CONVOCARE_CONSONARE: A DUET FOR FOUR VOICES

David Gerhard
University of Regina
Dept. of Computer Science
Dept. of Music
Regina, SK Canada
gerhard@cs.uregina.ca

Ellen Moffat
Independent Artist
Saskatoon, SK, Canada
www.ellenmoffat.ca
moffat.e@sasktel.net

ABSTRACT

The human voice continues to provide source material for ongoing explorations in musical expression. As one of the oldest musical instruments, the human voice is simultaneously universal in availability, accessibility, understanding and mystery. It provides rich material for the subconscious interpretation of language and the construction of meaning. This paper presents the development of a phonetics-driven audio-visual synthesis engine, a physical interface to the synthesis engine, and an original sound composition, *convocare_consonare*, which fuses the synthesis engine and the physical interface as an artistic investigation of polyphonic composition.

As background, the phonetic components of voice are explored, considering historical and recent uses of phonetics as source material for musical expression. Technical, acoustic and artistic parameters of voice are discussed, including artificial generation by computer systems and the perception by humans. A set of phoneme classification taxonomies are developed which suggest possibilities for future exploration of voice as musical source.

*convocare_consonare* is the latest in an ongoing collection of works exploring phoneme-level voice in art. The development experiences are presented as a case study in the use of low-level linguistic material as source for music, interactive art, and expression.

1. INTRODUCTION

The human voice is an expressive and eloquent musical instrument. Song and other human-instrument music is pervasive in human history and is one of the primary forms of modern entertainment. Much of the “content” of current popular music is contained in the singer’s lyrics, expression and style. Modern electronic music is often seen to move away from the concept of voice and lyric to more abstract soundscape environments. While the two seem disparate, the notion of the voice can be included in modern electronic music as source material, with several obvious advantages.

Many modern new-music artists have made significant use of phonemes, language, and voice in their compositions [1, 9], and the use of individual components of language as notes or grains of music is not novel [14], but there remains opportunities for exploration. This work describes our experiences with the use of voice in a number of specific contexts.

The human voice has numerous advantages over other source material:

- Voice is recognizable in its original form and has perceptual and psychological immediacy, yet it also permits easy manipulation to create sounds which are familiar and yet not familiar.
- Voice is spectrally rich, allowing the creation of musically interesting content on a note-by-note basis.
- Voice is widely and freely available, not requiring algorithmic or copyrighted acquisition.
- The acoustic properties of voice are interpreted at a subconscious level but can be brought to conscious awareness with little effort or attention. This conscious/subconscious awareness can be used to the musician’s advantage.
- The computational, spectral, and physical characteristics of voice have been studied in other disciplines (notably speech recognition) for many years, yet the application of these studies to artistic expression continues to hold novelty.

1.1. The phoneme misnomer

Phonemes in language do not exist in isolation; rather, each is shaped by the previous and subsequent phonemes in the speech sequence. Indeed, words themselves are not separated as such in the acoustic stream, but are connected one to the next, and even overlap when the final phoneme of one word is identical to or related to the initial phoneme of the next. Given that phonemes do not exist in isolation, how might one effectively take advantage of preexisting cognitive models of the components of language? Section 3 details the experimentation that led to our particular implementations that address these concerns.

Another misnomer of phonetic deconstruction is the categorical perception and production of phonemes. The
particular characteristics of a sound produced as a specific intended phoneme varies widely between individuals, situations and accents. Even within a single individual, no two utterances of a specific phoneme or phoneme combination are identical. Indeed, languages commonly make use of phonemes that are not present in other languages, and are therefore indistinguishable to a listener whose experience does not include that language. Classic examples of this include the [r-ı] \(^1\) differentiation in English that is not present in Japanese, and the [d-ə] differentiation in Spanish that is not present in English [10]. Phonemes do not exist as archetypes of their category; rather, they exist throughout their categorical boundaries, and it is the human perceptual system that builds categorical perception on top of this reality. Every [s] we hear is acoustically unique, but we place them all in the same category. If we did not, we would not be able to create or interpret language.

1.2. Voice, Language, and Art

“The voice belongs first to a body, then to a language.” (John Berger).

