Voks: Digital instruments for chironomic control of voice samples

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

Voks: digital instruments for chironomic control of voice samples

Grégoire Locqueville[a], Christophe d'Alessandro[a], Samuel Delalez[b], Boris Doval[a], Xiao Xiao[a]

[a]*Institut Jean le Rond d'Alembert, Sorbonne Université, CNRS, UMR 7190, 4 place Jussieu, 75015, Paris, France*
[b]*Lunii,18 rue Dubrunfaut, 75012 Paris, France*

Abstract

This paper presents Voks, a new family of digital instruments that allow for real-time control and modification of pre-recorded voice signal samples. An instrument based on Voks is made of Voks itself, the synthesis software and a given set of chironomic (hand-driven) interfaces. Rhythm can be accurately controlled thanks to a new methodology, based on syllabic control points. Timing can also be controlled with other methods, including scrubbing and playback speed variation. Pitch, vocal effort, voice tension, apparent vocal tract size, voicing ratio, aperiodicity ratio of the voice samples can be modified thanks to a real-time high-quality vocoder. Different forms of chironomic control of the vocal parameters are proposed. Pitch is controlled by continuous hand motions using a stylus on a surface (C-Voks) or a theremin (T-Voks). Other interfaces can be used as well. Syllabic rhythm is controlled using a biphasic button. Scrubbing, playback speed and timbre related parameters can be controlled using the theremin, control surfaces or continuous controllers like faders. In addition to realistic imitation of speaking or singing voices, other playing modes yield new interesting sounds. Voks participated in comparative perceptual evaluation of singing synthesis systems. It has been demonstrated in a live musical settings, using different control interfaces. In addition to musical or poetic performances, applications of performative vocal synthesis to language learning and speech reeducation are foreseen.

Keywords: voice synthesis, singing synthesis, new interfaces for musical expression, performative synthesis, real-time vocoder

1. Introduction

1.1. Performative vocal synthesis

Performative vocal synthesis, or voice instruments, are the meeting point of voice synthesis and new interfaces for musical expression[1]. While the synthesis of acoustic musical instruments such as the pipe organ, the piano, strings and winds have reached high levels of realism, the speaking and singing voices remain a challenge for digital sound synthesis and new musical instruments research. Performative vocal synthesis is an important issue for voice research, with applications in the fields of music (new instruments, studio), language education, speech therapy (control of voice source substitution).

A fundamental difference between vocal synthesis and musical instrument synthesis is the additional linguistic content. The speaker or singer must in real time perform a musical task (pitch, rhythm and timing, voice force and quality) and a linguistic task (a stream of phonemes, syllables, words, sentences). Performative vocal synthesis requires two main types of processes: 1. selection and planning of the linguistic and musical material to be sung and 2. use of an external control device for sound synthesis to mimic the motions of the internal voice apparatus.

Voks is a new paradigm in the already long history of artificial voices. The voice is played like an instrument, allowing for singing or speaking with the borrowed voice of another [4] with realism, expressivity and musicality. In the Voks system, linguistic material is prepared in advance, by speech recording

[1]Part of this work has been presented at NIME 2017 [1], Interspeech 2017 [2] and NIME 2019 [3, 4]

and labelling. Any utterance is composed of voice signal samples (ranging from a single syllable to entire sentences) enriched with syllabic marks that allow for accurate rhythmic control. A Voks set, for a given performance, is a set of linguistic utterances, that can be selected and played on the fly. The synthesis is achieved by real-time control of a high-quality vocoder. The vocoder allows for real-time fundamental frequency (f0) scaling, time scaling, vocal effort and voice quality modifications. These modifications are driven by the player's gestures, using various chironomic (hand controlled) interfaces. The first performative speech synthesis system, allowing for synthesis of any text, was Glove Talk [5, 6]. It converts hand gestures to speech, based on a neural network gesture-to-formant model. The gesture vocabulary is based on a correspondence between hand shapes and articulators positions. Synthesis is based on a formant synthesizer. 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" [6]. Though it enables the real-time performative synthesis of speech, neither version of Glove Talk can be considered a singing synthesis system, as melodic and rhymes control was highly limited. The formant model has been used in several singing synthesis systems. This approach has the advantage of granting the user total control over the generated sound. A relatively small number of parameters drives a synthesis algorithm [7, 8, 9, 10, 11, 12].

