetic (top) and natural (bottom) voice cause the fused image to split; in m. 32, the natural and synthetic voice respectively vocalize the vowels [o̞ˈ] and [ɐ̰ˈ], where the natural and synthetic voice subsequently transitions towards the unstressed vowels [ɐ̰] and [o̞]. *Enokian Soupe I*



- EACH REPETION, SLUR YOUR PRONOUNCIATION MORE. LET THE TONGUE REALLY SLAP AGAINST YOUR TEETH (USE MORE saliva if needed)
- WITH EACH REPETITION, BOTH VOICES ALSO GET MORE BREATHY..

Figure 3.8: Shows 5 repetitions of one motif, with each repetition, the phonation is more slurred (part three, section 8). *Enokian Soupe I*

Eventually, in the fourth part, timbral fusions happen as short-lived events. The probability that both voices fuse, or even come close to a single perceptual entity, is low. Two self-constructed rules lay down the foundation for this ending. The first section adheres to the "Consonant Trajectory" (CT) rule. In this mode, both voices improvise vowels in between fixed consonants for a given duration of time. For instance, in Figure 3.9(a), the CT's consonant is [p] and is enveloped by transient vowels. One can clearly hear [p] take on different acoustic forms as it is being coarticulated by [i]'s and [a]'s (excerpt 17). In Figure 3.9(b), the voices follow the "Vowel Trajectory" (VT) rule, which is nothing more than inverting the roles of consonants and vowels in the CT. The vowels are now the temporary fixed phonetic units, swarmed and barraged by arbitrary consonants, as heard in excerpt 18. Both frameworks operate within a constrained improvisation setup, in order to stray even further away from the monotone quality present in part one. Intriguingly, strings of random tones and clicks expose the timbral opulence of a single phoneme. In speech perception this is known as the "Lack of Invariance Problem," where speech sounds are coarticulated into the context of an utterance, so their patterns change in different phonetic contexts (Heald, Klos, and Nusbaum 2016, 197-198).[8]



(a)

(b)

Figure 3.9: (a) The Consonant Trajectory (CT) and (b) Vowel Trajectory (VT) in *Enokian Soupe I*

3.2.1.3 *Enokian Soupe II*

I decided to extend *Enokian Soupe II* by sequencing synthetic and natural consonant-vowel clusters in various configurations and patterns.[9] The first three parts from the first iteration serve as interludes between the four new parts, repurposing large and unchanged sections in a "collage"-like manner. Parts one, two, and three of the original piece are inserted between sections A,B,C, and D of the new one. The previous contextual links are cut and recombined to form new narratives, possibly enhancing or disrupting the continuity of the musical whole.

[8] Also known as "Phonetic Context Variability".

[9] In a similar way, the poet Enzo Minarelli describes the approach to vocal poetic experimentation of Raoul Hausmann and the Russian Futurists in an interview as "...oriented towards the exploration of single phonemes in a rational pattern." (Minarelli 2019)

In section B, the interplay between vowel timbres of synthetic and natural voices is explored through slow interpolations between vowels. Vowels are usually described by three formants (i.e., F1, F2, and F3) that encapsulate the acoustic energy surrounding certain frequencies. Notably, evidence shows the possibility to distinguish vowels by presenting only F1 and F2 (Sakayori et al. 2002), and is supported by studies on vowel identification using synthetic vowels, where F1 and F2 were found to be the primary cues in speech comprehension (Dorman et al. 1989). Within phonetics, vowel height and vowel frontness are two dimensions used to classify vowel sounds and are respectively related to F1 and F2. They describe where in the mouth the tongue is positioned when producing a vowel sound, both vertically (i.e., height) and horizontally (i.e., frontness). Each dimension is classified in three categories as shown left in Figure 3.10 and the corresponding F1/F2 map.[10]

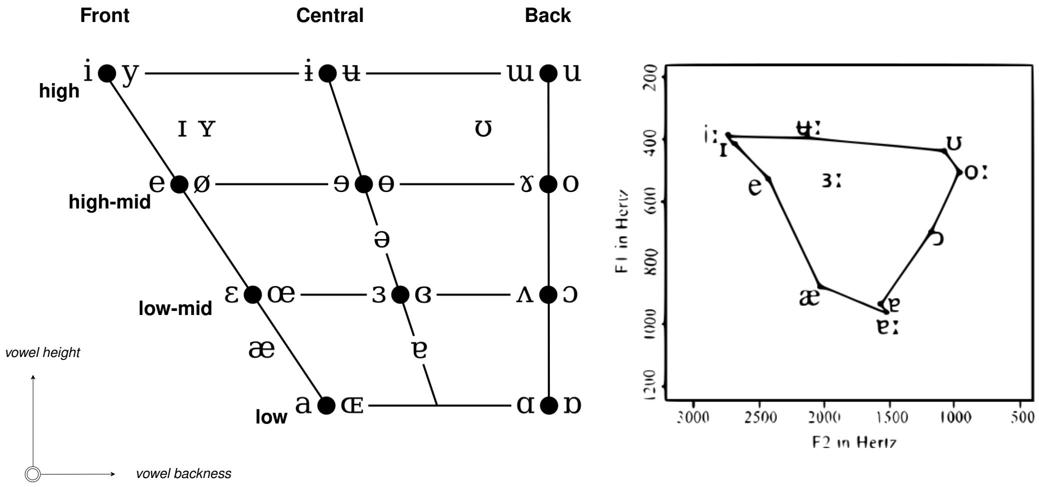


Figure 3.10: Left, a vowel diagram of cardinal vowels ([a],[i], etc. . .) and diphthongs or glide vowels ([ɪ],[ʊ], etc. . .); Right, a vowel map (Cox and Warner 2017).

In this section, both voices explore the vowel space in a systematic fashion. In each phrase, the voices interpolate between the same pair of cardinal vowels and/or semivowels. In m. 4a, from [ø] to [i], and in m. 3b, from [e] to [ɑ] (excerpts 19 and 20). Both transitions are shown in Figure 3.12. The vowels are chosen by a serialisation procedure. Two spirals are drawn on the vowel diagram, one for each iteration. In Figure 3.11, the tangential points on the spiral curve are used as the next value in the series, resulting into two series of eight vowels:

a) [**ɑ → ɔ , ɶ→ ə , ɛ → u , ø → i**]

b) [**a → æ , œ → ɜ , e → ɑ , o → y**]

Timbral fusion occurs when the individual formants of both voices align, and relative loudness and relative fundamental frequency (f0) are small. Notably, the natural voice relates to a female vocal

[10] A vowel diagram for 60 adolescent female speakers.

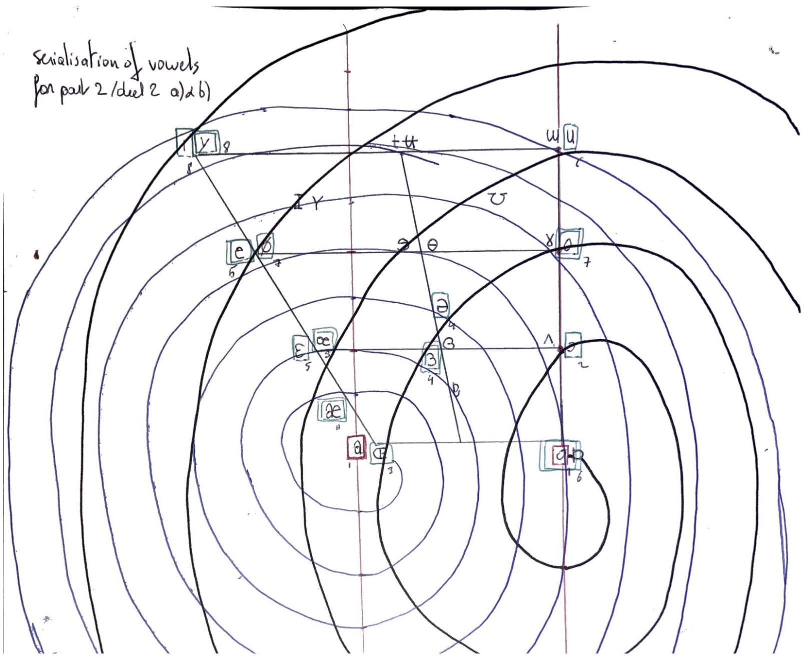


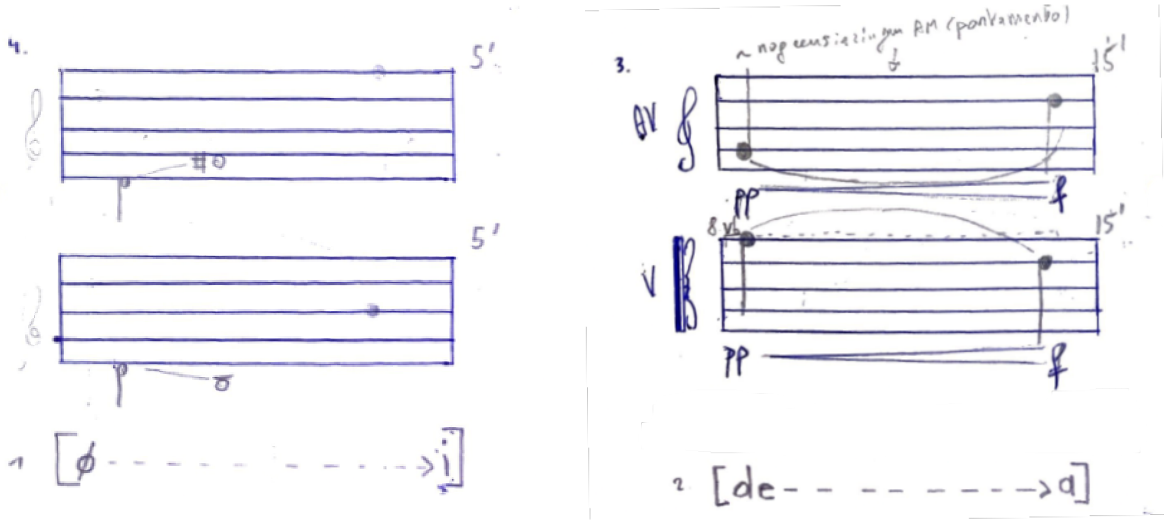
Figure 3.11: Serialisation procedure of tangential points to the spiral curve on the vowel diagram for section B. *Enokian Soupe II*

archetype, while the synthetic voice tries to embody a male vocal archetype. Research shows that formant frequencies appear to be higher in females than in males due to differences in the shapes and sizes of the vocal tract (Smorenburg and Chen 2020). When going from one vowel to another, the frequencies and bandwidths also change. Here, the partials of the synthetic and natural voices could momentarily overlap. In other words, the F1 or F2 trajectories intersect and intermodulate, resulting in short events of timbral fusion.

I would argue that for timbral fusions to occur, two additional features need to be similar in both voices, being vocal jitter and shimmer patterns. Voice jitter is the amount of f0 variation from vibrotory cycle to cycle, while shimmer relates to the amplitude variation of the sound wave (Teixeira, Oliveira, and Lopes 2013, 2).[11] To achieve this, the synthetic voice uses pitch and intensity values extracted from recordings of the natural voice. These recordings contain vocalizations meant for the synthetic voice as notated in the upper staff line in the score in Figure 3.12. The extracted data contains the minimal variations in fundamental frequency and amplitude (i.e., jitter and shimmer), and are idiosyncratic to each person. The resynthesis of these frequency and intensity time series is done by the synthesis system NKOAPP.[12] The resulting artificial voice has a vocal quality similar to the natural voice, where the naturalness and presence stand out as the principal shared features with the archetype of a mezzo-soprano voice (i.e., the natural voice performer). The recordings of the natural voice can be heard in

[11] Vibrotory cycle is the open and closing phase of the vocal folds ("Understanding Voice Production - THE VOICE FOUNDATION" 2013).

[12] discussed in Chapter 4



(a) M. 4a from [ø] to [i]

(b) M. 3b from [e] to [ɑ]

Figure 3.12: Measure 4a (a) and 3b (b) from Section B in *Enokian Soupe II* show the synthetic and natural voice phonetically interpolating between vowels

excerpt 21 from m. 2a, and the resynthesized version in excerpt 22. In excerpt 23, the recordings are slightly panned left and right in order to increase the probability for timbral fusions to occur.

Describing timbral fusion purely through analytical means proves to be a tricky problem, as multiple factors play part in its process; acoustic and perceptual indices of the natural and synthetic voice have to align in order for this phenomenon to occur. As stated in Section 2.2.1, the thing we are trying to measure between the natural and synthetic voice is a perceptual *distance of likeness*. And, in this type of fusion, this *distance* comes from a *similarity* in vocal timbre. Timbre can be defined as "the perceptual quality" of sound, and therefore, trying to find a metric to describe an auditory phenomenon necessitates an understanding of the neural mechanisms involved in auditory processing within the brain. Through computational analysis, an efficient way to describe verbal sounds in a compressed and perceptually inclusive manner, is by computing its "Short-term Power Spectrum" or "Mel-Frequency Cepstrum" (MFC). This description is made up out of a low number of parameters, called the "Mel-Frequency Cepstral Coefficients" or MFCC's. This analysis is often used with verbal tasks because it mimics the human auditory system's response to sounds; in other words, it models the cochlea. Two time series of 13 MFCCs of the natural and synthetic voices are analyzed, and their mean Euclidean distances $d_{i,n}$ are calculated for each frame n with $i \in \{1, 13\}$. Then, the arithmetic mean of the sum of distances $d_{i,n}$ is calculated in Equation (3.1) to get the timbral distance D_n or *distance of likeness.*

$$
\begin{aligned}
d_{i,n} &= \sqrt{(MFCC_{nv}[i] - MFCC_{av}[i])^2} \quad (\forall n) \\
D_n &= \sum_{i=1}^{13} \frac{d_{i,n}}{13}
\end{aligned}
\tag{3.1}
$$

A real-time implementation in Supercollider for the timbral distance calculation between natural and synthetic voice is shown in Listing 3.1. If D_n is low (< 20), a timbral fusion is likely to happen only if the relative fundamental frequency $\Delta f0_n$ and loudness ΔL_n are small. This approach is still in its experimental stage. Further investigation and testing with other sounds still needs to be done. For now, this timbral distance is used as a suggestive measure in predicting when a timbral fusion might occur.

Listing 3.1: Supercollider implementation for real-time calculation of timbral distances D_n between synthetic and natural voice

```
// the more similar the timbres the lower the measured "distance" between them will be
(
~playSynAndNatVoice = {|synVoiceBuf, startCoeff = 1, numCoeffs=13, updateRate=100|
    var synVoice = PlayBuf.ar(1,synVoiceBuf,BufRateScale.ir(synVoiceBuf),loop:0);
 var natVoice = PlayBuf.ar(1,natVoiceBuf,BufRateScale.ir(natVoiceBuf),loop:0);
 var impUpdate = Impulse.kr(updateRate); // 10 ms frame duration
 var sigComb = [synVoice, natVoice];
 // extract 13 mfcc's, don't include loudness -> startCoeff = 1
 var mfccNatVoice = FluidMFCC.kr(natVoice,numCoeffs:numCoeffs, startCoeff:startCoeff);
 var mfccSynVoice = FluidMFCC.kr(synVoice,numCoeffs:numCoeffs, startCoeff:startCoeff);
 // mean value of 13 distances between 13 mfcc's
 var meanDist = Mix((mfccSynVoice - mfccNatVoice).squared).sqrt;
 // send to sclang
 SendReply.kr(impUpdate,"/dists",meanDist);
 Out.ar(0,sigComb);
};
{~playSynAndNatVoice.value(
 synVoiceBuf:~synVoiceBuf,
 natVoiceBuf:~naturalVoiceBuf,
 startCoeff:1)
}.play;
// receive distances and poll in post window
OSCdef(\dists,{
 arg msg;
 {"*".post} ! (msg[3] / 5);
},"/dists");
)
```

3.2.2 Linguistic Fusion

Certain spectral and frequency-based musical techniques for achieving timbral ambiguities are extended by introducing fusions that rely on psycho-acoustic and psycho-linguistic phenomena, called *Linguistic Fusions*. Linguistic fusions are experimented with in *Enokian Soupe II*. Most of the inspiration for expanding this investigation in fusions of natural and synthetic voices into the domain of speech, originates from a 70s paper titled “Auditory and Linguistic Processes in Speech Perception: Inferences from Six Fusions in Dichotic Listening” (Cutting 1976) by cognitive scientist James E. Cutting. In this research, he examines six types of fusions by how the brain integrates and processes auditory information from both ears. Notably, most of his experiments are carried out by sounding synthetic and natural voices in separate ears.

a) **psycho-acoustic fusion** or “Fusion of Proximal Acoustic Features by Perceptual Averaging”

b) **phonological fusion** or “Perceptual Construction of Phoneme Clusters”

Linguistic fusions are seen as a subset of spectral an timbral fusions, where a linguistic context is the additional mental component present in the fused auditory percept. These fusions are perceptually incidental (e.g., in psycho-acoustic fusion) or accidental (e.g., in phonological fusion) in their combinatorial processes. In contrary to the dichotic nature of the experiments trialed by Cutting—hearing two different stimuli separately in the left and right ear—, *Enokian Soupe II* brings these experiments out of the laboratory and into an *acoustic space.* A space that reflects and absorbs sound. Instilling a perception of distance, which stands in contrast to the direct, loudspeaker-to-ear approach of Cutting in the presentation of his vocal stimuli. This space invites the mixing and masking of the synthetic with the natural voice in a binaural way. In rehearsing this piece with a vocal performer, the spatial setup (i.e., amount of speakers, seating area for audience, directivity and direction of speakers, and speaker symmetry) turned out to be critical in facilitating and sustaining the fusion-like moments. This presence of a spatial awareness by the composer is necessary, not only for achieving linguistic fusion, but also for timbral fusion.

3.2.2.1 Psycho-acoustic fusion of phonemes

Combining the synthetic phoneme [gɒ] with a natural voice [bɒ] at a relatively similar fundamental frequency, intensity, and onset-timing, results in the perception of a single, intermediate phoneme [dɒ] (Figure 3.13). The vocal percept is intermediate, because the second-formant transitions, or F2 transitions, of [dɒ] lie between the F2 transitions of [bɒ] and [gɒ].[13] (Figure 3.14)(Cutting 1976, 118).

