T-Voks: the Singing and Speaking Theremin

Xiao Xiao, Grégoire Locqueville, Christophe d'Alessandro, Boris Doval

T-Voks: the Singing and Speaking Theremin

Xiao Xiao
Grégoire Locqueville
Christophe d'Alessandro
Boris Doval
LAM - Institut Jean le Rond d'Alembert
xiaosquared@gmail.com
{gregoire.locqueville,christophe.dalessandro,boris.doval}
@sorbonne-universite.fr

ABSTRACT

T-Voks is an augmented theremin that controls Voks, a performative singing synthesizer. Originally developed for control with a graphic tablet interface, Voks allows for real-time pitch and time scaling, vocal effort modification and syllable sequencing for pre-recorded voice utterances.

For T-Voks the theremin's frequency antenna modifies the output pitch of the target utterance while the amplitude antenna controls not only volume as usual but also voice quality and vocal effort. Syllabic sequencing is handled by an additional pressure sensor attached to the player's volume-control hand.

This paper presents the system architecture of T-Voks, the preparation procedure for a song, playing gestures, and practice techniques, along with musical and poetic examples across four different languages and styles.

Author Keywords

theremin, vocal synthesis, gesture

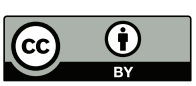
CCS Concepts

•**Applied computing** → **Sound and music computing;** Performing arts; •**Human-centered computing** → *Gestural input;*

1. INTRODUCTION

Invented in 1920, the theremin is one of the world's first electronic musical instruments [19]. It is particularly visually attractive as there is no contact between the player and the instrument, only the dance of both hands around the two antennas. Proximity to one antenna controls the frequency while the other controls the amplitude of the output sound, which is traditionally generated by an analogue oscillating circuit using the heterodyne principle.

Despite its relatively simple waveforms, the theremin's expressivity has often been compared to the human voice due to the sensitivity of its control interface. Like the voice the theremin allows for subtle melodic variations (vibrato, portamento, etc.) and refined volume control. The present research aims to further the analogy between the theremin and the singing voice, taking advantage of recent work in performative singing synthesis.



NIME'19, June 3-6, 2019, Federal University of Rio Grande do Sul, Porto Alegre, Brazil.

One strategy for high-quality performative synthesis is based on vocoding and modifying pre-recorded voice samples, as demonstrated by Vokinesis [13, 12]. In this system, intonation and vocal effort are controlled using a graphic tablet. Articulation timing and rhythm are defined by syllabic control points, which can be triggered by the press and release of a control button or through the continuous variation of a fader pedal.

The present work is built upon Voks, a new performative synthesizer based on the WORLD vocoder [28, 26] that improves upon the capabilities of Vokinesis. While Voks can be used with the same graphic tablet interface as Vokinesis, we have mapped the control of intonation and vocal effort to the pitch and volume antennas of the theremin. The addition of a pressure sensor attached to the player's volume modulating hand triggers the advancement of syllabic control points. This combination of Voks, the theremin and pressure sensor is called T-Voks[1].

This paper first summarizes prior work in vocal synthesis and theremin augmentation. Next, it presents the technical side of Voks, describing underlying principles of voice analysis, synthesis and control, as well as the software and hardware implementation. Considerations from the performer's side are then discussed, including playing gestures, practice techniques, and a set of musical and poetic examples across languages and styles.

2. RELATED WORK

2.1 Speech and Singing Synthesis

While speech synthesis manages to achieve convincing results, the synthesis of expressive voices remains hard. *Performative vocal synthesis* approaches this problem by using human gestures in real time as an input to add expressivity to an artificial voice. Several such speech or singing synthesis systems have been proposed in the past. These systems can be differentiated by their synthesis algorithms, the level of control for the performer, and the control interfaces used.