Another 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 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 singing synthesizers [13, 14, 15, 16]. A performative text-to-speech synthesis environment based on Hidden Markov Models synthesis has been presented [17, 18]. In this case, the linguistic materiel is entered in text form on a computer keyboard or from a text file. More recently, several neural network based synthesis systems appeared, with application to singing synthesis [19].

Several control interfaces for melodic control have been proposed. The discrete nature of the traditional piano-like, MIDI keyboard, makes it ill-suited to the control of voice [20]. The graphic tablet has been used in singing synthesis systems [10, 11, 9, 12]. It has the advantage of reusing the expert gestures of writing, allowing for similar or even higher precision for singing pitch control [21]. Other interfaces sending continuous data streams can be used, particularly those using the MIDI (or Multidimentional) Polyphonic Expression protocol [22], including the Continuum [23] and the Seaboard [24].

1.2. Chironomic control of voice samples

As it does not seem possible to control all the aspects of voice production through only hand (or feet) gestures, chironomic control of voice samples is a good compromise between modification capabilities and sound quality [25]. The present work focuses on accurate melodic and rhythmic control. Melodic control is based on earlier work on chironomic control [12, 21]. The paradigm of syllabic control points introduced in the Vokinesis system [1, 2] offers accurate control over voice rhythm and timing. Voice quality and vocal effort are processed using a high quality vocoder [26]. Voks architecture is presented in Figure1.

This architecture contains three main blocks: data preparation (left, green frame), chironomic control (top, green frame) and real-time processing (center, blue frame). Voks is based on WORLD [26], a powerful vocoder, whose underlying signal model is described in section 3.1. WORLD allows for analysis of a speech signal in a spectral-domain representation for further processing and synthesis (see section 3.3). Prior to the performance, an audio sample, represented in the left green panel in Figure1, is recorded. This linguistic material is then labelled, resulting in a text file containing a representation of the syllable locations. Details about the labelling method and the resulting *label file* can be found in section 2.3.

The top, green panel in Figure1 represents the user off-line and real-time controls. As a musical instrument, Voks takes real-time data as input. Some of those data are acquired via gestural interfaces; additionally, some other values can be set using a graphical interface. The acquisition of real-time data, and their interplay with the rest of the system, are discussed in section 5.2.

The center, blue panel in Figure1 represents data processing and audio synthesis. At each instant of the performance, the real-time data and the data contained in the labelling file are converted, using *time*

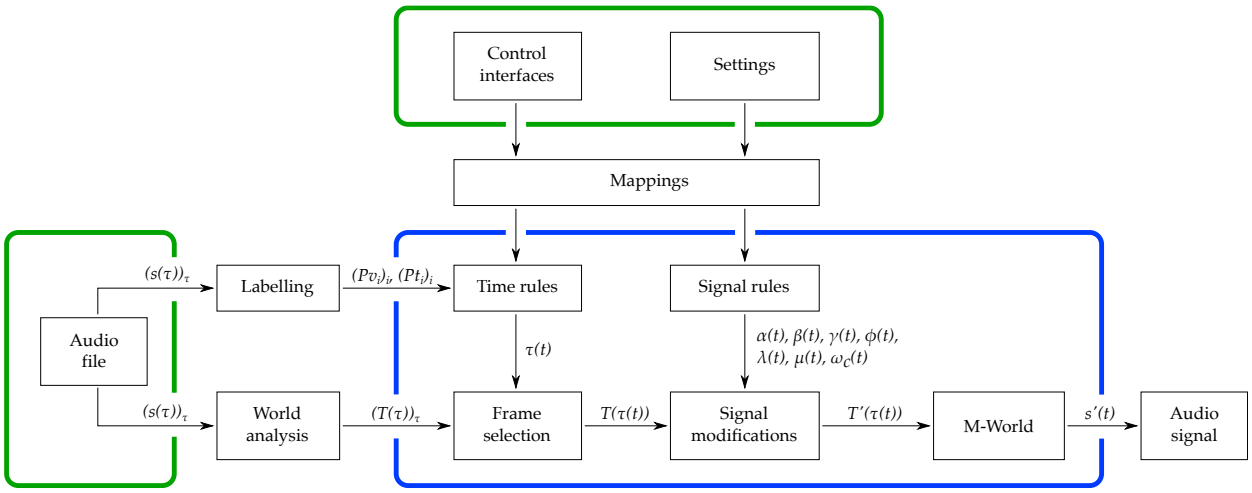

Figure 1: General architecture of Voks. The green boxes enclose input data; the blue box encloses the real-time software processing that takes place at the time of the performance. The meaning of the time index τ is explained in section 2; that of other Greek letters is explained in section 4.