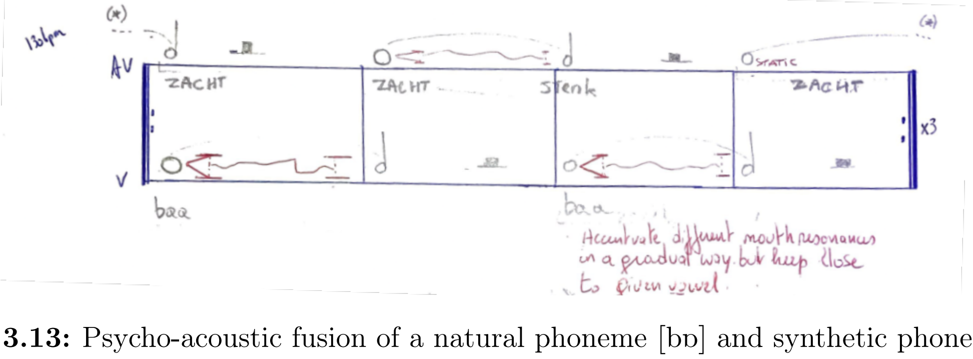

130bpm
(*)
(*)
AV
V
ZACHT
ZACHT
Sterk
STATIC
ZACHT
x3
baa
baa
Accentuate different mouth resonances in a gradual way but keep close to given vowel

Figure 3.13: Psycho-acoustic fusion of a natural phoneme [bɒ] and synthetic phoneme [gɒ] perceived as [dɒ] (*Enokian Soupe II*, Section A, measure 6)

[13] Cutting states that “. . . that the *proximity* of the second-formant transitions in the to-be-fused pair to the /da/ boundaries is important to the phenomenon.”, and “it is the averaging of optimally similar transitions that is crucial.”

This is called a psycho-acoustic fusion of phonemes and can be heard in excerpt 24. Cutting says "it appears that this fusion is psychoacoustic rather than psycholinguistic, since it is quite sensitive to phonemically irrelevant acoustic variation in the stimuli."(Cutting 1976, 123) These fusions occur as well with other neighbouring vowels, such as [ieɛɔou]. The mental image that emerges when the syllable pairs fuse, is ambiguous, and exhibits the oscillatory behaviour akin to perceptual rivalry. If the phoneme pairs' relative onset time and/or intensity difference between natural and synthetic voices are large (i.e., $> 20 milliseconds$), fusions are less likely to happen (Cutting, 125-130). A relative intensity difference of 10 dB is enough to hear the two voices as separate sounds (Cutting, 130-131). Aiding further in the ambiguous quality of this phenomena, is the apparent difference in vocal timbre between the synthetic male and natural female voice. One linguistic unit is perceived, yet this vocal sensation carries with it a hybrid sound halo. An aura that emerges from a dissonance in vocal features and impregnates the sound with a sense of multiplicity.

In addition to the careful manipulation of vocal parameters present in *Enokian Soupe I*—such as phonation types, articulation, pitch height, dynamics, and temporal synchronicity—, psycho-acoustic fusion serves as a tool to orchestrate the convergence of the natural and synthetic voices in *Enokian Soupe II*. A vocal experience is enriched by allowing extra-musical and linguistic relationals to appear in the listener's mental space.

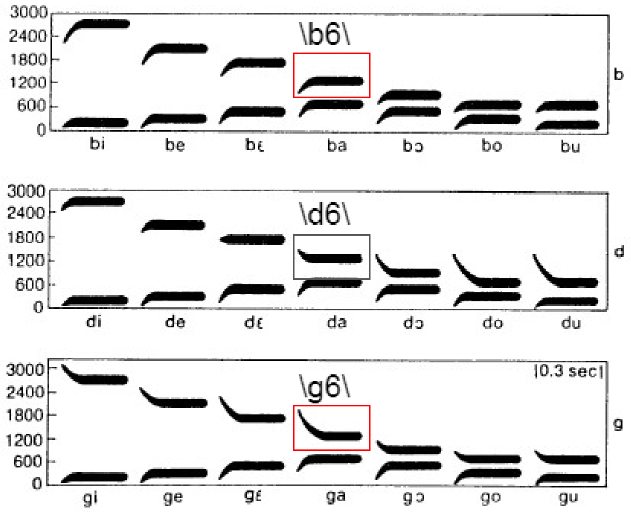


Figure 3.14: Formant transitions for voiced stop-vowel syllables [ieɛɔou] ("3.2. Acoustic Aspects of Consonants – Phonetics and Phonology")

3.2.2.2 Phonological fusion of phonemes

As mentioned before, I use phonic material—mostly sub-phonemic—to convey a sense of meaning or thought in an ambiguous form. In the timbral fusions, sounds containing similar acoustic and psycho-acoustic features are composed to achieve moments of a singular vocal texture, while still expressing traces of their original constituents. Another type of linguistic fusion is introduced, namely the *phonological fusion of phonemes.* Phonological fusion is described by Cutting as follows:

> Phonological fusion occurs when two inputs, each of n phonemes, yield a response of $n+1$ phonemes. (Cutting 1976, 121)

When the nonsense words *banket* and *lanket* are simultaneously heard separately in each ear, one is likely to perceive the word *blanket.* The two words fuse to form a longer and linguistically more complex word than either of them. This is interesting due to the fact that a perceptual reconstruction occurs, where the percept is singular and lexical. Two meaningless vocal entities merge to form a semantic unit, and this only occurs when both strings contain enough phonetic information to construct the word. In *Enokian Soupe II*, this linguistic fusion is experimented with in the final part (Figure 3.15). The synthetic and natural voice both start of with repetitively vocalizing the initial phonetic segments of respectively [r] and [l]. After a duration of four staccato vocalizations of the same phonetic segment, both voices aggregate the following phonetic segments [o] and [a] in the string. When the the natural and synthetic voice phonetically arrive at their nonsense word, only then does the fusion occur. An effect that is experientially similar to the sudden comprehension of the text in Ablingers' *A Letter from Schoenberg.* Suddenly, it becomes clear that both voices are trying to speak, and their message can only be understood when both voices *give* the listener enough phonetical information.

Phonological fusions enter the phonetic, phonemic, morphemic, and lexical domain of speech. Words are segmented into phonemic clusters, which are then atomized into phonic units and recombined to build an abstract musical language with an undefined grammar. This approach to speech, particularly verbal sounds, perceptually amplifies the phonetic properties of language and the linguistic energies associated with vocal articulations. In my artistic endeavors, I felt compelled to strip language of its lexical semanticity. This process is described by Lettrist and sound poet Isidore Isou as the "chiselling phase," which naturally follows the "amplic phase," where "form is expanded and pushed to incorporate whole swathes of experience and existence." (Cooper 2019) In other words, this is where language reaches its limit to grow in expressive potential. This post-structuralist approach to speech inspired me to deconstruct language and reformulate it into an abstract language based on the sonic particularities of speech. The synthetic voice aids in this abstraction by blurring the line between nonsense utterances and sensible language. In the case of nonsense utterances or gibberish, there is no systematic grammar to construct semantic content. Here, semanticity emerges from various linguistic, pragmatic, and cognitive mechanisms, such as associations evoked by phonetic and phonological patterns or by paralinguistic cues.

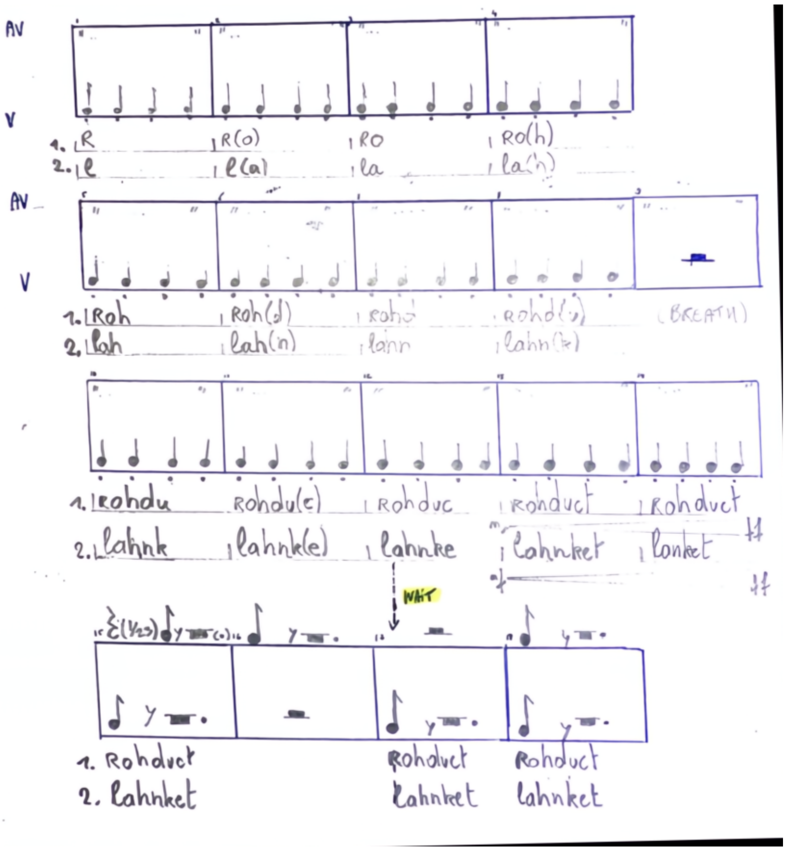

AV
V
1. R R(o) Ro Ro(h)
2. l l(a) la la(h)
AV
V
1. Roh Roh(d) roh rohd
2. lah lah(n) lahn lahn(k)
(BREATH)
1. rohdu rohdu(c) Rohduc Rohduct Rohduct
2. lahnk lahnk(e) lahnke lahnket lanket
WAIT
1. Rohduct Rohduct Rohduct
2. lahnket lahnket lahnket

Figure 3.15: A phonological fusion of a synthetic phoneme string *lanket* with the naturally vocalized string *banket* into the word *blanket* (*Enokian Soupe II*, Section D)

3.3 Conclusion

This chapter discussed different types of fusion that can occur when working with synthetic and natural voice. Ideas and techniques from the Spectralist movement and frequency-based composers regarding fusion and instrumental synthesis are framed and reinterpreted within the scope of the physical modelling paradigm. The analysis and approaches presented in the duets *Enokian Soupe I* and *Enokian Soupe II* led to the conclusion that while vocal fusions can take on many perceptual forms, they share commonalities in how they are perceived sonically. Whether as a timbral or linguistic fusion, they produce what I term 'vocal ambiguities,' which, through their evoked mental images, enrich the vocal experience rather than diminish it through obscurity. The synthetic voice serves as an investigative musical tool to express alternative meanings through vocal sounds, and not just through language, highlighting its linguistic and timbral relationship to the natural voice.

Chapter 4

NKOAPP: An Articulatory Speech model

NKOAPP is a program written in Python, that utilizes open source programs Praat[1] for sound analysis and the synthesis backend of VocalTractLab[2], which together allow for an advanced control of the vocal tract shape and glottis properties over time. It functions as a tool for the synthesis of mouth sounds, such as accurate speech, singing voices, nonsense utterances, and non-human articulations. NKOAPP was used to generate all the synthetic material for the artifical voice in the iterations of the *Enokian Soupe* series.[3] As mentioned in Section 2.1.2, the system should be able to express a wide sounding scale from abstraction to realistic image. To allow for this wide sounding scale, NKOAPP operates in various modes to address a multitude of vocalizations (e.g., phonemes, glottal clicks, whispers, mouth pops) and articulatory transitions between phonatory targets (e.g., extremely slow or fast movements). Relating back to Bergsland's notion of a *maximal and minimal voice continuum*, NKOAPP invites the user to experiment around the perceptual thresholds of a stable vocal image or proto-voice. In Appendix C, an instruction manual is provided on how to use the synthesis and control system.

4.1 Technical development

The NKOAPP program was initially designed to generate vocalizations and articulations ranging from natural to unnatural. It achieves this through continuous or stepped transitions between presets representing different mouth and glottis configurations. NKOAPP facilitates both automated and manual

[1] `https://github.com/praat`

[2] `https://github.com/TUD-STKS/VocalTractLab-dev`

[3] An abbreviation for *Neus-Keer-Oor Application*. Where *Neus*, *Keel* and *Oor* respectively mean nose, throat and ear.

manipulation of the physical properties of the vocal tract and glottis (i.e., invariants). The application allows for the adjustment of acoustic parameters (i.e., the variants), such as fundamental frequency, intensity, and vocal tract shapes.

4.1.1 Synthesis and Control program

Similar to the system employed in composing *Keelcore*, the synthesis procedure follows the source-filter paradigm, where the vocal fold vibrations serve as the excitations of the filter. The vocal tract acts as a filter for the glottal pulses. The vocal fold vibrations and the vocal tract filtering are both modelled in VocalTractLab. VocalTractLab is developed as a software for high precision articulatory speech synthesis and analysis. The fully parametric human speech model is meant to be used by speech scientists and phoneticians in order to simulate and analyze the articulator movements involved in various speech tasks in a non-intrusive manner, unlike high-speed laryngoscopy and nasolaryngoscopy. It came to my attention that, as a music-making tool, the vanilla VocalTractLab interface limits the musician by imposing ranges that are designed to simulate non pathological, natural speech as accurately as possible. The motivation to hack this control system, is to break free from the humanized constraints of the available control parameters, which are especially present in the form of simulating the systems' physical inertia. Dissolving the idea of needing a spatiotemporal resolution, specific for each type of vocal sound, allows for the synthesis of humanly impossible articulations. For example, the spatiotemporal resolution of tongue movements[4] produced by a person lies between 50 and 125 milliseconds — at a spatial resolution of 2.8 to 3.8 mm^2 (Lingala et al. 2016, 16-19). In the system, the temporal resolution for any kind of speech task is bounded solely by the chosen sample rate. At a sample rate of 44.1 kHz, the highest resolution is 2.5 milliseconds, or 400 Hz — at a spatial resolution dependent on initial physical constraints. For example, the spatial constraints for the jaw movement on the X axis (JX) are $[-0.5mm; 0.0mm]$, and the constraints of the jaw angle are $[-7.000deg; 0.000deg]$. See Figure 2.2 for the spatiotemporal resolutions of certain speech tasks (Ibid., 17). Transitions between phonation modes can therefore occur at rates reaching up to 400 Hz (excerpt 25).[5]

In the control section of NKOAPP, the user defines preset sequences for the glottis and vocal tract models. Each element in the sequences starts with its *source-preset* and ends with a *target-preset.* A preset is a list of physical parameters that is needed to synthesize a particular phone. In order to produce the vowel [a] in a modal phonation, a vocal tract preset is required for the vocal tract shape, and one for the glottis excitation mode (e.g., modal phonation), forming a *preset pair.* The presets are retrieved from what is called a *speaker file*, which defines a model speaker. This file comprises the

[4] Such as in vowel-to-consonant (VC) transitions.

[5] Phonation modes such as modal, pressed, hoarse, among other.

anatomy of the speaker, the vocal tract shapes used to produce individual phones, as well as model properties for the glottal excitation (Figure 4.1). Adding and removing presets is done by a module that edits and reconfigures the entries of the speaker JSON file. In Table 4.1, the variable names are listed for vocal tract shape and glottis excitation modes. One of the three glottis geometries—*Geometric*, *Two-mass*, and *Triangular*— is selected before starting the synthesis procedure. This option can not be changed during synthesis of the vocalizations.

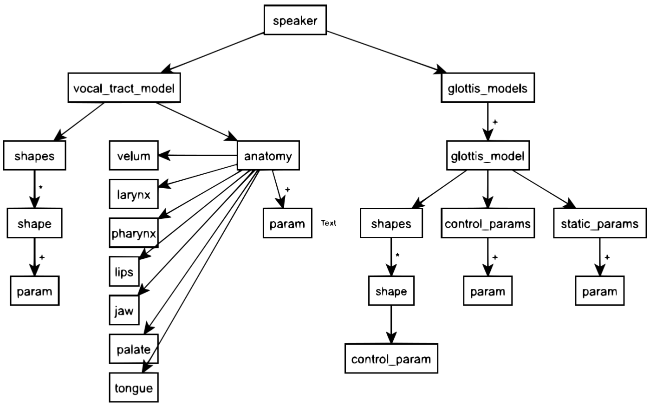


Figure 4.1: NKOAPP *speaker file* tree diagram structure

The musician can generate N number of source-target pairs either by manual input or by automated sequencing. Interpolating between N source-target pairs, is performed by a spline interpolation algorithm, namely the "Akima spline" (Akima 1970). Each source-target pair can also be differentiated to achieve cubic, squared, linear, or stepped curves between the glottis and vocal tract source-target pairs. The trajectory between presets or physiological states of the articulators is now numerically calculated by this algorithm. Notably, this overwrites the interpolation algorithms active in the vanilla VocalTractLab synthesis system, ignoring coarticulative effects during synthesis. Coarticulation refers to how the pronunciation of one sound is influenced by the preceding and following sounds; a process

Table 4.1: Control parameters for glottis and vocal tract contained within a preset in NKOAPP

Vocal Tract	**Glottis Geometric Model**$_{GM}$	**Glottis 2-Mass Model**$_{2M}$	**Glottis Triangular Model**$_{TRI}$
Horz. hyoid pos.(HX)	f0	f0	f0
Vert. hyoid pos.(HY)	Subglottal Pressure	Subglottal Pressure	Subglottal Pressure
Horz. jaw pos.(JX)	Lower displacement	Lower rest displ.	Lower rest displ.
Jaw angle (JA)	Upper displacement	Upper rest displ.	Upper rest displ.
Lip protrusion (LP)	Chink area	Extra arytenoid area	Extra arytenoid area
Lip distance (LD)	Phase lag	Damping factor	Aspiration strength
Velum shape	Relative amplitude		
Velic opening	Double pulsing		
Tongue body X	Pulse skewness		
Tongue body Y	Flutter		
Tongue tip X	Aspiration strength		
Tongue tip Y			
Tongue blade X			
Tongue blade Y			
Tongue side elev. 1			
Tongue side elev. 2			
Tongue side elev. 3			

specific to the human vocal apparatus. See Figure 4.2 for an example of the synthesis of the phrase [aoɢ] (excerpt 26). Excerpt 27 demonstrates a vocalization with an N number of random interpolations between the same set of glottis modes and vocal tract presets as the previous excerpt. A helpful measure for understanding the rate at which synthetic articulations occur, is the interpolation rate IR(Hz). This value describes the chosen amount of interpolations N(-), to the total duration T_{total}(seconds) of the vocalization (Equation (4.1)):

$$IR = \frac{N}{T_{total}} \tag{4.1}$$

The individual durations between pairs are either manually chosen or generated. The program then combines all the glottis and vocal tract frames[6] into a single text file, called a *tractSequence* file. This serves as the *score* for the VocalTractLab synthesis engine. The interfacing and synthesis are done by changing variable values and running the program within the Python environment (i.e., in the *NKOAPP.py* file).