Some performative voice synthesis systems generate sound from the ground up, usually using a formant synthesis algorithm [16, 17]. This approach allows for free speech, but the numerous parameters involved can be difficult to control. For that reason, other systems make use of a pre-recorded voice [22, 13]. Intermediary approaches have also been devised, such as Glove-Talk I [15], where specific gestures are mapped to specific short words, whose characteristics are used to determine the speaking rate and the stress of each word. Another intermediary approach is diphone-based concatenative synthesis, which offers both the flexibility of pure synthesis and the realism of re-synthesis. In this method, speech is synthesized by the concatenation of short

[1]T-Voks demo video: `http://youtu.be/jJdVsv_-WIo`

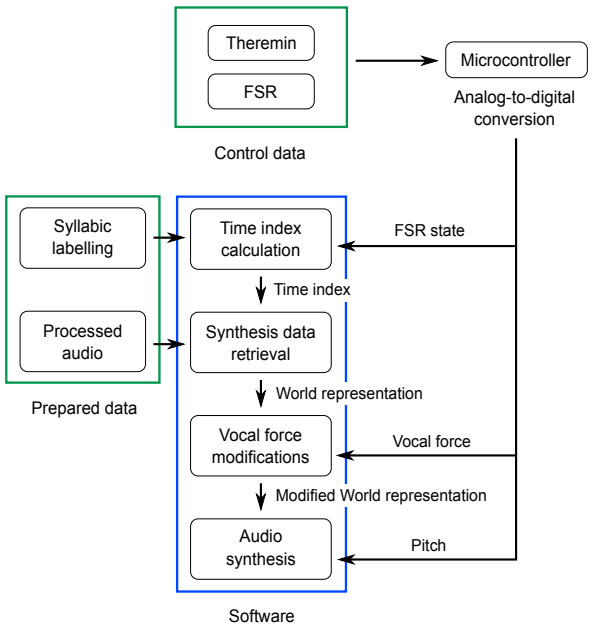


Figure 1: *T-Voks principles and architecture. Control gestures by the player are in shown the top green frame (FSR stands for force sensitive resistor). The left green frame represents the linguistic data prepared for performance (recorded voice and syllabic control points). The blue frame represents real-time processing of gestural and linguistic data for performative singing synthesis.*

sound pieces, from a dictionary of all possible phoneme to phoneme transitions in a language (e.g. about 1200 units in French). Such algorithms are used in offline synthesizers such as Vocaloid [21].

Speech control involves numerous parameters. For real-time use with input from human gestures, systems must make a tradeoff between degree of control and ease of use. On one end of the spectrum, Glove-Talk II [16] offers direct control over several vocal articulation mechanisms. However, even a well trained performer (accomplished pianist, over 100 hours of training) "finds it difficult to speak quickly, pronounce polysyllabic words and speak spontaneously" [16]. The resulting sound quality is similar to early formant-based text-to-speech systems. On the other end, Calliphony [22] lets a performer reproduce a pre-recorded signal with a speed and fundamental frequency chosen in real-time. The sound quality and intelligibility are excellent, but the linguistic material is fixed for a given performance, and timing control is unnatural.

Vokinesis [13] offers more direct control over rhythm while still being based on existing audio. The timing and rhythm of consonants are controlled through sequencing syllabic sized chunks using various methods like tapping or continuous expression pedal motions. Vokinesis enable the control of continuous voice parameters (e.g. pitch and vocal effort) and can be used with continuous control surfaces, such as the Seaboard, the LinnStrument, the Soundplane, the Hakken Continuum, or any other interface that outputs Multidimentional Polyphonic Expression (MPE) data. The interface of choice for Vokinesis is the graphic tablet, which offer a two-dimensional continuous surface, as well as stylus-pressure detection. It enables the reuse of familiar gestures from handwriting, which most people have learned at a young age [8, 11, 17].

In summary, it does not seem possible to control all the aspects of voice production through only hand (or feet) gestures. Re-sequencing a pre-recorded voice is a good compromise between modification capabilities and sound quality. Diphone based concatenative singing synthesis sounds very natural, but it can hardly be applied to real-time performative synthesis because it does not seem possible to select any arbitrary sequence of diphones on the fly (the task would be to select one diphone among about 1200 in less than about 150-200 ms). The syllable seems a better candidate as the basic unit for real time speech or singing control, where syllables more or less correspond to the musical notes in a score. This is the solution chosen for the augmented singing theremin.