rules, to a *time index* that indicates the position in the original audio upon which to base the synthesis. This method for temporal control, detailed in section 4.1, allows the extraction of a *WORLD frame*, the WORLD representation of the original audio at one specific instant. This frame then undergoes some or all of the modifications described in sections 4.2, 4.3 and 4.4. Those modifications are driven by parameters that are functions of the real-time inputs; the rules used to turn input data into parameters for modification are described in section 4. The resulting representation is then fed to the synthesis module mentioned in section 3.3.

Note that Voks is based on the same principles as the earlier Calliphony and Vokinesis systems [15, 1, 2]. The architecture in Figure1 represents Calliphony and Vokinesis as well. The main difference between these systems is the vocoder used, RT-PSOLA [27] for Calliphony and Vokinesis, vs WORLD for Voks. As WORLD offers an explicit spectral decomposition of the speech signal, it has been preferred to RT-PSOLA: more flexibility in signal processing and transformations is allowed. For pitch and rhythm control the quality and functions of Voks and Vokinesis are equivalent.

2. Sample preparation and labelling

Voks is based on rhythmic sequencing, pitch and voice quality modifications of pre-recorded speech samples. The principles for vocal rhythm control are presented, and a new method using syllabic control points is proposed in this section.

2.1. Speech material

To use Voks, the first step is to record a suitable speech sample. The suitable speech samples are monophonic recordings. Note that the vocoder may show some limitation for processing non-standard vocal techniques, such as growling, that would disrupt the harmonicity of the recording too much. Excessive reverberation may also jeopardize the results. Other monophonic, harmonic sounds (such as many auto-oscillating instruments, bird songs, etc.) may give satisfactory results. Clearly articulated syllables are necessary for accurate syllabic rhythm control, but are not required for non-syllabic playing modes (e.g. *scrub* and *speed*).

The recorded samples need to be *prepared* before being of any use to the performer, this involves two different steps:

- *syllabic labelling* for marking of the signal with rhythmic anchors or control points.

- *signal analysis* for transformation of the time-domain samples to a source-filter representation suitable for signal modification in the spectral domain and pitch scaling.

The Voks system is capable of using samples provided by a Text-to-Speech system, which can deliver both the signal samples and syllabic labels at the same time. The sound quality is often poorer than natural speech. Therefore recording natural voice is generally preferred for high quality musical performances.

2.2. Rhythmic model

Speech and singing rhythm is defined by the position in time of consonants and vowels. Following [28], the syllable can be regarded as the minimal suprasegmental unit for rhythm control. According to Wagner ([29], p. 41-53) the syllable can consistently be associated to rhythmic beats in speech and music. However, while some languages (e.g. French, Chinese) have syllable-timed isochrony, others have different ones. Mora-timed isochrony (e.g. Japanese) is based on units smaller than the syllable, and stress-timed isochrony (e.g. English) is based on units larger than the syllable. In the present work, an utterance is organized into a syllable stream. Uttering a syllable creates a musical note, a rhythmic unit. Of course one syllable can carry several musical notes (this is called a melismatic syllable, or melism), or a same note can extend over several syllables. Accurate control of syllable positions and timing is required for controlling the vocal rhythm. A simple idea would be to use only one control point for each syllable, for instance the perceptual centers (P-centers) of the syllable [29, 30]. P-centers are generally defined by tapping experiments, i.e. by synchronisation of a manual gesture with the perceived position of the syllable. This would correspond to hand-clapping or foot tapping, which are common ways of marking or following the perceived rhythm of music. P-centers are a *perception* concept, but for rhythm *production* it seems that only one point per syllable is not sufficient.