The voice is the original human instrument and the body is the original resonant chamber. The disembodied voice, however, reflects the mediated communication of contemporary experience, as is prevalent in telecommunications, radio, and contemporary vocal music. Polyphony of voice, particularly when achromatic, anharmonic or arrhythmic, suggests the (metaphorical) schizophrenia and cacophony of daily life. It may also be a trope for democracy and equity, and a metaphor for the cooperative logic of community and the movement “from isolation to community, the problematic of community, the repeatability of structures (and) the collectivization of the self” [8].

There are, of course, many examples of the artistic use of language and the components of voice. This short and incomplete list of examples demonstrates some of the primary influences for our work on this subject.

Truax [13] has made use of phonemes as source material for granular synthesis, and presents the phoneme as a metaphor for the grain as the base unit of sound [14]. Reich’s compositions [11, 12] evoke rhythm and phasing. Monk [9] uses voice directly as an instrument. Other influences for this work include La Barbara [6], Kosugi [4], and Bok [1].

2. PHONEME AS NOTE EVENT

As is suggested by Truax, phonemes can be considered the fundamental unit of language. As an abstract concept, this is sufficient for the study of language itself, but how can the misnomers described above be addressed such that individual units of language can be re-purposed and applied to the construction of new music? Phonemes themselves can be used as individual note-events in a number of different ways. The primary interpretation of the phoneme as note-event depends on the analysis of the phoneme unit itself. Because phonemes are spectrally rich and laden with pre-purposed meaning, the selection of a phoneme at any time in a musical piece brings more than just the pitch-base or the spectral content. The full internalized meaning of the phonemes must be deeply understood before they can be repurposed to music.

2.1. Phoneme taxonomies and musical usage

The selection of a particular phoneme in a particular musical context depends on the musical intent of the composer or performer. Because of their rich and quasi-familiar spectral content, phonemes can be used for many different musical intentions. Which taxonomy is used depends on the musical need.

2.1.1. Pitch

There are two classes of phonemes as they relate to pitch: voiced and unvoiced. Voiced phonemes make use of a pseudo-periodic glottal pulse as a driving function shaped by the spectral characteristics of the vocal tract. When voiced phonemes are sung, the perceived pitch is related to the frequency of the glottal pulse. Unvoiced phonemes cannot be sung in that sense, since they are driven by a chaotic function instead of a periodic function. Voiced phonemes, whether in isolation or extracted from spoken text, still have an associated pitch, but this pitch is not typically perceived as musical unless it is repeated and/or extended with time-stretching systems such as granular synthesis. Strict repetition of recorded phonemes can be perceived as artificial or “robotic,” so care must be taken when playing recorded phonemes in a loop. It should be noted that voiced phonemes are not restricted to vowels. Voiced consonants include some fricatives ([v],[f]), nasals ([m],[n],[ŋ]) and approximates ([l],[l]).

2.1.2. Formants

The perception of vowels in spoken language is achieved by the relative frequency of a set of formants, peaks in the spectral filter shape of the vocal tract. \(F1\) and \(F2\), the two lowest-frequency formants, are often sufficient for identification of the vowel being spoken or sung. The location of the formants in the frequency spectrum is critical for the perception of the phoneme, so if the phoneme itself is musically important, these peak locations cannot be altered by any significant amount. The change in location of the formants in time-shifted recordings is responsible for the “chipmunk” or “demon” effect of sped-up or slowed-down speech or songs. Small adjustments can be made without altering the perceived human quality, and these changes are frequently made by popular singers during or after recording sessions to clean up minor pitch errors in the singing. Care must be taken here not to over-correct, or again the vocalizations will sound artificial.

\(^1\) Phonemes in this paper will be indicated using international phonetic alphabet (IPA) symbols enclosed in square brackets.
Formants also act as band-pass filters, which can cause a sensation of pitch independent from the fundamental frequency, when heard in isolation, or driven by chaotic sources such as in the case of unvoiced phonemes. The $F1$ frequency location usually provides the strongest sensation of pitch; however, since each voiced phoneme has multiple formants, multiple simultaneous pitch centres are possible.