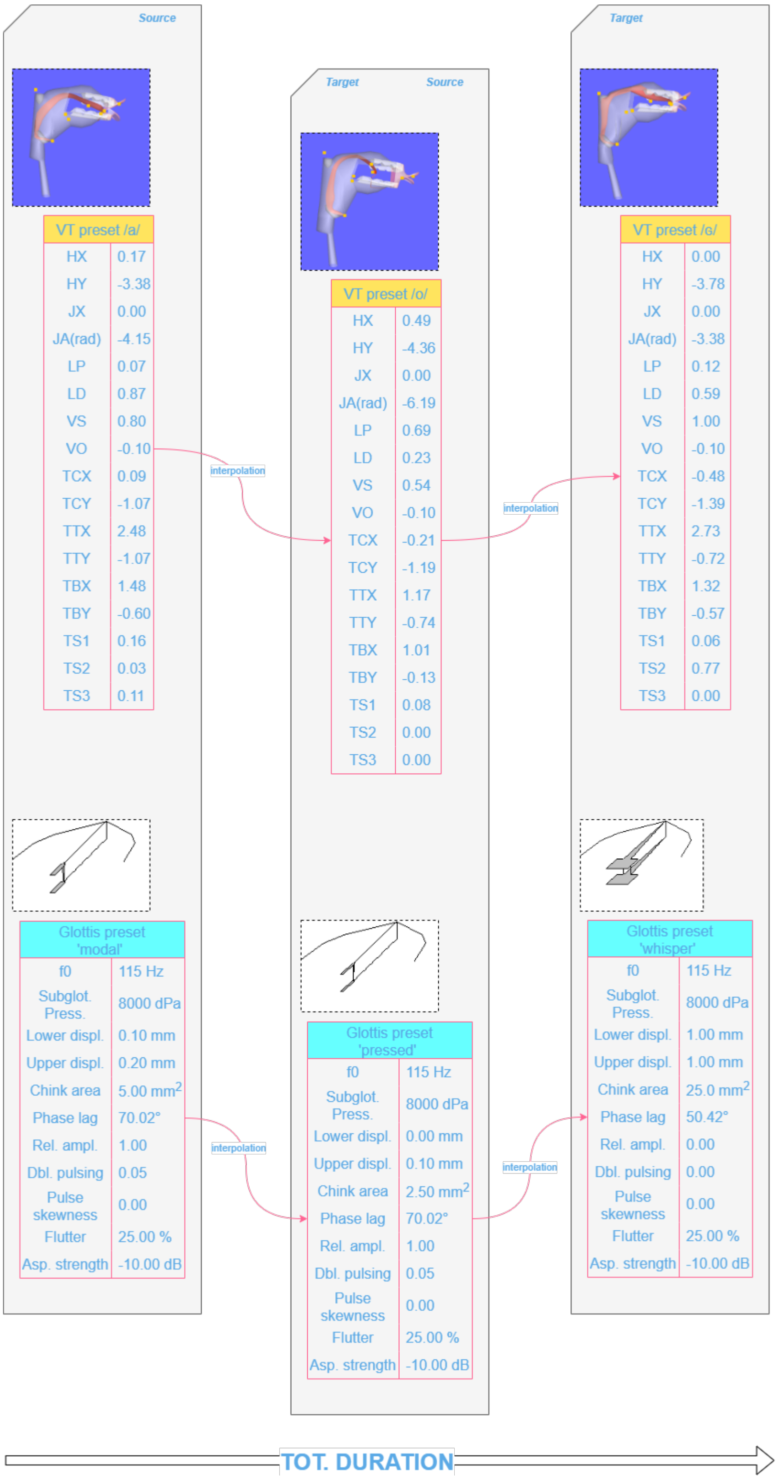
4.1.2 Resynthesis module

During the composition of the duets in *Enokian Soupe I* and *Enokian Soupe II*, there was a desire to incorporate certain acoustic features of the natural voice into the synthesis procedure of the synthetic voice. I decided to resynthesize the fundamental frequency and intensity contours to capture the prosodic, intonation, jitter, and shimmer patterns in the human performer's voice. The speaker's identity is incidentally represented in a hybrid and abstract form by the synthetic voice. The natural and synthetic voice are combined and layered sonically, and when the degree of likeness between voices is small enough, timbral fusion occurs under the right circumstances, as described in Section 3.2.1. The analysis of the fundamental frequency and intensity is performed by a script in the Praat software (Listing 4.1). This analysis file is then read by a text parser in NKOAPP and placed in the *tractSequence* file.

Using these pitch and intensity values, the control system can generate virtually infinite articulations. For instance, in excerpt 27—a vocal phrase from the fourth part of *Enokian Soupe III*—the natural voice vocalizes the syllables *tor*-*lel*-*anr*. In this passage, the pitches form a legato melody, with *tor* corresponding to the pitch C, *lel* to D, and *anr* to B.[7] The resynthesized versions can be heard in excerpts 28, 29, and 30. The resynthesized phonetic variations are audibly different, yet they express a form of contiguity with the natural voice. The synthetic voice is the minimal voice and the natural

[6] Each synthesis block is 110 frames long at a sample rate of 44100Hz, or 110/44100 = 2.5 milliseconds in duration.

[7] In 12 Tone equal temperament (12-TET).



VT preset /a/	
HX	0.17
HY	-3.38
JX	0.00
JA(rad)	-4.15
LP	0.07
LD	0.87
VS	0.80
VO	-0.10
TCX	0.09
TCY	-1.07
TTX	2.48
TTY	-1.07
TBX	1.48
TBY	-0.60
TS1	0.16
TS2	0.03
TS3	0.11

VT preset /o/	
HX	0.49
HY	-4.36
JX	0.00
JA(rad)	-6.19
LP	0.69
LD	0.23
VS	0.54
VO	-0.10
TCX	-0.21
TCY	-1.19
TTX	1.17
TTY	-0.74
TBX	1.01
TBY	-0.13
TS1	0.08
TS2	0.00
TS3	0.00

VT preset /ɢ/	
HX	0.00
HY	-3.78
JX	0.00
JA(rad)	-3.38
LP	0.12
LD	0.59
VS	1.00
VO	-0.10
TCX	-0.48
TCY	-1.39
TTX	2.73
TTY	-0.72
TBX	1.32
TBY	-0.57
TS1	0.06
TS2	0.77
TS3	0.00

Glottis preset 'modal'	
f0	115 Hz
Subglot. Press.	8000 dPa
Lower displ.	0.10 mm
Upper displ.	0.20 mm
Chink area	5.00 mm^2
Phase lag	70.02°
Rel. ampl.	1.00
Dbl. pulsing	0.05
Pulse skewness	0.00
Flutter	25.00 %
Asp. strength	-10.00 dB

Glottis preset 'pressed'	
f0	115 Hz
Subglot. Press.	8000 dPa
Lower displ.	0.00 mm
Upper displ.	0.10 mm
Chink area	2.50 mm^2
Phase lag	70.02°
Rel. ampl.	1.00
Dbl. pulsing	0.05
Pulse skewness	0.00
Flutter	25.00 %
Asp. strength	-10.00 dB

Glottis preset 'whisper'	
f0	115 Hz
Subglot. Press.	8000 dPa
Lower displ.	1.00 mm
Upper displ.	1.00 mm
Chink area	25.0 mm^2
Phase lag	50.42°
Rel. ampl.	0.00
Dbl. pulsing	0.00
Pulse skewness	0.00
Flutter	25.00 %
Asp. strength	-10.00 dB

Figure 4.2: Interpolations between glottis (Geometric Model) [*modal*:*pressed*:*whisper*] and vocal tract presets [[a]:[o]:[ɢ]] in NKOAPP

voice is the proto-voice in this context. A partial fusion of two vocal identities or personae occurs and results in an ambiguous perceptual quality. This quality is even more pronounced when heard simultaneously in a binaural or monaural setting (excerpt 31). In my composition *Laryngo-tauto* (2023), the NKOAPP system was primarily used to generate synthetic fricatives. The aim was to pronounce this contiguity between the familiar and unfamiliar voice (excerpt 32).

Listing 4.1: f0 and intensity analysis script in Praat for resynthesis in NKOAPP

```
sound = selected ("Sound")
    tmin = Get start time
    tmax = Get end time
    To Pitch: 0.0025, 75, 600
    Rename: "pitch"
    selectObject: sound
    To Intensity: 75, 0.0025
    Rename: "intensity"
    for i to (tmax-tmin)/0.0025
        time = tmin + i * 0.0025
        selectObject: "Pitch pitch"
        pitch = Get value at time: time, "Hertz", "linear"
        selectObject: "Intensity intensity"
        intensity = Get value at time: time, "cubic"
        appendInfoLine: fixed$ (pitch, 3), " ", fixed$ (intensity, 3)
    endfor
```

4.2 Aesthetical considerations

The idea of abstracting voice into a representation governed and described by mathematical equations and physical laws stands in contrast to how the Spectralists approach a *found vocal sound.* Initially, both modelling paradigms start with an observation in their perceivable environment. Then, in the physical modelling paradigm, the following step is to describe the components essential in producing sound pressure fluctuations specific to this found vocal sound. It is generally concerned with describing the process behind the source, and not taking its productive origin for granted, whereas, in the Spectralist frame of thought, the history of its acoustic cause is ignored. Once these physical laws are condensed into a set of equations, the system becomes less generalized and more specific, but simultaneously allows for a small to large complexity. Physical modelling also allows for a more compact description. In contrast with physical models, the spectral or acoustic models often used by the Spectralists involved a large operation-count/sample and its memory requirements were substantial (Smith 2008, 12). The

NKOAPP program's control and synthesis are flexible enough to sonically branch out far from the familiar sound classes of speech and into the uncharted domains of non-human vocalizations. This refers to the model's capability to handle varying levels of complexity, from simple to intricate. The synthesis can navigate between familiar and unfamiliar vocal images, evoking a phantasmagorical vocal interplay in the mind of the listener.

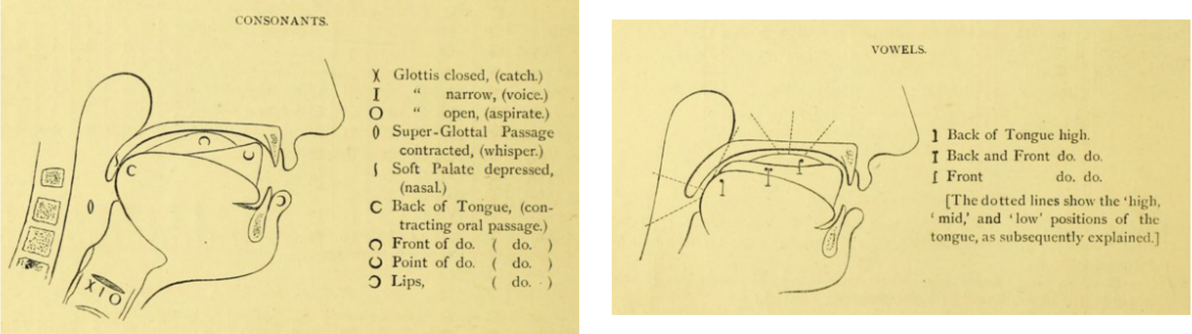
4.2.1 Imagining and Notating Mouth Sounds: "Visible Speech" and "Motor theory"

In developing and composing with the system, I became accustomed to identifying which articulators are involved in certain phonations and their corresponding physiological movements over time. However, I encountered a discrepancy between what I wanted the system to vocalize and how to document it in a compressed manner. This gap between my musical intention and the sound synthesis necessitated the use of a notation system that symbolically encoded mouth sounds in a kinaesthetically informed way. The eventual notation system that filled this gap is called "Visible Speech," and it was invented by British linguist Alexander Melville Bell (Bell 1867). It represents *visually* in what positions the speech organs (e.g., tongue, lips, etc.) should be in to produce a specific sound. Notably, in the International Phonetic Alphabet (IPA), the symbols only have vague visual cues for the shape of the articulators. For instance, the symbol for the diacritic <∘> in the partly voiceless [n] is [n̥̥]. It is not clear whether this visually represents partly open vocal folds and surely does not show in what positions the articulators have to be in order for the production of [n] in this type of phonation. In order to produce a vowel or consonant in the Visible Speech system, the symbols in Figure 4.3 are used (Ibid., 38). The direction and shape of the tongue arch are encoded visually in the mouth symbol.

When imagining a vocal sound that has to be synthesized, I usually vocalize the sounds internally. This mental simulation is further known as *inner speech.* Inner speech, or verbal thought, involves the mental rehearsal or articulation of words and sentences without vocalizing them aloud. This is supported by the hypothesis from "motor theory of speech perception"(Studdert-Kennedy et al. 1970), which posits that speech perception involves the activation of the vocal motor system[8], where the listener mentally simulates the movements of their own articulators in understanding speech sounds. Inner speech is also seen as a form of speech perception, and activates the same motor responses similar to speech production (Oppenheim 2013). In the context of motor theory of speech perception, the Visible Speech system could be seen as a way to visually represent the movements of speech articulators. When the performer sees these mouth symbols or diagrams representing speech sounds,

[8] The McGurk Effect is an evidence for supporting the motor theory. It demonstrates the integration of visual and auditory information in speech perception (Green 1996).

they may mentally simulate the corresponding articulatory movements as if they were producing those sounds themselves. This mental technique is an efficient way for both the synthesist, and the performer to write and read these symbols. This technique is used to imagine and notate phonations for the three iterations of the *Enokian Soupe* series.[9]


CONSONANTS.
Glottis closed, (catch.)
" narrow, (voice.)
" open, (aspirate.)
Super-Glottal Passage contracted, (whisper.)
Soft Palate depressed, (nasal.)
Back of Tongue, (contracting oral passage.)
Front of do. (do.)
Point of do. (do.)
Lips, (do.)
VOWELS.
Back of Tongue high.
Back and Front do. do.
Front do. do.
[The dotted lines show the 'high, 'mid,' and 'low' positions of the tongue, as subsequently explained.]

(a) Symbols for consonants

(b) Symbols for vowels

Figure 4.3: Diagrams showing the relation of the *primary organic symbols* to the organs for (a) consonants and (b) vowels.

[9] Only in the score of *Enokian Soupe I* (Appendix B) these symbols are notated. In *Enokian Soupe II* and *III* Visible Speech notation is notated elsewhere.

Conclusion and Future Work

Over the last four years, I have consistently worked on and listened to music that involved artificial voice. Throughout this formative period, recurring and poignant questions emerged from the contexts of composition, programming, and auditory perception and sensation: What elements in the experience and sensation of voice consistently draw me back to it? What attentions are at play during the audition of something vocal, and how can these attentions be composed? Synthesizing voice addressed many of the notions I was curious about, particularly in linking my musical intentions with the analytical and synthetic approaches to investigating vocal phenomena.

Through hearing and sounding the seemingly vocal—the synthetic—I gained a deeper understanding on what voice is and isn't, and how it operates perceptually and behaves musically different from other sounds, especially when heard together with human voice. In my music, I aimed to achieve what I call "vocal ambiguities" to create meanings distinct from those associated with language, allowing an equally potent yet abstract meaning to emerge solely from the sonority of voice. To achieve these vocal ambiguities, I reinterpreted ideas of fusion from Spectralist, frequency-based, and other composers into compositions that used physical models instead of spectral models. I explored experientially and cognitively informed frameworks for describing and analyzing voice, such as the minimal and maximal voice continuum and the degree of likeness between vocal sounds. Then, my intentions and design goals are defined for the vocal synthesis program NKOAPP in order to operate within these frameworks.

This project examines vocal ambiguities across micro and meso timescales of voice. In the micro time regime, both synthetic and natural voice are analyzed through features like timbre, pitch, formants, and phoneme articulation. The meso timescale includes aspects such as prosody, rhythm, intonation, and phrasing. Further investigation is needed to describe vocal ambiguities on a macro-scale, possibly enabling composers to experiment with lexical ambiguities, narrative structure, thematic development, and speech patterns over extended discourse. Looking into the future, combining vocal ambiguities across these timescales is certainly worthwhile to explore further. These ideas and prospects can be integrated into the design and development choices of the NKOAPP application, possibly guiding its future developments and additions.

Bibliography

"3.2. Acoustic Aspects of Consonants – Phonetics and Phonology". Visited on 05/04/2024. https://corpus.eduhk.hk/english_pronunciation/index.php/3-2-acoustic-aspects-of-consonants/.

Akima, Hiroshi. "A New Method of Interpolation and Smooth Curve Fitting Based on Local Procedures". *Journal of the ACM* 17, no. 4 (Oct. 1970): 589–602. ISSN: 0004-5411, 1557-735X, visited on 05/13/2024. https://doi.org/10.1145/321607.321609. https://dl.acm.org/doi/10.1145/321607.321609.

Bell, Alexander Melville. *Visible speech : the science of universal alphabetics, or self-interpreting physiological letters, for the writing of all languages in one alphabet.* London: Simpkin, 1867. Visited on 02/06/2023. https://wellcomecollection.org/works/a4u76ncp.

Bergsland, Andreas. "Experiencing Voices in Electroacoustic Music". Doctoral Dissertation, Norwegian University of Science and Technology, May 10, 2010.

Birkholz, Peter. "Modeling Consonant-Vowel Coarticulation for Articulatory Speech Synthesis". Ed. by Francesco Pappalardo. *PLoS ONE* 8, no. 4 (Apr. 16, 2013): e60603. ISSN: 1932-6203, visited on 06/11/2022. https://doi.org/10.1371/journal.pone.0060603.

Boersma, Paul, and David Weenink. "PRAAT, a system for doing phonetics by computer". *Glot international* 5 (Jan. 1, 2001): 341–345.

Bossis, Bruno. "Reflections on the analysis of artificial vocality: representations, tools and prospective". *Organised Sound* 9, no. 1 (Apr. 2004): 91–98. ISSN: 1355-7718, 1469-8153, visited on 03/17/2024. https://doi.org/10.1017/S1355771804000123.

– . "The Analysis of Electroacoustic Music: From sources to invariants". *Organised Sound* 11 (Aug. 1, 2006): 101–112. https://doi.org/10.1017/S135577180600135X.