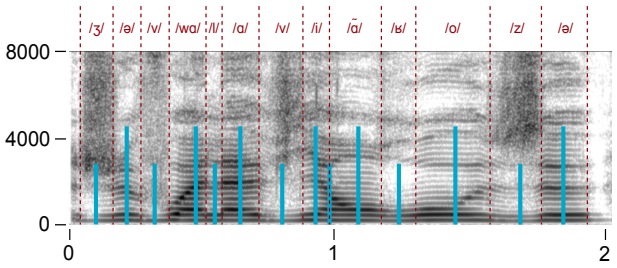


Figure 2: Syllabic control points
Spectrogram of a voice recording of the French sentence "Je vois la vie en rose" used as input for T-Voks (x-axis labelled in seconds, y-axis in Hertz). The phoneme boundaries are marked in red, and the location of control points (see section 3.2.3), in turquoise. When opening her hand, the performer triggers a gradual increase of the time index with a control point associated with a tall mark; when closing it, the target is set to a control point associated with a short one.

2.2 Augmented Theremin & Gesture Singing

To increase the sonic possibilities of their instrument, contemporary thereminists such as Dorit Chrysler and Carolina Eyck employ guitar pedals such as delay and reverb, as well as loopers to create additional layers with their voices [3, 1]. A vocal formant filter pedal can be used to make the theremin sound even more like the human voice, as exemplified by Rob Schwimmer [5]. The theremin has also been used as a controller for modular synthesizers, most notably by thereminist Coralie Ehinger [2]. Other players, such as Lyn Goeringer and Rachel Gibson, have used the theremin's control voltage (CV) output to manipulate textures generated on the computer [4, 18].

Recent research has begun to explore the intersection between singing and freehand gestural control. The Theremin Orchestra combines live vocal performance with capacitive sensors used in traditional theremins, some of which apply effects to the singers' voices while others translate the singers' hand gestures into traditional theremin sounds [9]. The Gestural Envelopes project employs wearable inertial motion sensors to track a performer's hand gestures, which control the syllable timing of pre-recorded singing, as well as some effects [23].

3. DESIGN AND IMPLEMENTATION

3.1 Control Principles

T-Voks is based on syllabic control and modification of pre-recorded voice utterances. There are three main parameters:

1. **Intonation** is altered by the vertical antenna of the theremin. Pitch scaling requires modification of the signal fundamental frequency (f_0) using a vocoder.

2. **Vocal effort** is adjusted by the loop antenna, or volume control of the theremin. Realistic vocal effort modification is a complex process: it involves joint sound intensity, spectral slope, voice/unvoiced ratio modification. This requires a specially designed vocoder and modification rules [29].

3. **Syllabic Sequencing** is managed by an additional interface, managed by the volume-control hand. This requires syllabic labels associated to the signal and time-scaling of the synthesized signal using a vocoder.

Syllabic control is the main difference between playing T-Voks and the usual theremin. It is rooted in the frame-content theory of speech production [24]. This theory postulates that speech utterances can be decomposed into syllabic "frames," associated with the opening/closing motions of the jaws, and "content," associated with smaller units like consonants. The speech rhythm is given mainly by the frames, although the content represents the details in articulation (i.e. the phoneme articulation).

For performative synthesis, controlling the frame timing is necessary and sufficient. Each frame is defined by two time control points: one for the open phase, and one for the closed phase. Figure 2 displays the placement of control points on a speech utterance. Note that a control point alone for each syllable is not sufficient: a biphasic control is needed [12].

3.2 Software

Written in C++ and Max, Voks is the performative singing synthesizer software behind T-Voks. Like its predecessors Calliphony [22, 10] and Vokinesis [13, 12], it is based on the time and frequency scaling of recorded voice utterances, with the possibility of some voice quality modifications. Rhythm is controlled by resequencing using the same syllabic control points as Vokinesis.

Voks is implemented using the WORLD vocoder, which features better sound quality and much improved robustness. It is designed with a modular architecture, with a control shell and an external procedure for syllable labeling. This modularity enables Voks to be incorporated into new programs. T-Voks is one such example. Voks can also be used with the same interfaces as used with Vokinesis, such as a graphic tablet or a fader pedal.

The following sections describe the functionality of the WORLD vocoder, how a new song is prepared, and how a song is synthesized during performance.

3.2.1 The WORLD Vocoder

Designed for speech synthesis, the WORLD vocoder [28, 26] is free software that allows real-time signal-level modification of a voice recording.

WORLD consists of two independent units, an analysis module and a synthesis module. Those units respectively allow for conversion from a monophonic audio file into a specific spectral representation (analysis), and conversion from that representation back into playable audio (synthesis). While analysis is only possible on an existing audio file on disk, the synthesis module can be used in real-time; the input representation to the synthesizer is updated just as sound is output from it.