### 2.1.3. Physiological characteristics

Historical taxonomies of phonemes begin with a high-level dichotomy between vowels and consonants (vowels having an open vocal tract and consonants having a restricted vocal tract), and then differentiate within vowels and consonants. English consonants, for example, are categorized by manner of articulation, which indicates the type of vocal tract constriction, place of articulation, indicating the location of the constriction and phonation indicating how and whether the vocal chords vibrate. Vowels are categorized by the height and back-ness of the tongue, as well as roundness of the lips. Each of these categories and sub-categorizations can be used to imply and infer the acoustic characteristics of the phoneme, and there is a correlation between the formant location (in vowels) and the location of the tongue. It should also be noted here that there is a dichotonic division between spoken vowels, which have an inherent musical pitch, and whispered vowels, which do not have a pitch. Consonants have pitch only when voiced (positive phonation).

This taxonomy of physiological characteristics is the one most commonly used in the study of phonetics and phonology, but at first glance seems obscure when applied to music. Knowing the physiology of phonemes can be informative and useful, however, if the appropriate context is observed. For example, the phonation of a consonant can be used to build rhythmic structures: pulmonic egressive plosives ([p], [t] for example) provide short spectrally rich bursts of noise, and careful choice of such plosives provides a range of noise bursts that can be used to simulate percussive events at different frequencies. Indeed, beat-box artists make use of these phenomena to simulate bass drum, snare, and hi-hat.

### 3. convocare - consonare AND THE USE OF VOICE

*convocare - consonare* is a four-part harmonic and rhythmic composition using phonemes from four distinct data banks as source material. Hand-held controllers allow for authorship through movement and interaction using distance sensors and other physical interface elements to create a collaborative musical performance. A video projection complements the audio composition using phonetic glyphs, providing an environment through visualization of the linguistic-musical script-score.

#### 3.1. The metaphorical synesthetic model

Synesthesia is a neurological condition wherein an individual perceives multiple sensory stimuli from a single external stimulus. Historically, synesthesia has also been used as a term to describe art which incorporates two disparate media, such as the visual and the acoustic. It is this sense of the word which we are using, as it implies a much tighter integration between visual and acoustic elements than mere multimedia. In the case of *convocare - consonare*, the synesthetic model corresponds to acoustic and textual representations of phonemes, as well as a tying together of the control and production of sound.

*convocare - consonare* explores composition, process, authorship, sound and its visualization using phonemes as linguistic note events, referencing choral music. A script-based sequencing function is employed for the score and a text fountain is used for visualization of the sounds as glyphs (the phonetic symbolization of the sound). As small particles of sound and grains of linguistic information, phonemes tease with meaning, suggesting semantic and acoustic characters. They create rhythm, rhyme, and harmony through juxtaposition and manipulation of the sound files in the performance process.

The dynamic, collaborative, interactive script moves the composition beyond a fixed score. Correlating the composition to the users’ movements adds a dynamic element to the compositional process as another form of collaboration. The goal is a kinesthetic work of interactivity, co-authorship and collaborative participation (as cooperation and/or competition) as a metaphor for dialogue and communication referencing four-part vocal harmony as metaphor.

The work is set either as a stage piece or an interactive installation. The main method of interaction is through novel physical devices, which are presented in more detail in Section 4. These devices consist of a central speaker surrounded by sensors and controllers, allowing each device both to be independent and to interact with the others. Each performer holds two such devices (one per hand) as dynamic interfaces and (in this work) each is associated with one set of selected phonemes. The devices allow the performers to manipulate and transform the phonemic script acoustically and semantically through their physical actions. Velocity, repetition, and duration of the sound files are variables. A real-time video projection of glyphs is presented above and behind the players as a circular-framed dynamic image. The project references German sound and visual poetry of the 1960’s and 1970’s [2, 3, 15] and contemporary voice painting [7].