Bregman, Albert. "Auditory Scene Analysis: The Perceptual Organization of Sound". *Journal of The Acoustical Society of America* 95 (Jan. 1, 1990). https://doi.org/DOI:10.1121/1.408434.

Chafe, Chris. "Case Studies of Physical Models in Music Composition" (Jan. 1, 2004).

Chion, Michel. “Guide to Sound Objects” (Jan. 1, 1983).

Cooper, Sam. “Introduction to Isidore Isou”. Sam Cooper | Research Blog | Avant-Garde Literature and Visual Culture, May 26, 2019. Visited on 05/15/2024. `https://situationistresearch.wordpress.com/2019/05/26/introduction-to-isidore-isou/`.

Cox, Christoph, and Daniel Warner, eds. *All Sound Is Queer - Daniel Drew*. 1st ed. Bloomsbury Publishing Inc, 2017. ISBN: 978-1-5013-1839-9 978-1-5013-1836-8, visited on 12/06/2023. `https://doi.org/10.5040/9781501318399`. `https://www.bloomsburymusicandsound.com/encyclopedia?docid=b-9781501318399`.

Cutting, James E. “Auditory and Linguistic Processes in Speech Perception: Inferences from Six Fusions in Dichotic Listening” (1976): 114–140.

Dee, John. “48 Claves Angelicae”. Sloane Manuscript 3191, London. Visited on 05/12/2024. `https://booksofmagick.com/wp-content/uploads/2020/10/1.-48-Claves-Angelicae-Sloane-3191.pdf`.

Deleuze, Gilles. *The Logic of Sense*. Ed. by Constantin V. Boundas. Columbia University Press, 1990.

Depalle, Philippe, Guillermo Garcia, and Xavier Rodet. “A Virtual Castrato ()”. In *ICMC: International Computer Music Conference*, 357–360. Aarhus, Denmark, 1994. Visited on 03/18/2024. `https://hal.science/hal-01105432`.

Deutsch, Diana. “An auditory illusion”. *Nature* 251 (Oct. 1, 1974): 307–9. `https://doi.org/10.1121/1.1919587`.

Dorman, Michael F., et al. “Identification of Synthetic Vowels by Patients Using the Symbion Multichannel Cochlear Implant:” *Ear and Hearing* 10, no. 1 (Feb. 1989): 40–43. ISSN: 0196-0202, visited on 05/01/2024. `https://doi.org/10.1097/00003446-198902000-00007`. `http://journals.lww.com/00003446-198902000-00007`.

Fitch, W., and Jay Giedd. “Morphology and development of the human vocal tract: A study using magnetic resonance imaging”. *The Journal of the Acoustical Society of America* 106 (Oct. 1, 1999): 1511–22. `https://doi.org/10.1121/1.427148`.

Garant, Dominic. *Tristan Murail, Les objets sonores complexes. Analyse de “LÉsprit des dunes”*. Paris: L’Harmattan, 2011.

Green, K.P. “Studies of the McGurk effect: implications for theories of speech perception”. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP ’96*, 3:1652–1655. Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP ’96. Philadelphia, PA, USA: IEEE, 1996. ISBN: 978-0-7803-3555-4, visited on 05/14/2024. `https://doi.org/10.1109/ICSLP.1996.607942`. `http://ieeexplore.ieee.org/document/607942/`.

Griffin, David. “The 49 Enochian Calls”. In *A Completer Course in Practical Magic*, 32. 2008.

Harvey, Jonathan. “Speakings | Faber Music”, 2008. Visited on 01/08/2024. https://www.fabermusic.com/music/speakings-5282.

Heald, Shannon, Serena Klos, and Howard Nusbaum. “Understanding Speech in the Context of Variability”. In *Neurobiology of Language*, 195–208. Elsevier, 2016. ISBN: 978-0-12-407794-2, visited on 04/30/2024. https://doi.org/10.1016/B978-0-12-407794-2.00017-1.

Hirs, Rozalie. “On Tristan Murail’s Le lac: Contemporary compositional techniques and OpenMusic (Dissertation, Columbia University, 2007)”. *On Tristan Murail’s Le lac: Contemporary compositional techniques and OpenMusic (Dissertation Doctor of Musical Arts, Columbia University)* (Jan. 1, 2007). Visited on 03/19/2024. https://www.academia.edu/103175766/On_Tristan_Murail_s_Le_lac_Contemporary_compositional_techniques_and_OpenMusic_Dissertation_Columbia_University_2007_.

Hirs, Rozalie, and Bob Gilmore. *Contemporary compositional techniques and OpenMusic*. Musique-sciences. [Paris] Sampzon: IRCAM-Centre Pompidou Delatour France, 2009. ISBN: 978-2-7521-0080-1.

Hunt, Jerry. “Gesture Modulation of Templates”, 2001. Visited on 04/21/2024. http://www.jerryhunt.org/gesture_mod.htm.

Jonathan Harvey - BBC Scottish Symphony Orchestra, Ilan Volkov - Speakings. 2010. Visited on 05/14/2024. https://www.discogs.com/release/2476683-Jonathan-Harvey-BBC-Scottish-Symphony-Orchestra-Ilan-Volkov-Speakings.

Lingala, Sajan Goud, et al. “Recommendations for Real-Time Speech MRI”. Publisher: NIH Public Access, *Journal of magnetic resonance imaging : JMRI* 43, no. 1 (Jan. 2016): 28. Visited on 05/14/2024. https://doi.org/10.1002/jmri.24997. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5079859/.

McAdams, Stephen. “L’organisation perceptive de l’environnement sonore”. In *Rencontres IPSEN en ORL,* ed. by Editions Irvinn, 91–102. 1997. Visited on 05/12/2024. https://hal.science/hal-01105540.

McNabb, Michael. “’Dreamsong’: The Composition”. *Computer Music Journal* 5, no. 4 (1981): 36. ISSN: 01489267, visited on 03/19/2024. https://doi.org/10.2307/3679505.

Minarelli, Enzo. “Every Poet Needs A Theory”. Enzo Minarelli, July 2, 2019. Visited on 04/27/2024. https://www.enzominarelli.com/every-poet-needs-a-theory-poetry-polypoetry-sound-poetry/.

Morrison, Landon. “Encoding Post-Spectral Sound: Kaija Saariaho’s Early Electronic Music at IRCAM, 1982–87”. *Music Theory Online* 27, no. 3 (Sept. 1, 2021). Visited on 05/11/2024. https://mtosmt.org/issues/mto.21.27.3/mto.21.27.3.morrison.html.

Oppenheim, Gary M. "Inner speech as a forward model?" *Behavioral and Brain Sciences* 36, no. 4 (Aug. 2013): 369–370. ISSN: 0140-525X, 1469-1825, visited on 05/14/2024. https://doi.org/10.1017/S0140525X12002798. https://www.cambridge.org/core/product/identifier/S0140525X12002798/type/journal_article.

Phil Legard. "Musical Interlude: David Dunn". Larkfall, July 24, 2013. Visited on 04/21/2024. https://larkfall.wordpress.com//?s=jerry+hunt&search=Go.

René, van Peer. "Common Ground: Jerry Hunt and Paul Panhuysen in conversation", 1993. Visited on 04/22/2024. http://www.jerryhunt.org/van_peer.htm#uncanny.

Repp, Bruno H. "The Auditory Processing of Speech: From Sound to Words". *Language and Speech* 37, no. 3 (July 1994): 337–340. ISSN: 0023-8309, 1756-6053, visited on 04/10/2024. https://doi.org/10.1177/002383099403700311. http://journals.sagepub.com/doi/10.1177/002383099403700311.

Russo, Nicole, et al. "Brainstem responses to speech syllables". *Clinical Neurophysiology* 115, no. 9 (Sept. 2004): 2021–2030. ISSN: 13882457, visited on 04/15/2024. https://doi.org/10.1016/j.clinph.2004.04.003.

Sakayori, Shuichi, et al. "Critical spectral regions for vowel identification". *Neuroscience Research* 43, no. 2 (June 2002): 155–162. ISSN: 01680102, visited on 05/01/2024. https://doi.org/10.1016/S0168-0102(02)00026-3.

Small, Christopher. *Musicking: The Meanings of Performing and Listening.* Wesleyan University Press, 1998.

Smith, Julius. "A Basic Introduction to Digital Waveguide Synthesis (for the Technically Inclined)" (Jan. 2006).

– . "Viewpoints on the History of Digital Synthesis", Perspectives of New Music (Dec. 8, 2008). Visited on 06/07/2023. https://doi.org/10.2307/833307. https://ccrma.stanford.edu/~jos/kna/kna.pdf.

Smorenburg, Laura, and Aoju Chen. "The effect of female voice on verbal processing". *Speech Communication* 122 (Sept. 1, 2020): 11–18. ISSN: 0167-6393, visited on 05/01/2024. https://doi.org/10.1016/j.specom.2020.04.004. https://www.sciencedirect.com/science/article/pii/S0167639319302560.

Story, Brad. "Speech synthesis by mapping articulator movement patterns to a shape-based area function model of the vocal tract". *The Journal of the Acoustical Society of America* 109 (May 1, 2001): 2444–2445. https://doi.org/10.1121/1.4744658.

Studdert-Kennedy, Michael, et al. "Motor theory of speech perception: A reply to Lane's critical review." *Psychological Review* 77, no. 3 (1970): 234–249. ISSN: 1939-1471, 0033-295X, visited on

05/14/2024. https://doi.org/10.1037/h0029078. https://doi.apa.org/doi/10.1037/h0029078.

Taverna, Andrea S. "Motherese in the Wichi Language (El maternés en la lengua wichí)". Publisher: SAGE Publications, *Journal for the Study of Education and Development* 44, no. 2 (May 1, 2021): 303–335. ISSN: 0210-3702, visited on 04/15/2024. https://doi.org/10.1080/02103702.2021.1889290.

Teixeira, João Paulo, Carla Oliveira, and Carla Lopes. "Vocal Acoustic Analysis – Jitter, Shimmer and HNR Parameters". *Procedia Technology* 9 (2013): 1112–1122. ISSN: 22120173, visited on 05/08/2024. https://doi.org/10.1016/j.protcy.2013.12.124.

Teodorescu-Ciocanea, Livia. "Timbre versus spectralism". *Contemporary Music Review* 22, no. 1 (Mar. 2003): 87–104. ISSN: 0749-4467, 1477-2256, visited on 05/12/2024. https://doi.org/10.1080/0749446032000134751. http://www.tandfonline.com/doi/abs/10.1080/0749446032000134751.

Tokuda, Isao, and Hanspeter Herzel. "Biomechanical simulation of chest-falsetto transitions and the influence of vocal tract resonators" (Jan. 1, 2009).

"Understanding Voice Production - THE VOICE FOUNDATION", Aug. 31, 2013. Visited on 05/08/2024. https://voicefoundation.org/health-science/voice-disorders/anatomy-physiology-of-voice-production/understanding-voice-production/.

Wishart, Trevor, and Simon Emmerson. *On sonic art.* New and rev. ed. Contemporary music studies v. 12. OCLC: ocm35617339. Amsterdam: Harwood Academic Publishers, 1996. ISBN: 978-3-7186-5846-6 978-3-7186-5847-3 978-3-7186-5848-0.

Appendices

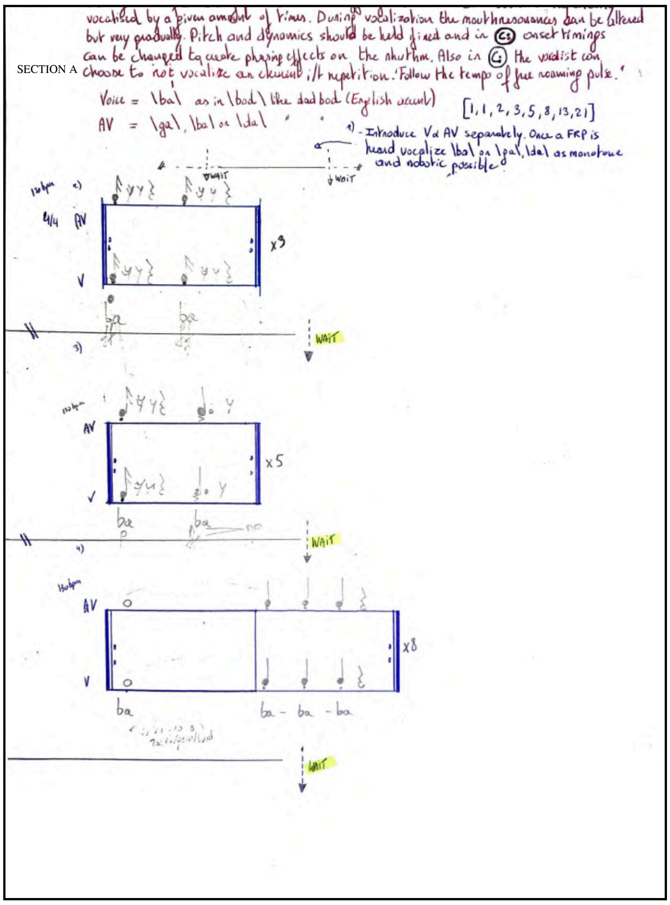
Appendix A:

Scores for ***Enokian Soupe I, II & III***

The score for *Enokian Soupe I* is partly contained within the *Enokian Soupe II* score as the interludes 1, 2 and 3. These are inserted in between sections A, B, C and D of *Enokian Soupe II.* Only part 4 is not used in the second iteration of the *Enokian Soupe* series. The score for *Enokian Soupe II* is shown below.

Listen to *Enokian Soupe I* and *Enokian Soupe II* in the sound excerpt listing numbered 33 and 34.

SECTION A



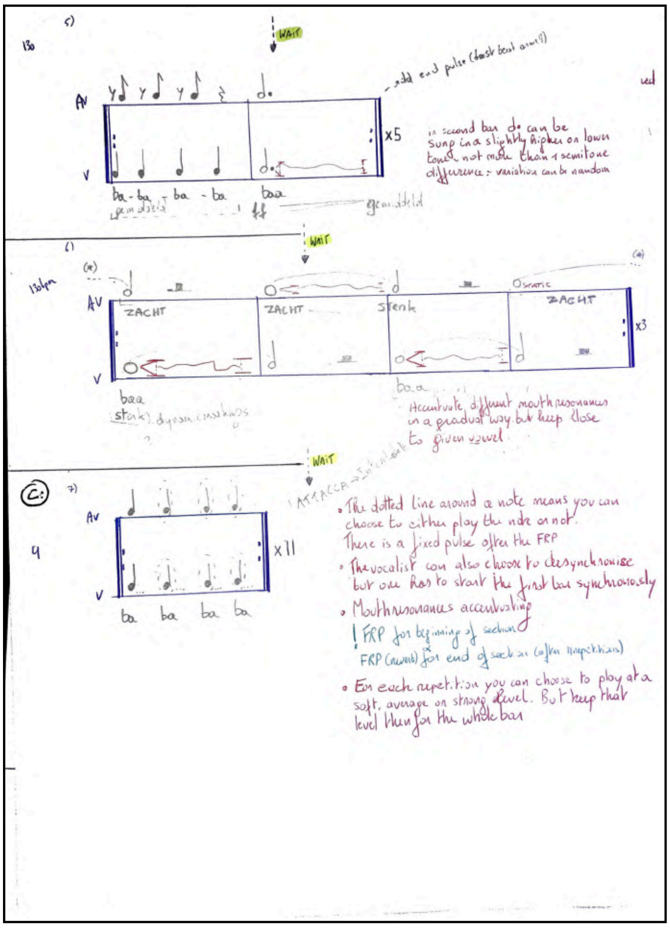
vocalised by a given amount of times. During vocalization the mouthresonances can be altered but very gradually. Pitch and dynamics should be held fixed and in C: onset timings can be changed to create phasing effects on the rhythm. Also in C: the vocalist can choose to not vocalize an element i/t repetition. 'Follow the tempo of free roaming pulse.'

Voice = \ba\ as in \bod\ like dad bod (English accent)

AV = \ga\, \ba\ or \da\

[1, 1, 2, 3, 5, 8, 13, 21]

1) - Introduce V & AV separately. Once a FRP is heard vocalize \ba\ or \ga\, \da\ as monotone and robotic possible.