Another approach draws from the frame-content theory of speech production [28], where syllables are related to oscillations of the mandible. In this theory, speech production, i.e. planning and realization of the articulatory motion for a given stream of phonemes and prosody, is considered to be the superposition of two coordinated processes. The relatively fast succession of segmental articulatory events defines the phonemic *content*, that is carried by the syllabic *frame*. The frame corresponds to the stream of syllables, each syllable being a biphasic cycle of opening/closing of the jaw. For each syllable, two points are needed, one corresponding to the opening, and the other to the closure of the mandible. Note that this is somewhat analogous to keyboard playing. When playing the pipe-organ for instance, a first motion is to depress the key, and a second motion, as important as the first, is to release the key. The duration of a note is a result of both "control points".

Syllabic rhythm depends on the syllable structure. The syllable is composed of three parts: the attack, the vocalic nucleus, and the coda. The attack and the coda correspond to zero, one or more consonants, and the nucleus generally corresponds to a vowel. A syllable always contains a vocalic nucleus, but not necessarily an attack and a coda. In an actual voice utterance, syllables are chained; attacks and codas of successive syllables correspond to consonants, i.e. the opening and closure motions of the vocal apparatus, while vowels correspond to the open positions. These opening and closing cycles can be exploited for rhythmic control. The concepts of *arsis* and *thesis* (derived from Greek prosody) are then very useful for our purpose. *Thesis* represents the stable part of the segment, in our case the vowel, or nucleus, and *arsis* represents the transient part between nuclei. The coda of one syllable and the attack of the next one (when they exist) are grouped to form the arsis. If there is neither a coda nor an attack, the arsis still exists, and corresponds to a short transition between two vowels. Controlling syllabic rhythm implies controlling those time points.

We define the *Syllabic Control Points* (SCP) as temporal marking points for rhythm production. *Vocalic Points* (P_v) are the SCP that correspond to the vocalic nuclei or thesis, and *Transient Points* (P_t), those that correspond to the transient phases, or arsis. These points define a target temporal location for each phase: when a vocalic phase is triggered (see section 4.1.1), the target timestamp is at the corresponding P_v until the next transient phase is triggered. Once this transient phase is triggered, the target timestamp evolves from the current P_v to the next P_t, and the synthesis signal stops at that P_t until the next vocalic phase is triggered, and so on. Controlling the timing of these points allows for accurate rhythmic control while

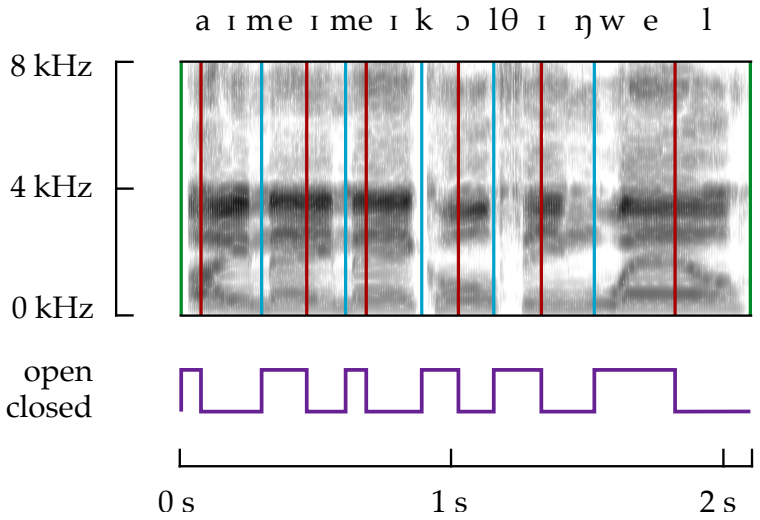

Figure 2: Syllabic control points for the sentence *"I may make all thing well"* (/aɪ meɪ meɪk ɔːl θɪŋ wel/). Nucleic control points are marked with red lines, transient control points with cyan lines, superimposed on the spectrogram of the audio sample. Green lines indicate the starting and ending points. The purple graph indicates which portions of the signal will be played when the controller is respectively open and closed.

preserving the correct articulation. Although the syllable is the main unit discussed herein, the SCP may in principle be used for mora-timed or stress-timed isochrony as well. In this case P_v and P_t would be associated to time spans smaller of larger than the syllable, corresponding to morae, feet of stress-groups.