#### 3.2. Source material collection

Existing collections of four different speakers performing pronunciations of phonemes were used as source material for the audio component of *convocare - consonare*. The visual component, closely coupled with the audio component, is a textual representation of the phoneme. The
sound-character pairs are hereafter referred to as “glyphs”. The character components of the glyphs are taken from English sounds and pronunciations identified using the International Phonetic Alphabet (IPA), a character set typeface which contains unique symbols for the phonemes produced in many languages, dialects and accents. The advantages of using IPA were that it presented a unique single glyph for each phoneme (as opposed to the character combinations that are required in some other phonetic representations), and that it offered a quasi-familiar presentation to the audience, one that was on the cusp of recognition without being overtly meaningless.

Since many consonant phonemes (such as the plosives [p] and stops [t] are not available in isolation, each such consonant phoneme performance was in the form of a consonant-vowel pair (cv), with the open back unrounded vowel [ʊ] as in father [fəːtʃə] used to sound the consonant (for example: [pʊ]).

3.2.1. Sources

There are four sets of phonemes (two female, two male) from three sources—the Department of Phonetics and Linguistics at the University College of London, England; the Cambridge English On-line British Council ELT Group; and wikipedia’s collection of phoneme samples from the articles on specific phonemes, which is a conglomerate of various recordings, the ones in question recorded under the GNU Free Documentation License. The use of existing sound recordings from established language institutions is by permission for University College of London, England; the Cambridge English On-line British Council ELT Group; and wikipedia’s collection of phoneme samples from the acoustic field, depending on the presentation of the work. If set as a performance piece, this is achieved by having each sound emanate from the speakers contained within the performer’s devices. If set as an installation, with the projection on the floor, then spatialization is achieved with a set of speakers on the floor and low on the walls, and the sound is made to come from the same location as the character.

In a further extension of the work, glyphs are located in a concentric pattern reminiscent of Zuverspaeteter-fidurinnemenswerttollos [5], with new glyphs emphasized and old glyphs fading into the background, as seen in Figures 2 and 1. As a time-based media, the video deposits a cumulative graphic of visual æsthetics as a palimpsest that references the phonemic sounds rather than the phonetics or text of language. The act of translating the linguistic sound-bites into graphic art

“relates to the manner in which linguistic signs, as [to] shape and spatially, no longer evolve as textual lines but as textual planes. In terms of concrete realization, they don’t appear as a message but as visual aesthetic information” [15].

Several competing technical methods for accumulating and fading glyphs were considered. The first method consisted of each glyph being driven by a unique and separate instance of [jit.gl.text3d] (the jitter wrapper for the OpenGL text3d method, with the above mentioned unicode problems), maintaining complete control over the attributes of each glyph, but requiring a large number of objects to be in memory at once. The second method consisted of rendering a glyph with [jit.gl.text3d], but then immediately merging that glyph with an accumulating background of all other text glyphs previously generated. The third method consisted of pre-rendering

http://fontforge.sourceforge.net/
glyphs as images, and then selecting and manipulating these instead of the text characters, requiring less memory but maintaining control over the objects. Difficulties in transparency layers, resolution controls and object FORK-ing led us to select the second method. The drawback of this method is that once a glyph is placed, it cannot be manipulated individually as it is now part of a conglomerate. This resulting conglomerate can be manipulated, however, and a de-saturation function is employed on the conglomerate to achieve the effect of the glyphs fading slowly into the background.

Two options for the manipulation of the conglomerate and are used for different purposes. If the piece is set as an installation, the conglomerate is de-saturated but not faded, leaving all glyphs present, as seen in Figure 1. If the piece is set as a performance, the conglomerate is both de-saturated and faded (Figure 2) so that only the more recent glyphs are visible. This allows the performers more freedom to move from section to section of a more constructed and composed piece.

Figure 1. Glyph imagery without decay. Each new glyph adds to the noise, and only by tracking new glyphs can a sequence be found.

Figure 2. Glyph imagery with decay. The glyphs fade into the background and are lost, as ethereal and temporary as the sounds they represent.

Figure 3. Glyph imagery in four voices. Each voice has a colour range and, as they decay, saturation is reduced causing the histories to become intertwined and confused.