- add end pulse (last beat accent)

in second bar 𝅗𝅥. can be sung in a slightly higher or lower tone, not more than 1 semitone difference - variation can be random

Accentuate different mouthresonances in a gradual way but keep close to given vowel

- The dotted line around a note means you can choose to either play the note or not. There is a fixed pulse after the FRP
- The vocalist can also choose to desynchronise but one has to start the first bar synchronously
- Mouthresonances accentuating

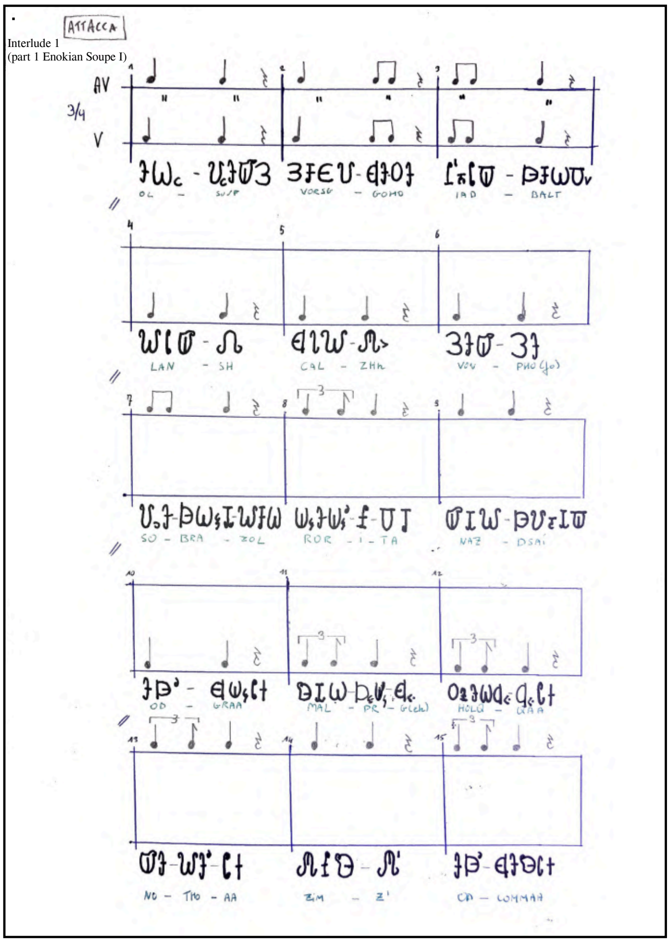
! FRP for beginning of section

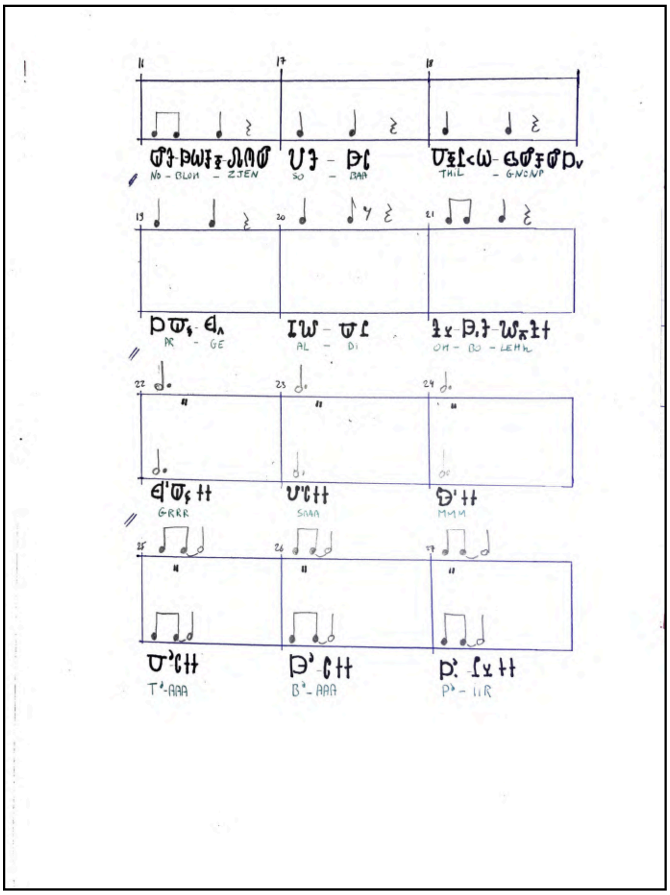
FRP (reverb) for end of section (after repetitions)

- For each repetition you can choose to play at a soft, average or strong level. But keep that level then for the whole bar

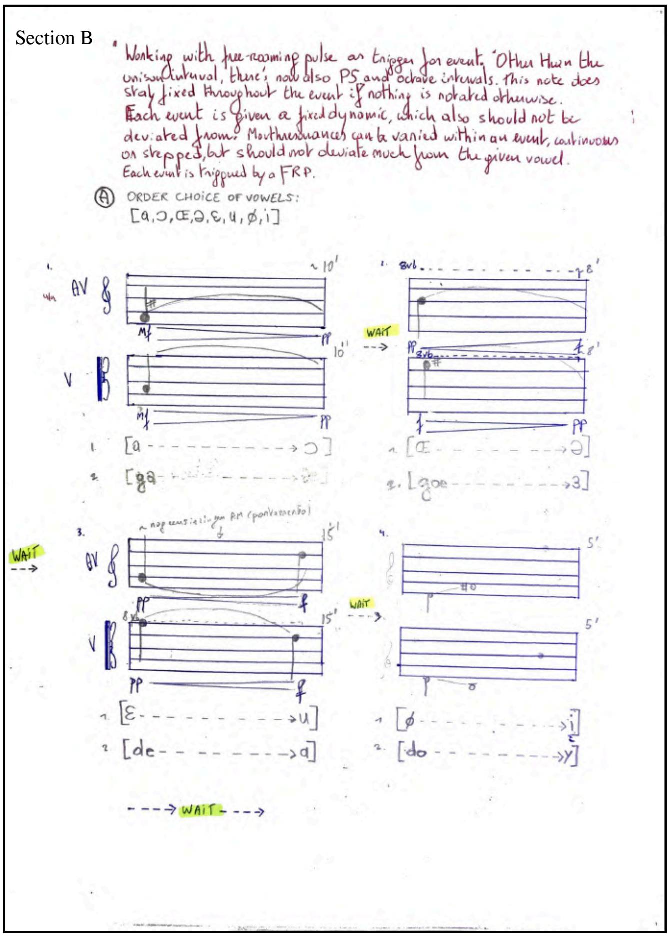
ATTACCA

Interlude 1
(part 1 Enokian Soupe I)

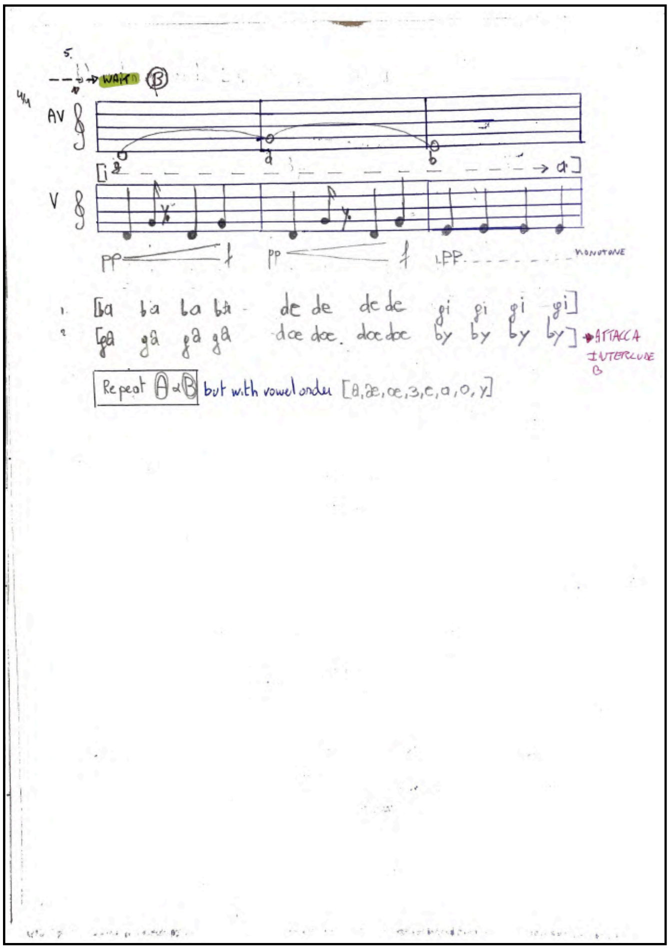




Section B


Working with free-rooming pulse as trigger for event. Other than the unison interval, there's now also P5 and octave intervals. This note does stay fixed throughout the event if nothing is notated otherwise. Each event is given a fixed dynamic, which also should not be deviated from. Mouthresonances can be varied within an event, continuous or stepped, but should not deviate much from the given vowel. Each event is tripped by a FRP.

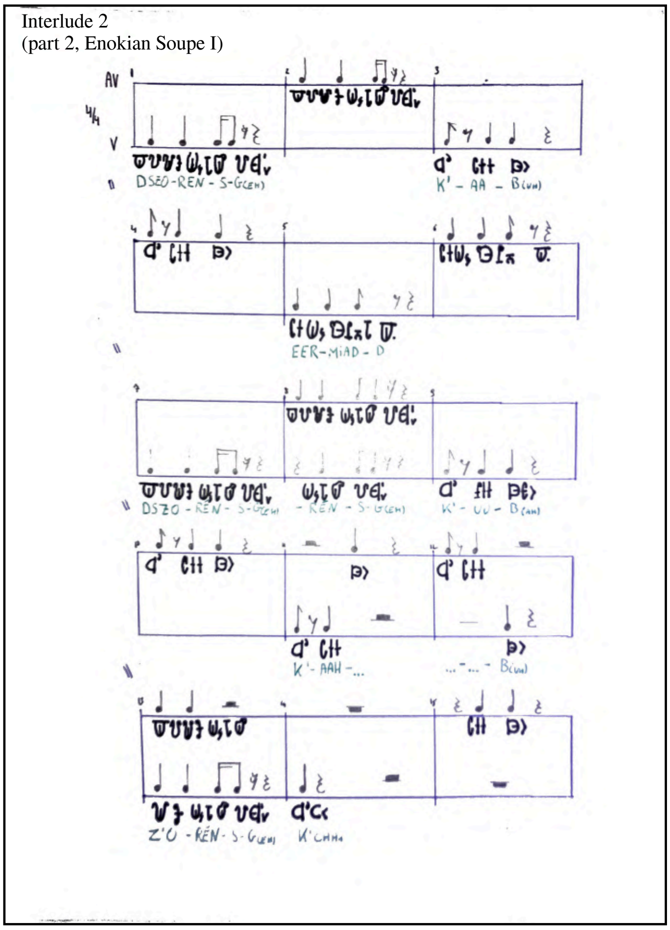
Ⓐ ORDER CHOICE OF VOWELS: [a,ɔ,œ,ə,ɛ,u,ø,i]

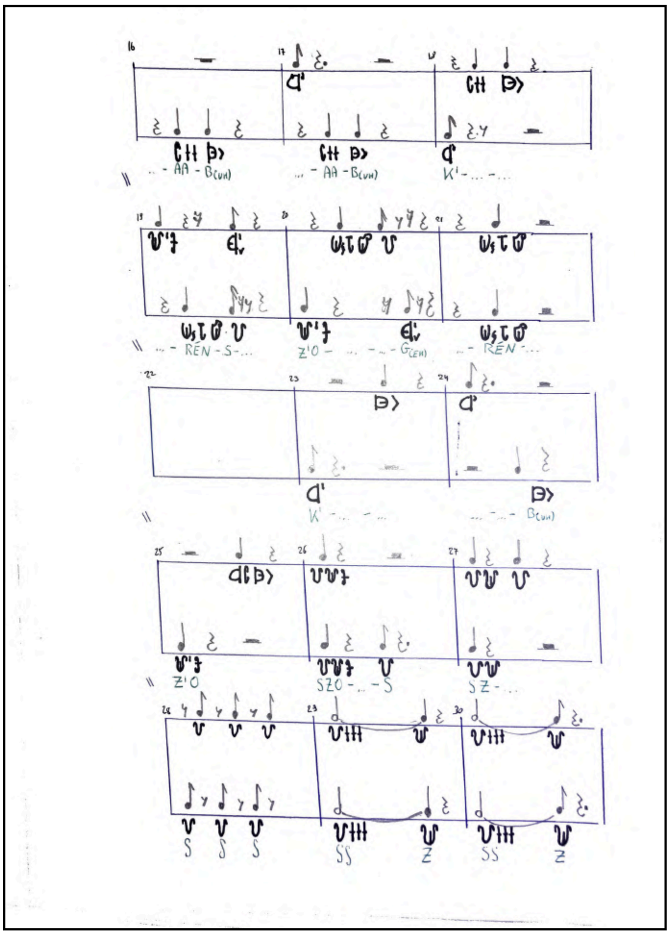


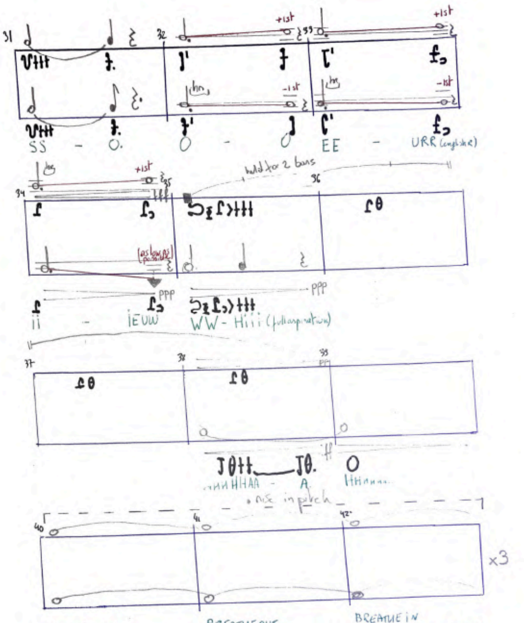
5.
WAIT Ⓑ
4/4
AV
V
[i → a]
pp
f
pp
f
PP
MONOTONE
1. [ba ba ba ba de de de de gi gi gi gi]
2. [ga ga ga ga doe doe doe doe by by by by]
ATTACCA INTERLUDE B

Repeat Ⓐ & Ⓑ but with vowel order [ɐ,æ,œ,ɜ,e,ɑ,o,y]

Interlude 2
(part 2, Enokian Soupe I)







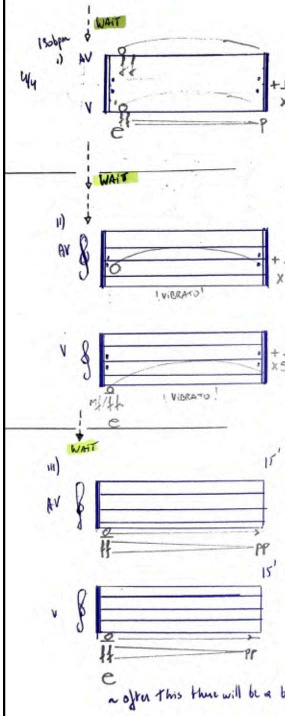
i) BREATHE IN WITH PROTRUDED LIPS OUTWARD — BREATHE OUT WITH PROTRUDED LIPS OUTWARD — BREATHE IN WITH PROTRUDED LIPS OUTWARD

ii) BREATHE OUT WITH LIPS IN NEUTRAL POSITION & OPEN MOUTH MORE — BREATHE IN " ... " — BREATHE OUT " ... "

iii) BREATHE IN HEAVILY LIKE YOUR OUT OF BREATH — BREATHE OUT "... " — BREATHE IN " ... "

Section C

After Interlude B, a focus will now rely on more noisier vocalizations. Producing vowels by gradually increasing and decreasing hoarse-like vocalizations. Still working with a free roaming pulse for triggering sections.



- First bar is breathing out but with the mouth configuration of \e\.
- Every repetition the vocal fold tension has to become bigger until the vowel \e\ is vocalized with a clean tone (speech like) and little to no breath should
- veel lucht → weinig lucht

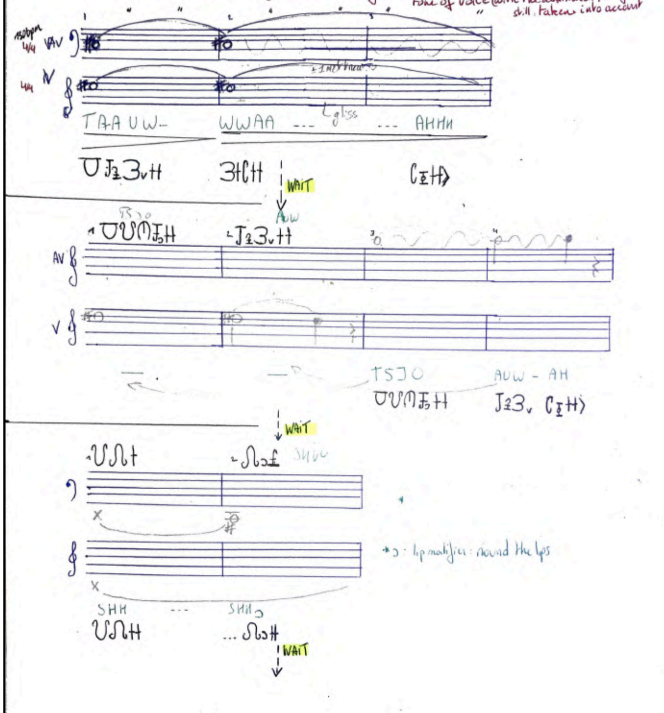
- Same as above, first bar is just breathing out in mouth configuration \e\
- " " until \e\ is vocalized with a clean singing tone.

- Maintain a growly, hoarselike texture of \e\
- Beginning has to be ~~an~~ an expulsion of energy and alot of tension has to be built up before vocalization and then loosen up this tension of the vocalfolds over 15'

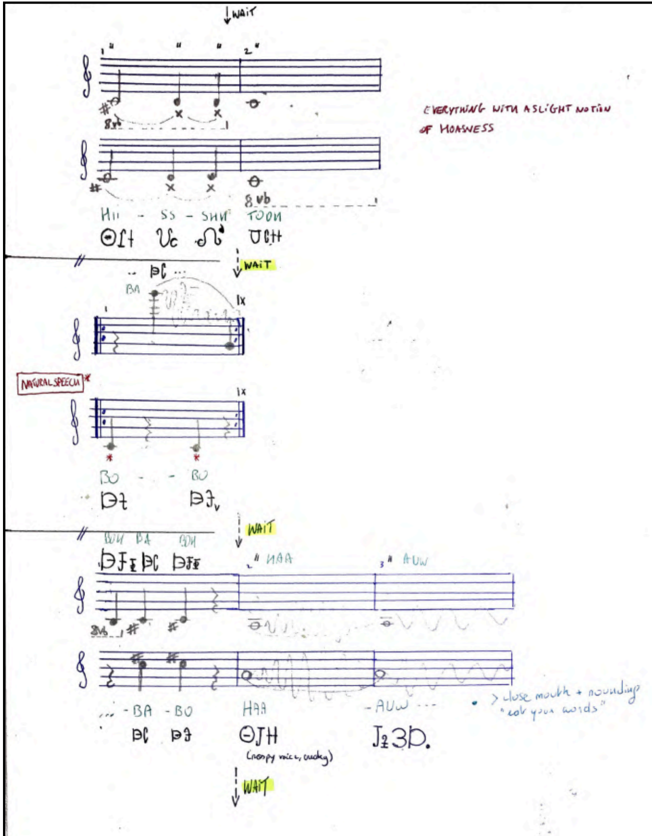
~ after this there will be a break of 10-15 sec (DRINK)

Interlude 3
(part 3 Enokian Soupe I)

Free roaming pulse in time. Before each part a whole bar will be heard before uttering the sentence/utterance/whatever. There will still be a steady tempo but when it's time to say or sing the part/bar, the part will have to be uttered in a steady way according to the notation. When you see an x under the staff, it means the utterance is unvoiced and when you see the marking NATURAL SPEECH, it should be uttered in normal tone of voice (with the additional modifiers still taken into account

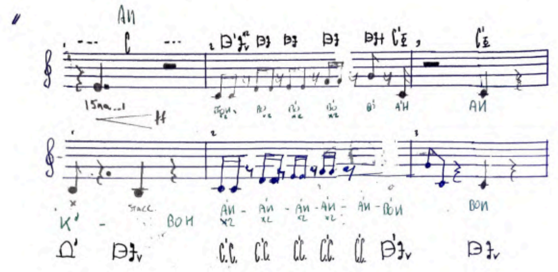


*ɔ : lip modifier: round the lips

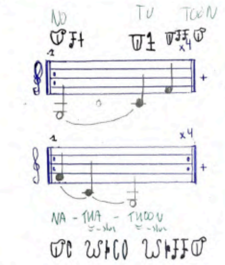


EVERYTHING WITH A SLIGHT NOTION OF HOARSNESS

• > close mouth + rounding "eat your words"

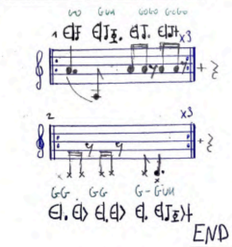


WAIT



- EACH REPETION, SLUR YOUR PRONOUNCIATION MORE. LET THE TONGUE REALLY SLAP AGAINST YOUR TEETH (use more saliva if needed)
- WITH EACH REPETITION, BOTH VOICES ALSO GET MORE BREATHY.