The WORLD representation is based on a modified version of the source-filter model of speech production [14] . In WORLD's version of that model, a voice signal is modeled as the sum of a filtered pulse train (the periodic part) and filtered white noise (the aperiodic part). To describe the input sound, the representation needs to include at each time

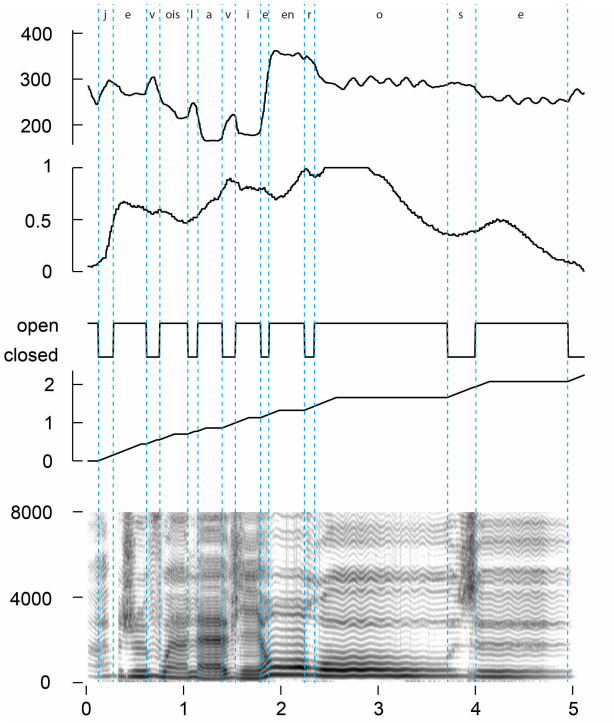


Figure 3: T-Voks performance
Time evolution (in second) of control parameters during a T-Voks performance of the sentence "Je vois la vie en rose." From top to bottom: syllable labels, fundamental frequency (in Hertz), vocal effort, FSR state and the resulting analysis time (in second), and a spectrogram of the output sound (ordinate labelled in Hertz). The opening/closing instants, marked in turquoise, do not correspond to control points, but trigger an increase in the time index until it reaches the next control point.

both the frequency of the pulse train and enough information to be able to recreate the two filters. In practice, those data are specified every 5ms, and the filter information is stored as two floating-point number arrays. One array contains a power spectrum for the whole signal (both the periodic and aperiodic parts). The other array specifies at each frequency bin the ratio of the power from the aperiodic part of the signal with the total power of the signal.

The f_0 is estimated at each frame by the Harvest algorithm [27]. Harvest acknowledges that the input audio might be entirely aperiodic at some frames, and does not specify a f_0 for those frames. The f_0 computed by Harvest is not directly used by T-Voks but is used by two subsequent algorithms, CheapTrick [25] and D4C [26]. CheapTrick computes an estimate of the power spectrum of the signal at each frame. D4C then estimates, at each frame and each frequency bin, how much of that power spectrum comes from the periodic part of the signal and how much comes from the aperiodic part.

To generate audio, the synthesis module takes a power spectrum and an aperiodicity ratio arrays as input, as well as a numeric value for the f_0. If fed with that input at a sufficiently high frequency, it can generate a realistic voice in real-time. The synthesis module has been ported into a Max object, allowing for its use in the T-Voks Max patch.

3.2.2 Voice Signal Modifications

In the WORLD representation, the input voice signal is split into three sets of parameters at the analysis stage: the f_0, the periodic spectrum and the aperiodic spectrum. At

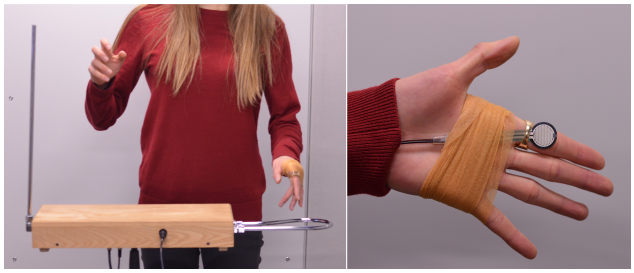


Figure 4: *Augmented theremin: a Theremin, a FSR, and an analog to digital interface. Analog control signals by the performer can then be processed using MAX/MSP. The whole interface being played is shown on the left. The right hand controls pitch, by subtle variations of distance between the hand to the upright antenna and hand shape. The Left hand controls voice quality, by variation of distance between the hand and the flat antenna. The left hand also controls biphasic syllabic sequencing, by pressure of the thumb on a FSR. The location of the FSR in the performer's left hand is shown on the right.*

the synthesis stage, these three sets of parameters are input into the synthesis engine. Modifications take place between analysis and synthesis.