3.5. Stochastic Variations

The relative location and brightness of each glyph appearing on the screen is driven by a stochastic process, which can be strongly controlled or gently directed by an audience or a performer, depending on the presentation context. The location of the glyph in space follows a brownian motion simulation algorithm (such as a random walk) constrained by the screen size and the concentric nature of the display. The random walk is performed in Cartesian co-ordinates, and the alignment of the character itself is set based on the angle towards the origin. The pseudo-randomness of the brownian simulation seems to present a path from one glyph to the next, implying further underlying meaning. The order and frequency of selection of the glyphs can also be pseudo-random, generating a cacophony, or be constrained using a score, as described in Section 3.6. Constraining the glyph selection to a score presents a more composed feel to the piece, implying further exploration of meaning.

Other characteristics of the glyph production system are driven by stochastic processes as well. The size, scale, and colour of the character are controlled by similar but independent pseudo-brownian processes, giving a sense of purpose or intent to the character path.
3.6. Scripting, scores and the hints of language

As a further addition, a scripting foundation was developed to allow strings of glyphs to appear in sequence. The scripts provide a set of constraints on the allowable range of the stochastic process of choosing the sequence. A single script entry consists of the glyph or set of acceptable glyphs, followed by the range of acceptable durations. Scripts are written to select glyphs for any number of purposes. An obvious first step is to build scripts that generate pseudo-language, for example, to select, in sequence, the glyphs [f], [l], [s], [o], [r], [z] to create an acoustic phrase that almost approaches the comprehension of the word “flowers”. Rhythmic characteristics of the acoustic phrase are accented by choosing appropriate stochastic ranges for the duration of each of the glyphs, thereby giving a slightly different sounding of the phrase each time it is played. The intent is to hint at the reconstruction of language from granular pieces within the greater context of the piece, without approaching artificial speech or generated linguistic forms.

The artistic intent is to build a sense of “almost-ness” or linguistic frustration, sitting on the cusp of understanding. Further, when the listener hears language-like sounds in repetitive combination, unintended phrases can begin to be perceptible, and the piece then begins to act as a sort of pseudo-Rorschach spot, where the perception of the piece is informed as much by what the audience brings to the piece as what the composer builds into the scripts.

3.7. Four Voices

The subtitle of the work is “a duet in four voices”. In the performance version, there are two performers, each controlling two distinct voices. There are two female and two male characters, allowing for a four-part harmony of soprano, alto, tenor and bass as a musical metaphor. Selected consonants—fricatives and plosives—are used for their acoustic qualities, as described above, as well as their semantic associations and their stability for pitch shifting. Vowels are selected for their pitch consistency, to provide a musical drone within the musical composition and to potentially trigger associative meaning for the audience.

4. INTERFACE

The previous section detailed the musical motivation and source material for *convocare_consonare*. The following section details the creation of a novel interface device that brings together the control and the production of sound into a single unit. This device is introduced in the context of *convocare_consonare* but is not restricted to this work.

4.1. The Device

In developing a theoretical basis for *convocare_consonare*, it became apparent that the standard modes of interaction would not suffice. The use of phonemes and the hinting at language is such an intimate and personalized aesthetic that there was a need to draw the speakers close to the performers. Although current spatialization techniques can effectively pan between a set of on-stage speakers, individual performers are still often disconnected from the sounds they create, when interacting with computer-driven music. For this reason, a novel physical musical interface was developed, consisting of raw mid-range speakers in a sparse enclosure serving as a substrate for interface controls. Figure 4 shows the design of these speaker devices as well as the fit to the hand. Each device can be held and operated by a single hand, somewhat in the style of a data glove, but with the primary objective of being a source of sound for the performer and audience.

![Figure 4](image-url)

Because the controls are drawn together with the sound production, this device is much more in the spirit of traditional musical instruments, with the flexibility of computer music synthesis. This is in fact one of the deeper motivations of the development of this device. Traditional computer music interfaces have relied on acquiring control data, with the assumption that the control data would be interpreted by a sound generation engine and the resulting sound would be played back through loudspeakers. One difficulty with this model is that the electronic musician is separated from the production of sound, and the sound is separated from the gestures and motions that created it. This device is an attempt to bring computer sound control and computer sound production together into a single compact device more in keeping with the spirit of acoustic instruments, where the control and the production of sound is intrinsically and inseparably linked.