WAIT

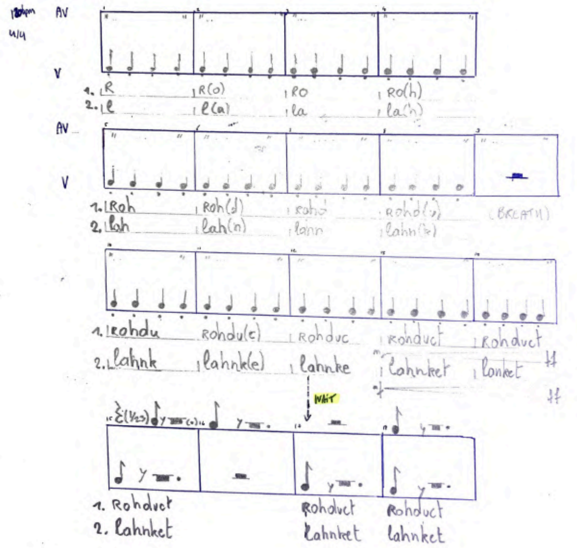


START WITH STACCATO PRONOUNCING OF THE 'G'S THEN OVER THE REPEATS, SLOW DOWN THE STACCATO' EFFECT AND LET THE TONGUE GET HEAVIER AND HEAVIER

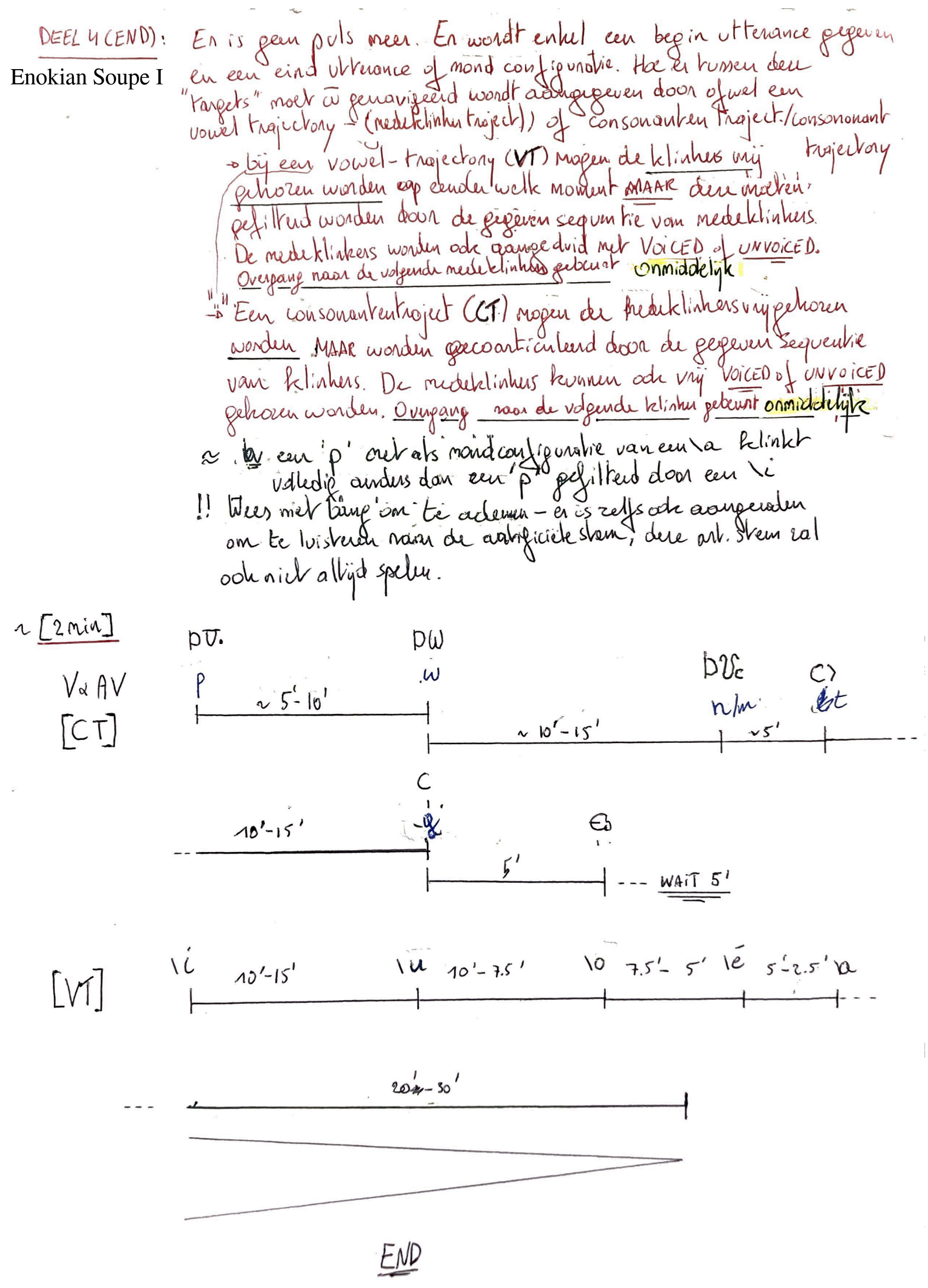
END PART 3

Section D

- After the last interlude, we'll slowly play with a phonological fusion. The natural voice will repeatedly say a part of the word \Rohduct. (pronounce in english way) And the artificial voice a part of the word \pohduct. Through simultaneous playback at a correct level, tone & synchronicity the listener could hear a fused word, "product".
- With more repetitions, larger chunks of the words are pronounced. Once the full word is vocalized we'll transition to \barhet for AV and \lanket for the natural voice, again eventually resulting in a percept of "planket".
- Be as monotone as possible, and speech like expression.
- FRP!



Now the last part (part 4) of *Enokian Soupe I* is shown below.



DEEL 4 (END): Er is geen puls meer. Er wordt enkel een begin utterance gegeven en een eind utterance of mond configuratie. Hoe er tussen deze "targets" moet ω genavigeerd wordt aangegeven door ofwel een vowel trajectory – (medeklinker traject)) of consonanten traject./consononant trajectory

Enokian Soupe I

→ bij een vowel-trajectory (VT) mogen de klinkers vrij gehoren worden op eender welk moment MAAR deze moeten gefiltred worden door de gegeven sequentie van medeklinkers. De medeklinkers worden ook aangeduid met VOICED of UNVOICED. Overgang naar de volgende medeklinker gebeurt onmiddelijk

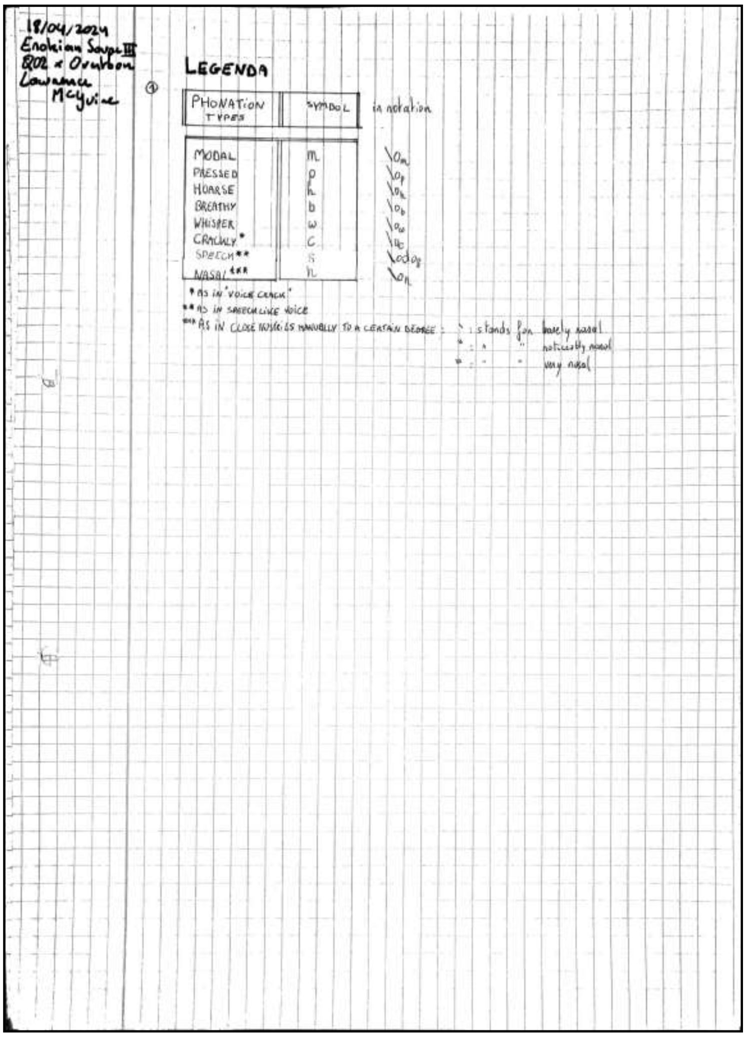
→ "Een consonantentraject (CT) mogen de medeklinkers vrij gehoren worden MAAR worden gecoarticuleerd door de gegeven sequentie van klinkers. De medeklinkers kunnen ook vrij VOICED of UNVOICED gehoren worden. Overgang naar de volgende klinker gebeurt onmiddelijk

≈ bv. een 'p' met als mondconfiguratie van een \a klinkt volledig anders dan een 'p' gefilterd door een \i

!! Wees niet bang om te ademen – er is zelfs ook aangeraden om te luisteren naar de artificiële stem, deze art. stem zal ook niet altijd spelen.

Shown below is the score for *Enokian Soupe III.* There are four parts and are structured around one quattrain of an ancient Irish Gaelic deibhide riddle poem. The pitch material is extracted from the Sean-nós singer Kitty Callaghers' *Keen for a Dead Child* (1998), and reintroduced by the synthetic voice with slightly offset pitches from the 12-TET system. All of the natural vocalizations are accompanied by a sine tone that follows the same pitch as the human vocalist.

Listen to a recording of *Enokian Soupe III* in the sound listings number 35.



18/04/2024
Enokian Soupe III
Q02 x Ovurtoon
Lawrence
McGuire

①

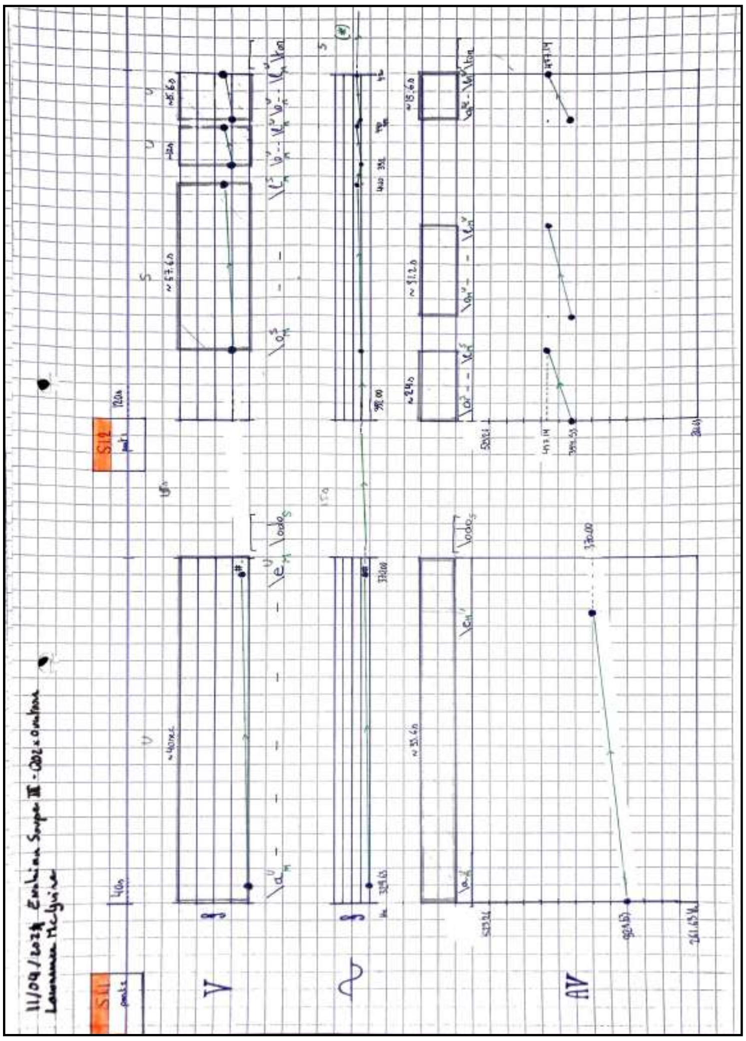
LEGENDA

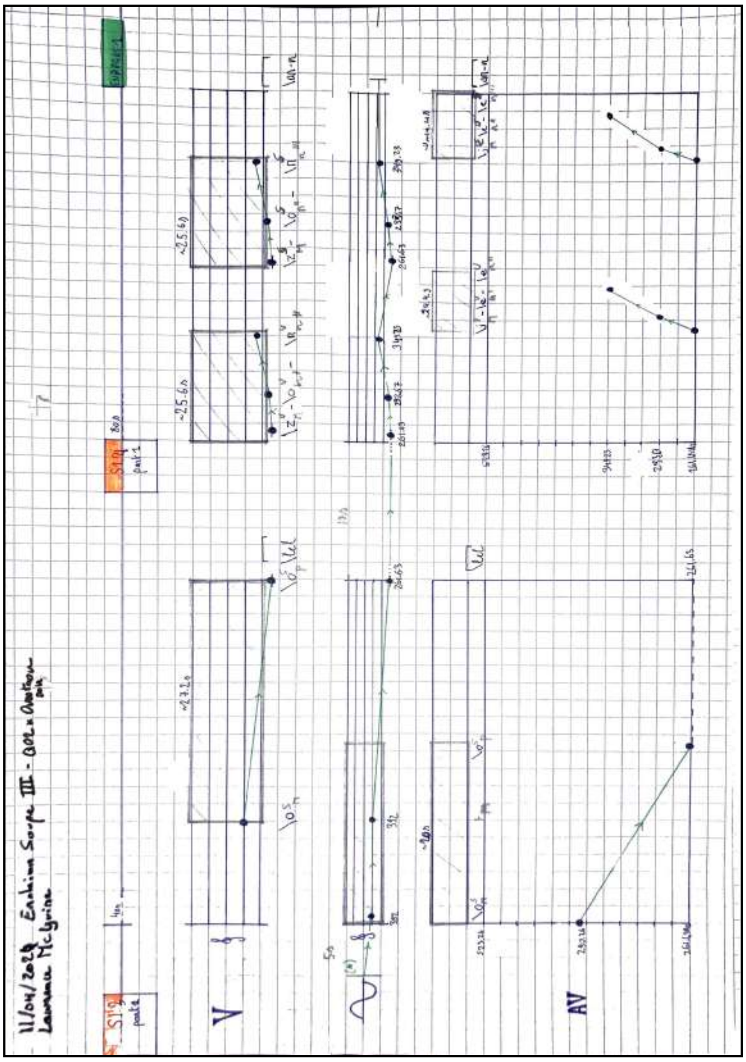
PHONATION TYPES	SYMBOL	in notation
MODAL	m	$\backslash o_{m}$
PRESSED	p	$\backslash o_{p}$
HOARSE	h	$\backslash o_{h}$
BREATHY	b	$\backslash o_{b}$
WHISPER	w	$\backslash o_{w}$
CRACKLY*	c	$\backslash o_{c}$
SPEECH**	s	$\backslash odo_{s}$
NASAL***	n	$\backslash o_{n}$

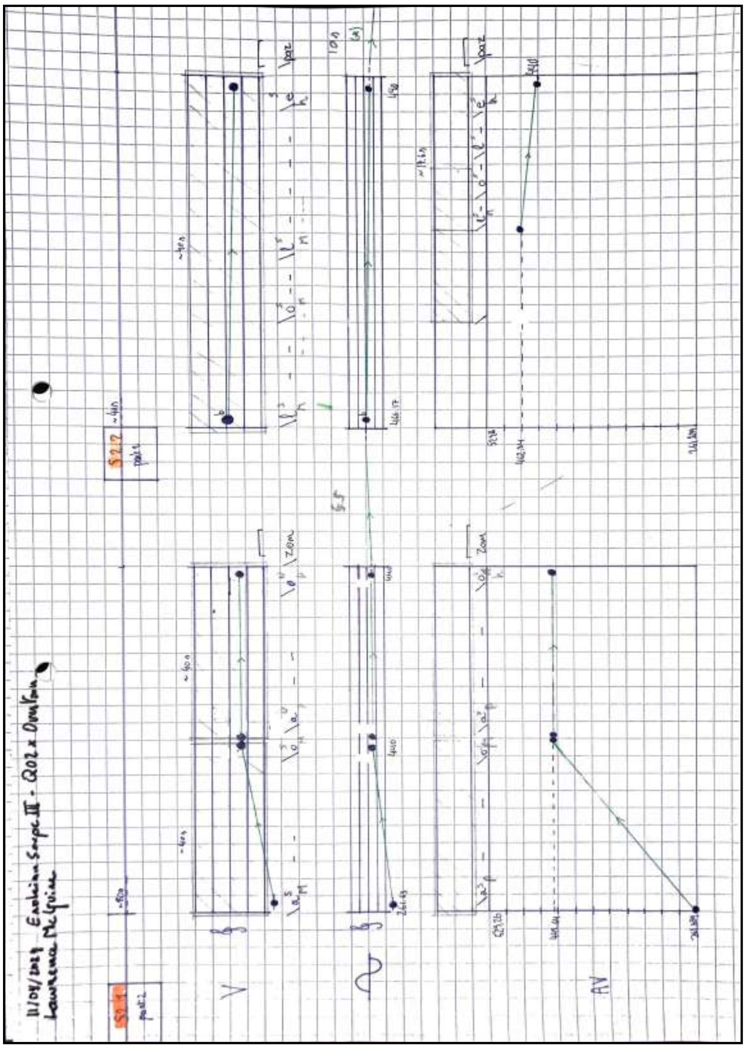
* AS IN "VOICE CRACK"