Pitch scale modification is straightforward: the synthesis f0 contour replaces the analysis f0 contour. Time scale modification is obtained by duplication or suppression of time frames for the three sets of parameters, without any parameter modification.

Voice quality modification is more intricate. It is obtained by spectral tilt modification (on the periodic and aperiodic spectra), periodic-aperiodic ratio modification, and signal amplitude modification. Note that the spectral representation in WORLD allows for other types of modification, that are not used in the present demonstration (e.g. formant modifications, vocal tract apparent length modification, etc.).

3.2.3 Preparing a Piece

Playing a piece on T-Voks requires three files to be prepared in advance: an audio sample, an analysis file, and a labeling data file.

The audio sample is the data that will be re-sequenced. It must be strictly monophonic, with a clearly defined pitch. Best results are obtained by ensuring that the target text is clearly spoken or sung in monotone, with a pitch close to the range that will be used during performance. Modifications of the recorded signal prior to the performance are possible, as long as their result complies with the conditions mentioned above. Resampling with a factor of about 1.2, for instance, allows for a "gender swap" effect while leaving the audio valid for use in T-Voks. Prior to use in T-Voks, the chosen audio file must be analyzed using the WORLD analysis module and stored for T-Voks to read during performance.

The labeling data file is a text file specifying the locations of all control points mentioned in section 3.1 As of now, the labelling process is done manually though a semi-automated procedure based on automatic detection of the phonemes in audio could be envisioned.

3.2.4 Synthesis during Performance

At the beginning of a performance, the analysis file, as well as the file containing a syllabic labelling of the audio, are loaded. Along with the control data received from the augmented theremin, their contents will drive the synthesis, which comprises three consecutive operations. Figure 3 shows the evolution of the control parameters when playing the verse "*Je vois la vie en rose*," as well as a spectrogram of the resulting audio.

First, a time index is received. That index, which we call the *analysis time*, specifies the temporal position in the original audio on which the synthesis should be based. An analysis time is associated with each instant in the time of the performance, or *synthesis time*. Once the current value of the analysis time is known, the WORLD representation at the corresponding time frame is queried in the loaded analysis file. That representation, while already a valid input to the WORLD synthesis module, then undergoes a few modifications to account for the changing values of vocal effort that are being received.

The vocal effort is varied by modifying the intermediate WORLD representation right before it is used in the synthesis. Among those modifications, only two affect the periodic part of the voice. One is a variation of the ratio between the first and higher harmonics, to simulate a change in the frequency of the glottal formant. The other is the use of a spectral slope. One last modification involves both the periodic and the aperiodic part, since it consists of varying the relative prominence of those parts: increasing the prominence of aperiodicity results in a breathier voice, thus reducing the perceived vocal effort.

3.3 Hardware and Computer Interface

The hardware of T-Voks is based on a Moog Etherwave Plus theremin, which features control voltage (CV) outputs for pitch and volume. The pitch CV ranges from -2.5v to 4.5v, with a change of 1 volt per octave of the theremin's pitch. The volume CV from our theremin measured from 0 to 5.5v. Each CV output is fed into an analog input pin of an Arduino Uno. For optimal function, the CV ranges should be regulated to the 0 to 5 volt range of the Arduino. A force sensitive resistor (FSR) is used to control the sequencing of syllables. Its data uses another analog input of the Arduino. Our FSR is attached to a signal conditioning circuit from Interface-Z[2], whose output vary from 0 to 5 volts between maximum pressure and no pressure.

The Arduino is connected to a computer running the T-Voks Max patch, and is programmed to convert the pitch, volume, and FSR analogue data to digital values ranging from 0 to 1024. The Max patch receives data from Arduino using the Arduino2Max object, communicating over Serial with a baud rate of 115200. In the T-Voks patch, prepared pieces are loaded by setting the proper audio, analysis, and syllable labeling files, with shortcut buttons to load files for several example pieces. Pressing the button also serves to restart a piece.