Each speaker device has a set of buttons controlled with the fingers, and a mini-joystick or a set of rockers controlled by the thumb. This design draws from current video-game controller pads which typically have a joystick controlled by the thumb as the primary controller. In *convocare_consonare*, the buttons control the use of individual phonemes, and the joystick controls parameters of performance such as repetition rate.

The final controller on the device is an infra-red sensor mounted on the facing of the speaker itself. This provides information about proximity to other objects or devices.
in the performance, and is used specifically to identify distance from another performer’s devices. The infra-red sensor is used to control the volume of the speaker, in a direct relationship so that when a speaker is close to an object, it will be quieter, mirroring the more subtle expressions and gestures, as well as the intimate interaction between two people talking closely. Alternatively, the volume could be inversely related to the distance so that when approaching an object, the speaker seems to become more aggressive and agitated, referencing our human desire for personal space.

Each sensor and controller is wired to a phidget\(^3\) interface board via a custom onboard controller substation which gathers all signals and sends them over a standard RS232 cable. The speaker is wired to a 4-channel audio interface via a standard 1/4” instrument cable. Because of these connections, there is a pair of wires that connects each device to the computer controlling the sound.

Figure 5. *convocare_consonare* in performance. Each performer wears two of the controller devices and controls two of the four voices.

4.2. Discussion of the device

Each speaker is heavy enough that holding and interacting with the device is somewhat fatiguing, so a strap system was devised to secure the device to the performer’s hand. With the strap system in place, the performer can maintain otherwise awkward positions without dropping the device or losing the ability to manipulate the controls.

The number and type of controls were selected to suit both human ergonomic motivations and the requirements of the piece. Many alternative configurations are available that allow different levels of control over specific performance parameters. In the current implementation, each finger is presented with a single button, but a pair of buttons for each finger might be more useful, and an accordion-like offset grid may also be effective. As an alternative to a single-pole button, a photoresistor or a force-sensitive resistor could be employed, which would provide varying levels of contact. The two-axis joystick employed in the initial design can also been replaced by a single axis vertical rotary control (essentially half a joystick) which may perform better depending on the application and the control required by the performer.

Future development will include migrating toward wireless or onboard processing\(^4\) as well as wireless or onboard sound production, gyroscopic inertia sensing, and infra-red or ultrasound identification allowing each device to recognize other devices in the performance. Future device development will also include investigating alternative packaging and integration, as well as alternatives for power supplies, since onboard sound production will require heavy batteries, increasing the strain on the performer’s hands, wrists, and arms. An intermediate option would be to employ a secondary processing and power unit that the performer would wear on a belt, with standard wired connectors from this unit to the devices.

A significant advantage of this device over other electronic musical controllers is in the inherent spatialization of sound that it provides. Each performer has complete control over where the sound is perceived to be coming from, and can use that control to full artistic effect. The spatial location of the sound production becomes part of the performance, with the sounds interacting with each other as well as with the performers and the audience.

5. CONCLUSIONS

The electronic musician has limitless source material from which to draw to create meaningful music. The phonemes that are granular components of human communication provide more than just spectrally rich content, rhythmic events or pitch contours, they bring underlying and often subconscious meaning and purpose to a piece.

*convocare_consonare* is a work developed and composed by the artists which makes direct and intentional use of the range of artistic expression available through the use of phonemes. Tight integration between the visual, the acoustic, and the physical, provide a deeper sense of artistic synæsthesia.

The concurrent development of a physical device, a synthesis engine, and a set of source material means that

---

3 phidgets are modular physical interface devices controlled through a USB connection. http://www.phidgets.com

4 using, for example, the Arduino open source physical computing platform. http://www.arduino.cc
the piece itself is cohesive and complete, but by keeping the development modular, each piece can be used in the development of new components, presentation styles, pieces, synthesis engines or physical interfaces.

6. ACKNOWLEDGMENTS

The authors wish to acknowledge the support of the Natural Sciences and Engineering Research Council of Canada, the Canada Council for the Arts, and the Saskatchewan Arts Board.

7. REFERENCES