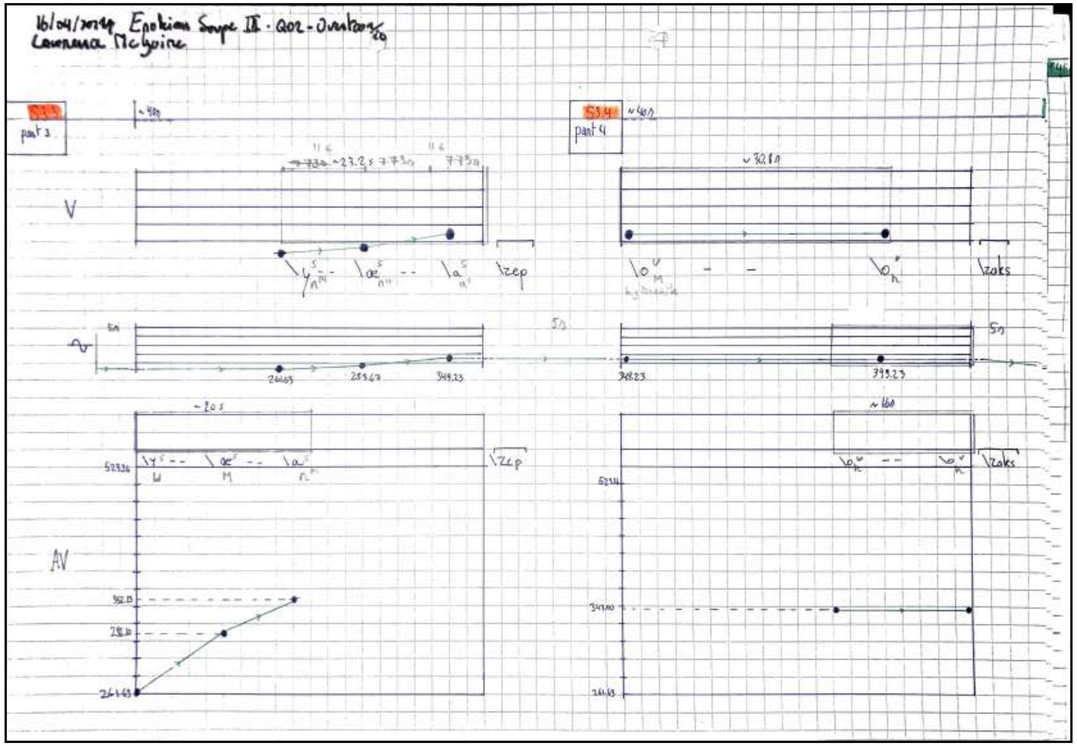
** AS IN SPEECHLIKE VOICE

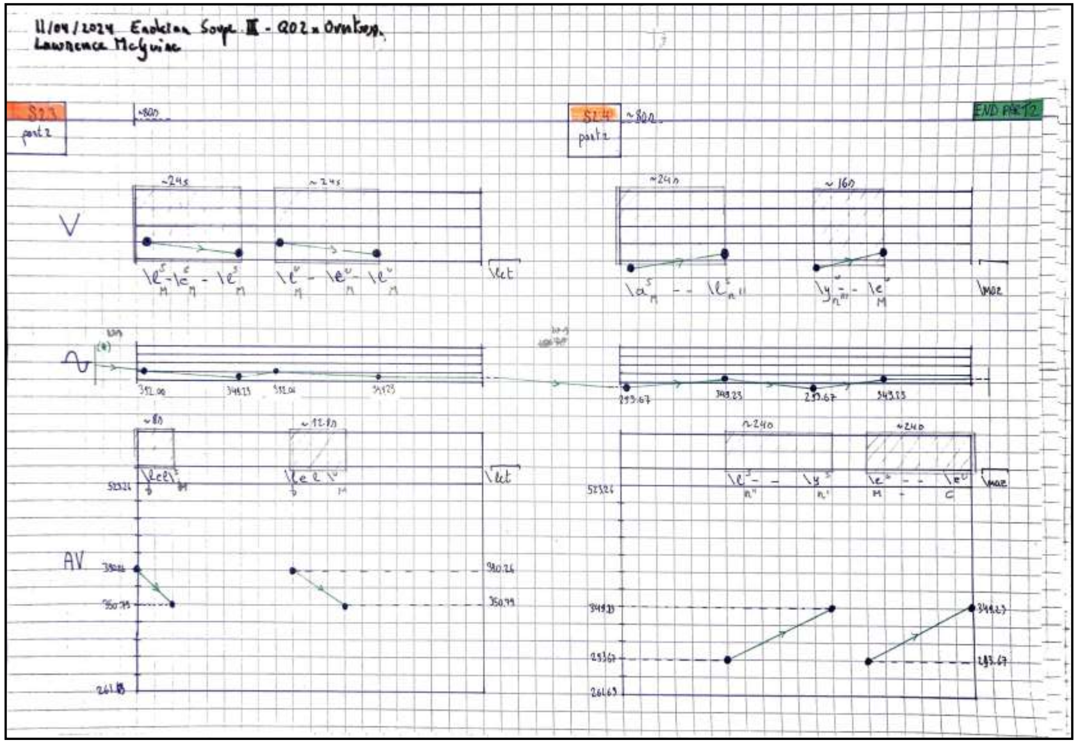
*** AS IN CLOSE NOSTRILS MANUALLY TO A CERTAIN DEGREE : * : stands for barely nasal
** : " " noticeably nasal
*** : " " very nasal

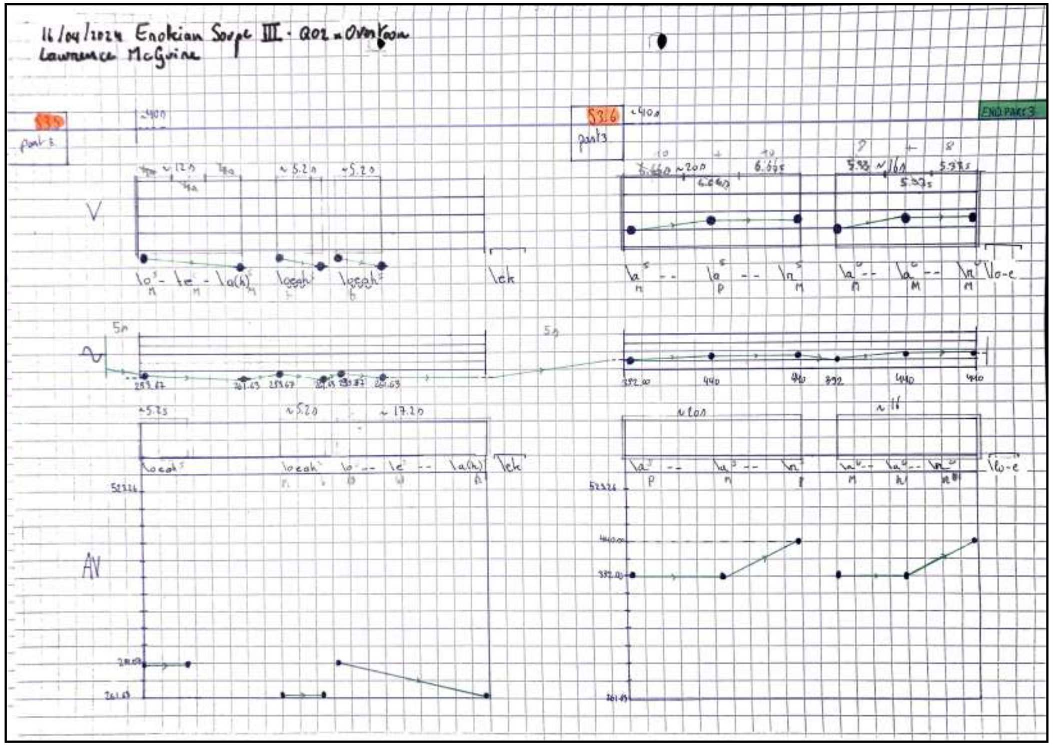


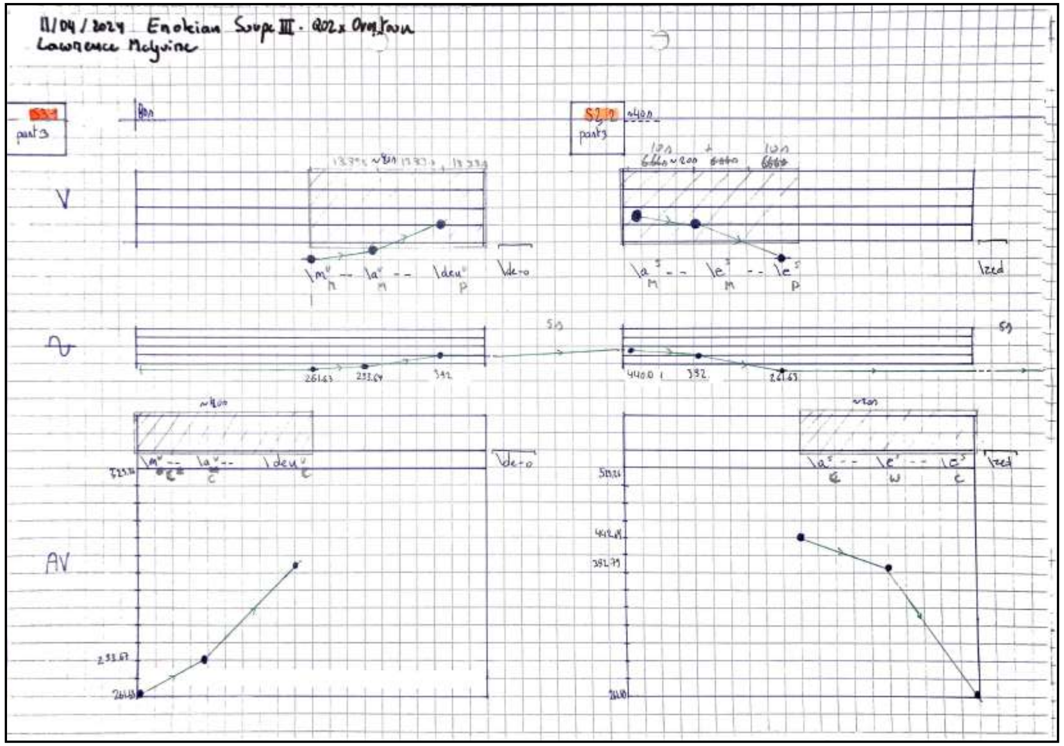




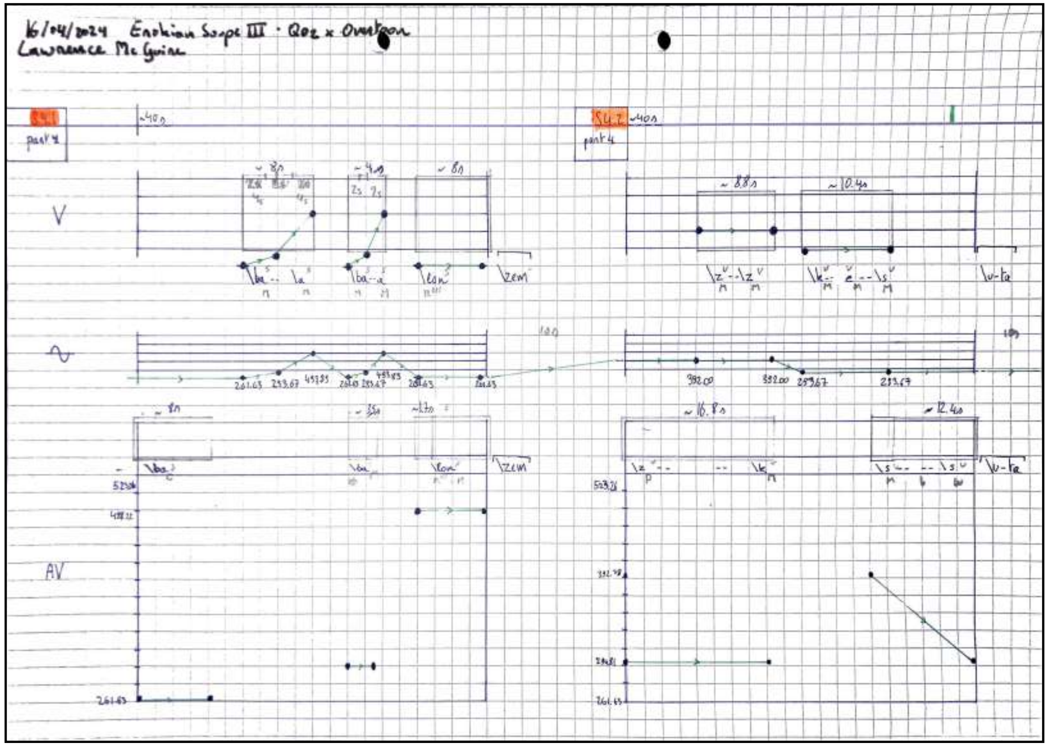

16/04/2024 Enokian Soupe II · Q02 - Overtone
Lawrence McGuire
S3.3
part 3
~40s
S3.4
part 4
~40s
V
~32.8s
5s
AV
~20s
~16s
523.26
261.63
293.67
349.23
\zep
\zaks


11/04/2024 Enokian Soupe II - Q02 = Overtone
Lawrence McGuire
S2.3
part 2
~80s
S2.4
part 2
~80s
END PART 2
V
~24s
~24s
~24s
~16s
AV
~8s
~12.8s
~24s
~24s
523.26
392.00
349.23
293.67
261.63
\ut
\moz

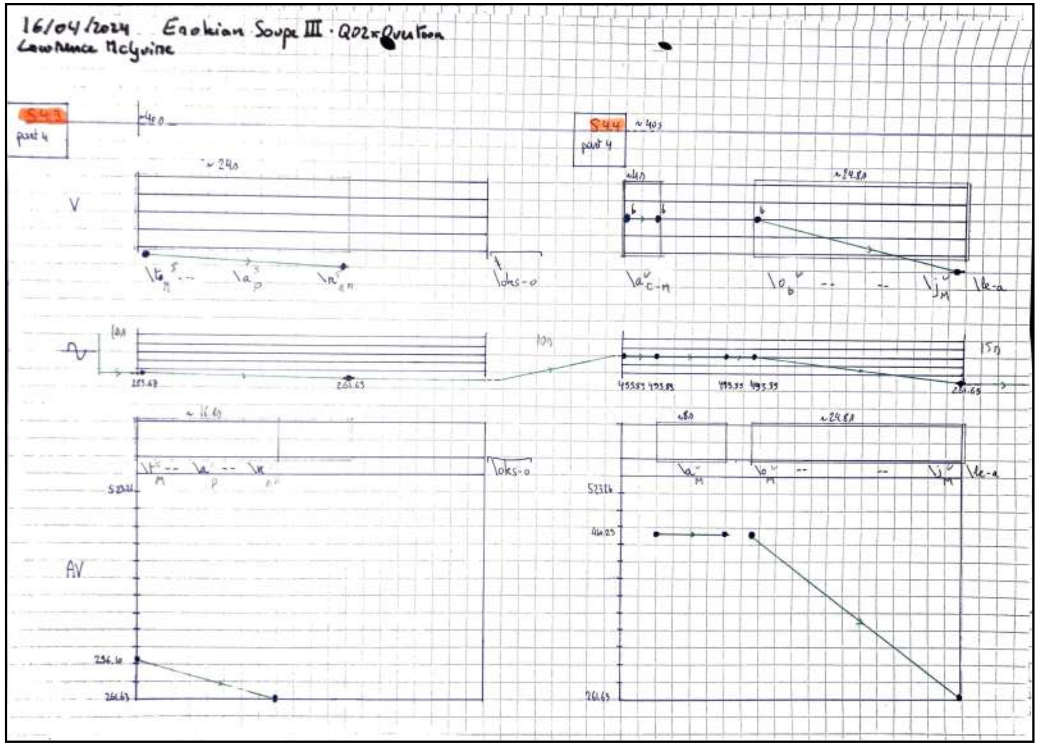

16/04/2024 Enokian Soupe III · Q02 = Overtone
Lawrence McGuire
S3.5
part 3
~40s
S3.6
part 3
~40s
END PART 3
V
~12s
~5.2s
~5.2s
~20s
~16s
5s
AV
~5.2s
~5.2s
~17.2s
~10s
~16
523.26
440
392.00
293.67
261.63
\ek
\o-e


11/04/2024 Enokian Soupe III · Q02 = Overtone
Lawrence McGuire
S3.1
part 3
~80s
S2.7
part 3
~40s
V
~20s
5s
AV
~40s
~40s
523.26
440.0
392
293.67
261.63
\de-o
\zed