Pitch data from the theremin modulates the fundamental frequency of a loaded piece at its current time index while volume data modulates the vocal effort. Although the output of the FSR belongs to a continuous range of values, it is only used as a binary switch, allowing the player to choose between two states, pressed and released. Each time the FSR changes from one state to another, a time index starts increasing until it reaches the next syllabic control point. Pressing the FSR triggers a transition from a syllabic nucleus of the original recording to a syllabic liaison, whereas releasing it triggers the inverse transition, loosely mimicking the movements of a jaw during speech, as described in the frame/content model of speech production.[24]

[2]https://www.interface-z.fr/pronfiture/contact/148-pression-force-fsr.html

4. MAKING MUSIC WITH T-VOKS

4.1 Theremin Playing Techniques

Theremin performance is a highly individual art, with no single, standardized technique, though some educational resources do exist [6]. Despite widely held views on the theremin's difficulty, many players (including our thereminist, the first author, who is self-taught) manage to accurately play melodies within several months of practice.

For most right-handed players, the right hand controls pitch and the left hand controls volume. While only the nearest point between the body and each antenna directly modifies the output tone, the rest of the player's body also influences the response of the antennas' capacitive fields. To reliably find notes and intervals, some thereminists have developed hand and finger gestures, such as opening and closing the hand toward the antenna for the span of an octave. Shaking the pitch-control hand produces a vibrato.

Unlike most traditional instruments, which must be actuated to produce a sound, the theremin outputs a tone by default and must be explicitly silenced. Raising and lowering the volume-control hand is used to cleanly delineate notes by removing unwanted glissandos in between. The quality of these movements sculpts the attack, duration and decay of each note, enabling a wide range of articulations, though legato across large intervals and sharp staccatos are difficult to achieve. Larger movements of hand, wrist and arm defines the dynamics across phrases.

4.1.1 Controlling Syllables

Syllable sequencing is controlled by the hand responsible for volume modulations (for our thereminist, the left hand). The movement to press the sensor must be comfortable enough to perform repeatedly, reliable enough for the system to detect, and fast enough to articulate several syllables in rapid succession. Moreover, it should not interfere with other gestures of the same hand, wrist, and arm for articulation and dynamics.

After experimenting with several different sensor placements, consistently satisfactory results were found with the FSR positioned against the first knuckle of the index finger, held in place by a ring and pressed by the thumb. Its cable runs down the palm, along the arm, around the shoulder and out behind the player. It is secured by an elastic band around the palm and easily hidden by a long-sleeved shirt. Only the thumb and forefinger of the volume hand are involved in syllabic control, leaving free range of motion for the rest of the hands and fingers, as well as the wrist, forearm, and elbow.

The addition of syllable control alters the volume hand's techniques. A syllable change at the same time as a note change hides the usual glissando to the new note, and removes the need for the volume hand to dip between the notes. Syllables also add more attack and textural variation, liberating the volume hand to focus on phrase-level dynamics rather than note-level articulation. While the pitch hand is not involved in syllabic control, the addition of articulated syllables has inspired new pitch-control gestures, as described in the next section.

4.2 Musical and Poetic examples

To showcase the versatility of T-Voks, four demonstration pieces were created, practiced and recorded. Each features a different language and musical or poetic style. Here we describe how each example was created as well as playing techniques specific to each style. These examples are demonstrated in the accompanying video (see footnote 1).

La Vie en Rose. To recreate the famous chorus of the 1940s French song by Edith Piaf, we recorded a vocal sample from a female speaker. This speaker has no musical background, showing the capability of T-Voks to work with untrained voices.

As French is a syllable-timed language [7], articulation of each new syllable is done with a quick tap to the FSR. The tap release begins the syllable rhyme, which is held by Voks until the next syllable begins. During sustained vowels, playing technique is no different from the unmodified theremin, which can replicate key features of Piaf's signature vocal style, including dramatic vibratos and small glissandos at the start and end of phrases.

My Funny Valentine For our version of the jazz standard made popular by Chet Baker in the 1950s, a vocal sample from a female American English speaker was resampled by a factor of 1.22 to yield a male voice. The modulation of vocal effort along with volume change lends a "breathiness" to the synthesized voice, inspired by Chet Baker's singing.

Unlike French, English is a stress-timed language [7]. Syllable control requires paying more attention to stress timing and inter-syllable transitions. One example is when to repress the FSR after a vowel to trigger a consonant word ending. For phrases ending in a consonant, volume fades must be carefully controlled to not reach silence in order to hear the final articulation (e.g. "...favorite work of *art*").

Pierrot Lunaire - Der Kranke Mond *Sprechstimme* is a vocal technique where singing imitates the continuous pitch contours of speech. A classic example is Arnold Schoenberg's *Pierrot Lunaire* suite, where a narrator, typically a soprano, recites poetry in German accompanied by a small instrumental ensemble. Schoenberg indicated that notated rhythms should be closely followed while notated pitches, once found, should be quickly altered by rising or falling sweeps.

An excerpt from one movement of *Pierrot Lunaire* was recorded by a male speaker, which was transformed into a female voice by a resampling factor of 0.8. As a stress-timed language, German shares the same considerations for syllable advancement as English. For a convincing *sprechstimme*, pitch slides and their volume curve must also correspond to the correct stress pattern. Pitch slides are achieved by small displacements of the fingers or by pivoting around the wrist.

Chun Xiao Mandarin Chinese is a tonal language, where the same syllable pronounced with different frequency contours changes in meaning. Each syllable can be pronounced with one of four tones, which is carried by the syllabic rhyme [20]. Classical poetry is typically recited with exaggerated tone enunciations.

A well-known Tang dynasty short poem was recorded by a native speaker pronouncing each syllable in monotone. The poem was then "recited" using T-Voks, with each tone shaped entirely by the theremin. Each syllable was triggered by a quick tap and release of the FSR.

Tones were created mostly with the pitch-hand, with the volume hand creating a gradual fade in and fade out. The pitch hand rests in place for tone 1, whose pitch stays steady. Other tones, whose pitch change in different ways, were produced using fluid wave-like gestures of the entire hand, pivoting at the wrist. These hand sweeps are larger than those required by *Pierrot Lunaire*, with the forearm remaining largely stationary.

4.3 Practice Methods

The addition of syllable advancement introduces a significant cognitive load to an instrument that already requires full concentration for playing. In early stages of using T-Voks, the thereminist invented exercises for syllable advancement to get used to the new task. These exercises

involved triggering new syllables at different rhythms while performing simple pitch changes with the other hand (e.g. scales and arpeggios).

For each musical example, the thereminist would isolate the difficulties of syllabic control and pitch control, practicing only the correct rhythm of syllable advancement, or only the notes and their phrasing on any sustained vowel.

From the early stages of learning a melody, the thereminist played along with example recordings to help stay in tune. Playing along with recordings and attempting to imitate the singer as closely as possible also helps to inform interpretations. At later stages of learning a song, the thereminist would alternate between playing along with a singer and playing with only an instrumental accompaniment track in order to find her own expression. To practice the Chinese poem, the thereminist (a Chinese speaker) alternated between vocal pronunciation and T-Voks replication in order to find gestures that replicate the tonal contours of her actual voice.

5. CONCLUSIONS

The intrinsic vocal quality of the theremin is particularly well suited to singing synthesis control. This work presents, to the best of our knowledge, the first singing theremin, i.e. the first encounter between performative singing synthesis and the (augmented) theremin. Expressive, accurate, and precise singing is obtained, when the instrument is played by a well-trained theremin performer.

It is interesting to compare singing synthesis using the augmented Theremin (T-Voks) and singing synthesis using a graphic tablet and a stylus (C-Voks). The synthesis engine and control principles are similar, but control interfaces are different, resulting in different types of expressive gestures and different musical styles.

Natural sounding singing is obtained by re-sequencing of recorded utterances, and therefore to the expense of freedom in terms of linguistic content. The performer is able to play any score (i.e. any rhythm and tones, but only with the fixed text loaded in the synthesis engine). Future work will address the question of free text singing.

6. ACKNOWLEDGMENTS

Part of this work has been done in the framework of the SMAC (FEDER IF0011085) project. Xiao Xiao was partially supported by the MIT-France program during her stay at Sorbonne Université.