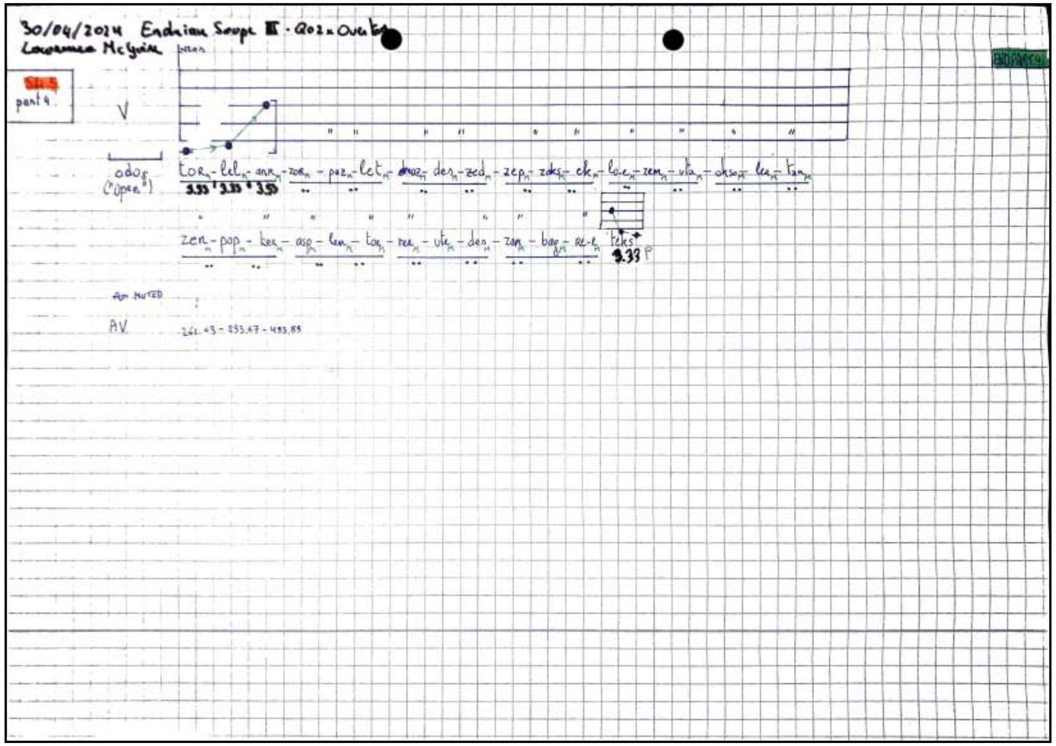
16/04/2024 Enokian Soupe III · Q02 x Overtoon
Lawrence McGuire



16/04/2024 Enokian Soupe III · Q02 x Overtoon
Lawrence McGuire



30/04/2024 Endrian Soupe III · Q02 x Overtoon
Lawrence McGuire


S4.5
part 4
V
odos ("open")
tor - lel - ann - zor - paz - let - draz - der - zed - zep - zoks - ek - loe - zem - vta - okso - lea - tan
3.33 3.33 3.33
zen - pop - ker - asp - len - tor - ree - ute - den - zon - bag - re-e
teks
3.33 P
AV MUTED
AV
261.63 - 293.67 - 493.83

Appendix B:

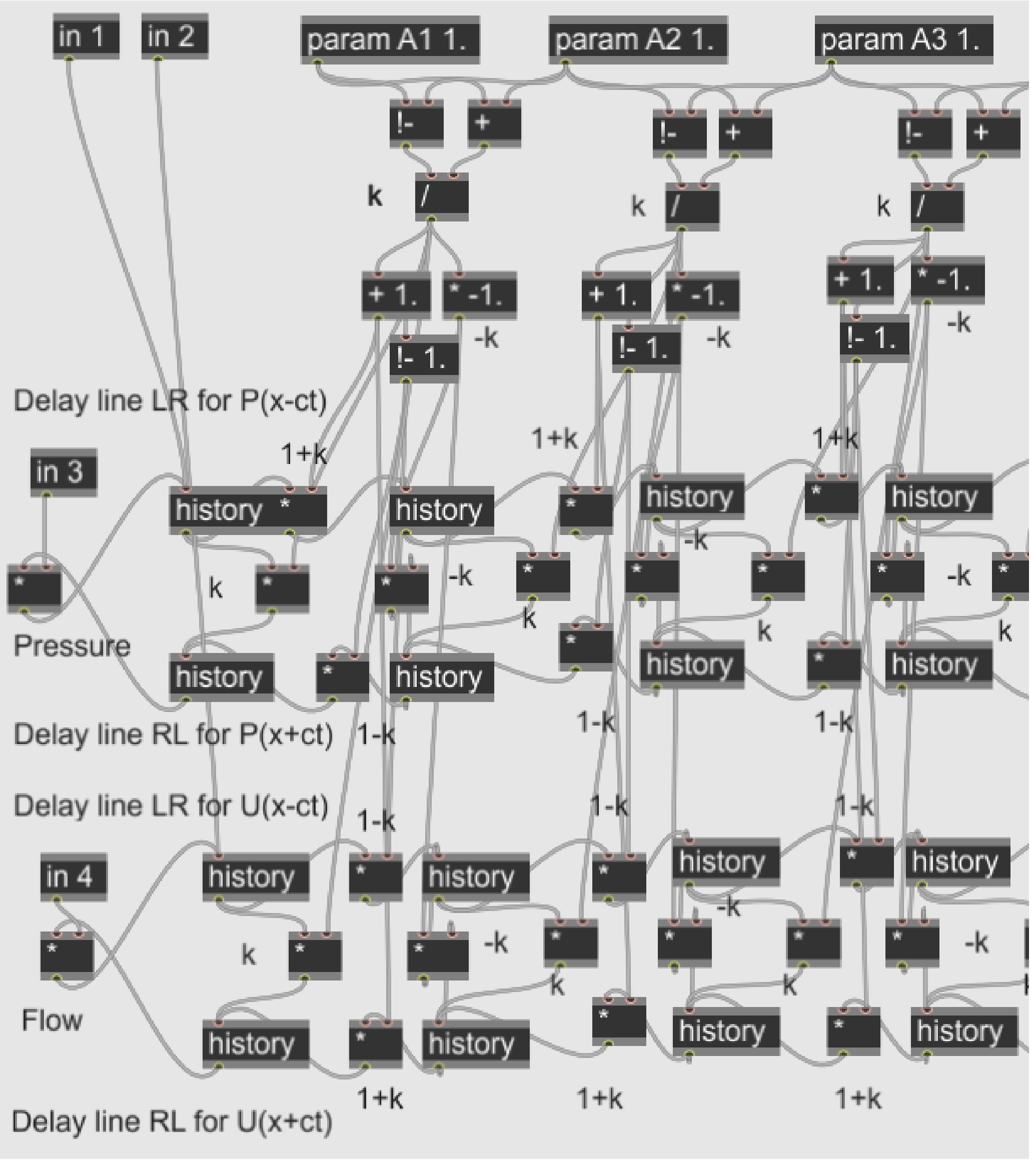
MaxMSP patch ***Keelcore***

The following github link contains the patch used to compose material for *Keelcore* (2021). Read the *README.md* file to understand how to use the program.

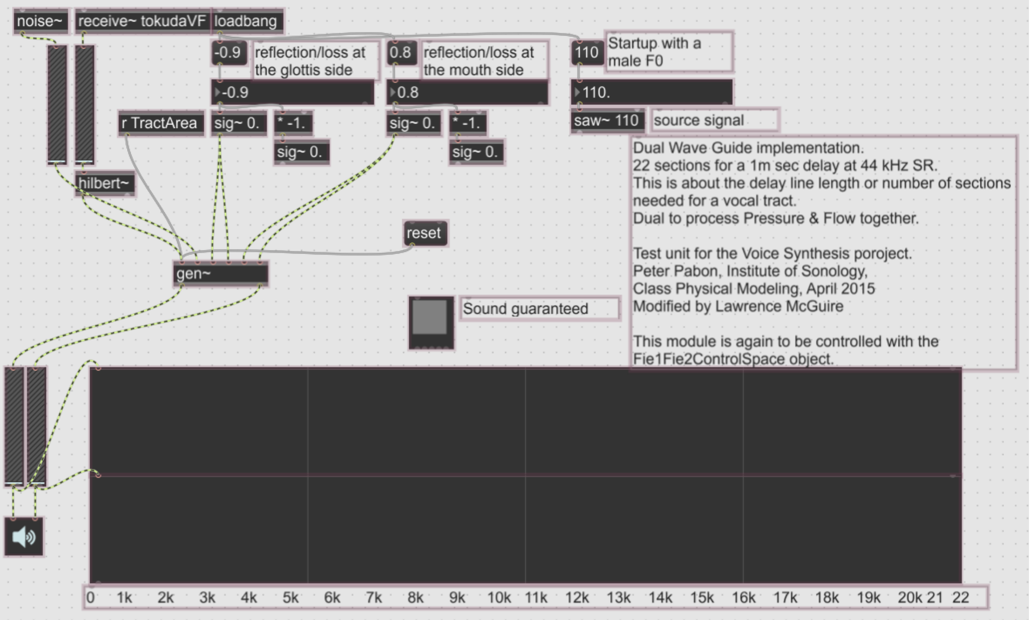
`https://github.com/hogobogobogo/Keelcore-2021-WGF/tree/main`

A 25 tube Kelly-Lochbaum Waveguide was used. A control program was devised to allow for *independent* movements of the first and second formants (F1 and F2). Some snapshots of the program are shown below.

Listen to a recording of *Keelcore* in the sound listings number 36.


in 1
in 2
param A1 1.
param A2 1.
param A3 1.
!-
+
k
/
+ 1.
* -1.
-k
!- 1.
Delay line LR for P(x-ct)
in 3
1+k
history
*
k
-k
1-k
Pressure
Delay line RL for P(x+ct)
Delay line LR for U(x-ct)
in 4
Flow
Delay line RL for U(x+ct)

An implementation in the ∼*gen* environment in MaxMSP of a Kelly-Lochbaum Waveguide filter.

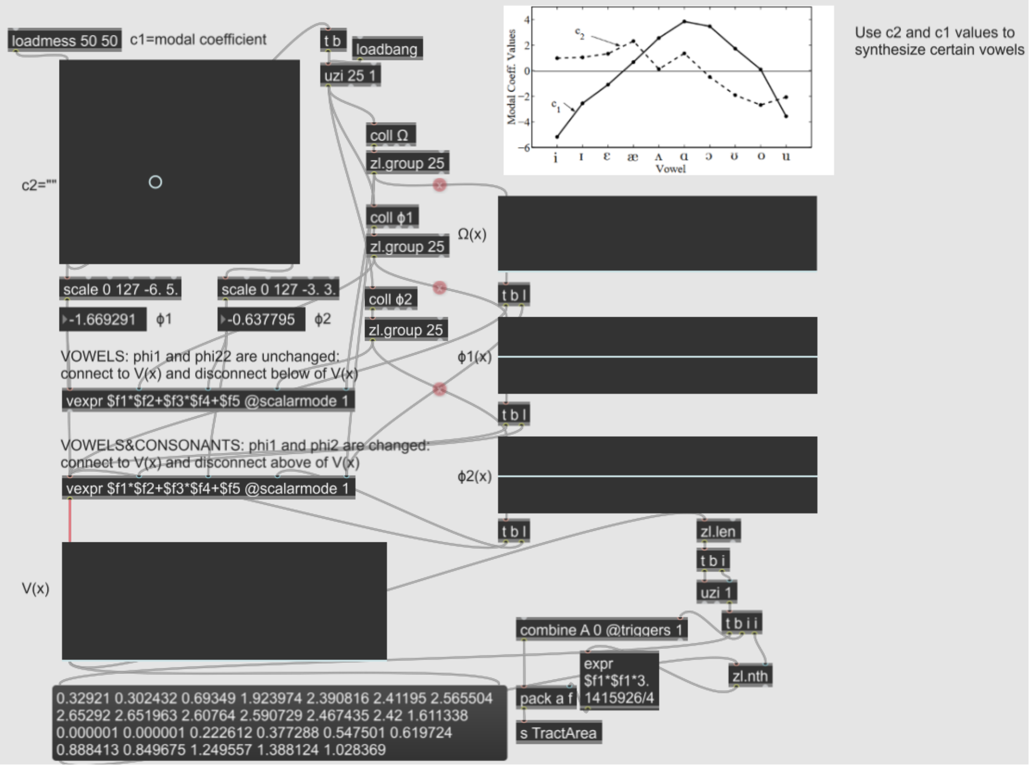


Dual Wave Guide implementation.
22 sections for a 1m sec delay at 44 kHz SR.
This is about the delay line length or number of sections needed for a vocal tract.
Dual to process Pressure & Flow together.

Test unit for the Voice Synthesis poroject.
Peter Pabon, Institute of Sonology,
Class Physical Modeling, April 2015
Modified by Lawrence McGuire

This module is again to be controlled with the Fie1Fie2ControlSpace object.

The central program used in *Keelcore*



Use c2 and c1 values to synthesize certain vowels

VOWELS: phi1 and phi22 are unchanged:
connect to V(x) and disconnect below of V(x)

```
vexpr $f1*$f2+$f3*$f4+$f5 @scalarmode 1
```

VOWELS&CONSONANTS: phi1 and phi2 are changed:
connect to V(x) and disconnect above of V(x)

```
vexpr $f1*$f2+$f3*$f4+$f5 @scalarmode 1
```

```
expr $f1*$f1*3.1415926/4
```

```
0.32921 0.302432 0.69349 1.923974 2.390816 2.41195 2.565504
2.65292 2.651963 2.60764 2.590729 2.467435 2.42 1.611338
0.000001 0.000001 0.222612 0.377288 0.547501 0.619724
0.888413 0.849675 1.249557 1.388124 1.028369
```

The control program used in *Keelcore* to change the vocal tract shape

Appendix C:
NKOAPP instructions

The following github link contains the NKOAPP program used to compose material for *Enokian Soupe I* (2023), *II* (2023) and *III* (2024). Read the *README.md* file to understand how to use the program.

`https://github.com/hogobogobogo/NKOAPP`

---

1. Download the latest version of the **Praat** (`https://github.com/praat/praat`) and **VocalTractLab2**(`https://github.com/TUD-STKS/VocalTractLab-dev`) software.
2. Then fork this repository locally on your computer.
3. If you want to use f0 and intensity data from an audio recording, then follow the following steps:
   - Open `Praat`> *Open* > *Read from file...* and select the audio file you'd like to analyze
   - Then select this sound in the *Object box*, go to the toolbar option *Praat* > *Open Praat script...*
   - Navigate to the folder in directory `NKOAPP\Praat Scripts\` and choose the `f0_intensity.txt` file. Then click on *Run* in the script popup window.
   - Wait a bit for the script to finish, then save the file as a txt-file with an arbitrary name in the folder `NKOAPP\f0andIntensity`. For example as `f0andIntVoxAdam.txt`
4. Open `NKOAPP.py` file and the `VocalTractLab2` software

- In `NKOAPP.py` you have to choose the glottis model `glChoice = glGeomOptions['GM']` with three choices `'GM'`,`'2M'` or `'TRI'`.

  - In `VocalTractLab` choose the same glottis model as `glChoice` (see above): *Synthesis models* > *Vocal folds model* > Choose one of the three > then click on *Use selected model for synthesis*

- If you want to use the f0 and intensity data from the recording, you should add the filename in this part of the code `f0andIntFileName = directory + '/f0andIntensity/**filename.txt**'`

- If we want to manually generate the tractSequence or use the f0 and intensity values from the praat script

  - `manual = True` *go to Section A in the code*

  - `manual = False` *go to Section B in the code*

1. Two ways of working with NKOAPP are either through manual input of glottis parameters, vocal tract shapes and durations of the interpolations between *sources* and *targets*

SECTION A

- **Option A1** for glottis and tract targets (glOption,trOption) and the durations between them

- **Option A2** to generate random interpolations between a set of chosen glottis and tract targets (glOption, trOption) and random durations

- Follow the comments in the code. Don't forget to comment out the parts you don't need when working in another section and another Option

SECTION B

- Same as above, but now the amount of frames in the tractSequence-file amount to the same duration as the audio file. This can not be changed.

1. Change the durations between targets by proportioning the segments in different ways

- Make the durations shorter and shorter starting from the longest part: `durModulationG = arithmetic_progr` or the interpolations get longer and longer starting from the shortest part: `durModulationG = arithmetic_progression(1,valT)[::-1]`

- Change the exponent of the duration segments

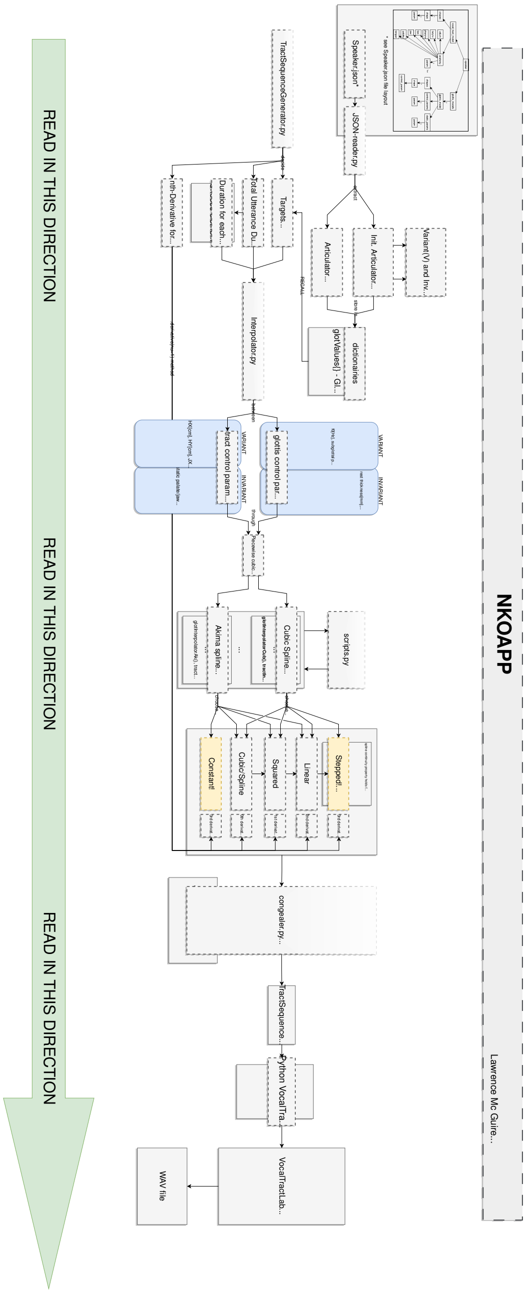
1. Take the time discrete derivative of the splines. The splines are cubic polynomials, so the highest possible derivative is 2. So, three options are available for both the glottis and tract interpolations in `derivativeGlot` and `derivativeTract`. This is still in its experimental state, use at own risk.
2. Choose the upper and lower boundary `devHi` and `devLo` for random uniform number generation as a factor to randomize the glottis parameters, such as f0, intensity, jaw height, tongue height, and more).

Now build the NKOAPP.py file

- After building, there should be a tractSequence file in the `\TractSequence` folder that should look something like this `2024-05-13_02-50-43_TractSequence....Manual.txt`

- Go to `VocalTractLab2` and navigate to *Synthesis from file > Tract sequence file to audio.* Then find the tractSequence file in the beforementioned folder and select it.

- The bigger the tractSequence file, the longer it takes to render.Ignore the warning ‘‘`The Program is not responding`’’.



Low level signal flow diagram of NKOAPP program