7. REFERENCES

[1] Carolina Eyck. `https://www.carolinaeyck.com/`.

[2] Coralie Ehinger. `https://coralieehinger.ch`.

[3] Dorit Chrysler. `http://www.doritchrysler.com/`.

[4] Lyn Goeringer. `http://www.lyngoeringer.com/portfolio/?page_id=57`.

[5] Rob Schwimmer. `http://www.robschwimmer.com/`.

[6] Theremin world: Learn to play the theremin.

[7] D. Abercrombie. *Elements of General Phonetics*. Edinburgh University Press, 1984.

[8] M. Astrinaki. *Performative statistical parametric speech synthesis applied to interactive designs*. PhD thesis, University of Mons, 2014.

[9] M. Blasco. The theremin orchestra. `http://half-half.es/the-theremin-orchestra`.

[10] C. d'Alessandro, A. Rilliard, and S. Le Beux. Chironomic stylization of intonation. *JASA*, 129(3):1594–1604, 2011.

[11] N. d'Alessandro and T. Dutoit. Handsketch bi-manual controller: Investigation on expressive control issues of an augmented tablet. In *Proc. NIME07*, pages 78–81. NIME, 2007.

[12] S. Delalez and C. d'Alessandro. Adjusting the frame: Biphasic performative control of speech rhythm. In *Proc. INTERSPEECH 2017*, pages 864–868, 2017.

[13] S. Delalez and C. d'Alessandro. Vokinesis : syllabic control points for performative singing synthesis. In *Proc. NIME 2017*, pages 198–203, 2017.

[14] G. Fant. *Acoustic theory of speech production*. Mouton, 1970.

[15] S. S. Fels and G. E. Hinton. Glove-talk: A neural network interface between a data-glove and a speech synthesizer. *Neural Networks, IEEE Trans. on*, 4(1):2–8, 1993.

[16] S. S. Fels and G. E. Hinton. Glove-talkii-a neural-network interface which maps gestures to parallel formant speech synthesizer controls. *IEEE Trans.on Neural Networks*, 9(1):205–212, Jan 1998.

[17] L. Feugère, C. d'Alessandro, B. Doval, and O. Perrotin. Cantor digitalis: chironomic parametric synthesis of singing. *J. Audi. Speech Mus. Proc.*, 2017.

[18] R. Gibson. The theremin textural expander. In *Proc. NIME18*, pages 51–52. ACM, 2018.

[19] A. Glinsky. *Theremin: Ether Music and Espionage*. University of Illinois Press, 2005.

[20] P. Hallé. Evidence for tone-specific activity of the sternohyoid muscle in modern standard chinese. *Language and Speech*, 73:103–124–1043, 1994.

[21] H. Kenmochi and H. Ohshita. Vocaloid-commercial singing synthesizer based on sample concatenation. In *INTERSPEECH07*, pages 4009–4010.

[22] S. Le Beux, C. d'Alessandro, A. Rilliard, and B. Doval. Calliphony: A system for real-time gestural modification of intonation and rhythm. In *Speech Prosody*, 2010.

[23] A. Lough, M. Micchelli, and M. Kimura. Gestural envelopes: Aesthetic considerations for mapping physical gestures using wireless motion sensors. In *Proc. ICMC07*, pages 60–64. ICMC, 2018.

[24] P. F. MacNeilage. The frame/content theory of evolution of speech production. *Behavioral and Brain Sciences*, 21(4):499–511, 1998.

[25] M. Morise. Cheaptrick, a spectral envelope estimator for high-quality speech synthesis. *Speech Communication*, 67:1 – 7, 2015.

[26] M. Morise. D4c, a band-aperiodicity estimator for high-quality speech synthesis. *Speech Communication*, 84:57 – 65, 2016.

[27] M. Morise. Harvest: A high-performance fundamental frequency estimator from speech signals. In *Proc. Interspeech 2017*, pages 2321–2325, 2017.

[28] M. Morise, F. Yokomori, and K. Ozawa. World: A vocoder-based high-quality speech synthesis system for real-time applications. *IEICE Trans. on Information and Systems*, E99.D(7):1877–1884, 2016.

[29] O. Perrotin and C. d'Alessandro. Vocal effort modification for singing synthesis. In *Proc. INTERSPEECH 2016*, pages 1235–1239, 2016.