2.3. Syllabic Control Points

The SCP allows for accurate rhythmic control over the unwinding of syllables. Such control requires a *labelling* representing the temporal arrangement of syllables in the sample, in the form of a text file containing a series of timestamps corresponding to the *syllabic control points*. Placement and positions of the two SCP per syllable P_v and P_t are displayed in Figure2. Each syllable begins with a P_t, followed by a P_v in the syllabic nucleus. P_v are placed near the center of the vowel corresponding to the vocalic nucleus to ensure a correct pronunciation. The P_t are placed at a stable or silent zone in the cluster of consonants between two vowels. This guarantees an accurate control of the instant of occurrence of the next vocalic phase. When the consonant cluster ends with an unvoiced plosive, the corresponding P_t must be placed during the silence prior the explosion. Special treatments for the phrase initial and final P_t are needed, in order to ensure the pronunciation of every phoneme entirely. The phrase initial P_t have to be placed at the end of the silence prior to the first phoneme, and the phrase final P_t must be placed at the beginning of the silence following the last phoneme.

SCPs have to be set beforehand. It can be done manually, identifying the locations of the SCP with the help of an audio editor and reporting them in a text file. This task can be automated, or semi-automated, as has been done in Vokinesis [31], with the help of automatic speech recognition and segmentation algorithms, and simple rules for locating control points based on that segmentation.

3. Signal representation and Vocoder

Labelled speech samples need to be processed in real-time. This is achieved by a real-time voice coder (vocoder) system based on the source-filter voice signal decomposition. In Vokinesis [2, 1] a real-time PSOLA vocoder was used [27]. Other real-time high-quality vocoder systems can be used as well. In Voks the WORLD vocoder has been chosen and modified, because of its high sound quality and ability to perform various spectral modification in real-time. The following subsections focus on the WORLD vocoder.

3.1. Source-filter model

WORLD is based on the linear source-filter model of speech production [32].

$$\begin{aligned} s(t) &= (\text{Ш}_{T_0} g(t) + n(t)) * v(t) * l(t) \quad (1) \\ &= (\text{Ш}_{T_0} g(t) + n(t)) * c(t) \quad (2) \end{aligned}$$

where $*$ is the convolution operator, and $\text{Ш}_{T_0} = \sum_n \delta(t - nT_0)$

The source filter model is made of a source component, corresponding to phonation, and a filter component corresponding to the vocal tract and lip radiation. The source component is modelled as the sum of $\text{Ш}_{T_0} g(t)$ a periodic (harmonic) component and n an aperiodic (noise) component. g represents a glottal pulse. The aperiodic component accounts for the sound corresponding to the fast motion of articulatory organs during consonants (transient noise), but also for the turbulent airflow (breath noise, aspiration or frication noise) that accompanies periodic vibrations of the vocal folds during voiced sounds or fricative consonants. The evolution of both the periodic and aperiodic components are assumed to be slow compared to the fundamental periods. $v(t)$ is a time-varying linear filter accounting for the vocal tract action on the source signal. $l(t)$ represents the effect of lip radiation. v and l can be associated in a same filter c.

WORLD is based on a slightly different formulation of the source filter model. The glottal source g is regarded as a filter, and associated to c to form the periodic component filter $h_p(t)$. A second filter is defined for the aperiodic component h_{ap}. With this formulation, the harmonic component is written as a simple Dirac comb Ш_{T_0}, with period $T0$, filtered by a filter with response $h_p(t)$; similarly, the noise component is seen as white noise $n(t)$ filtered by a filter with response $h_{ap}(t)$:

$$s(t) = \text{Ш}_{T_0}(t) * h_p(t) + n(t) * h_{ap}(t) \quad (3)$$

In the spectral domain, that equation becomes

$$S(\omega_i) = \text{Ш}_{F_0}(\omega_i) H_p(\omega_i) + N(\omega_i) H_{ap}(\omega_i) \quad (4)$$

where H_p and H_{ap} are the Fourier transforms of h_p and h_{ap}. WORLD does not actually deal with H_p and H_{ap}; it uses parameters closely related to those. Those parameters are a total spectral envelope ($E(\omega_i)$) and another filter, representing the "aperiodicity ratio" $R(\omega_i)$. They are related to H_p and H_{ap} by the following relations:

$$\begin{aligned} E(\omega_i) &= \left|H_p(\omega_i)\right|^2 + \left|H_{ap}(\omega_i)\right|^2 \\ R(\omega_i) &= \frac{\left|H_{ap}(\omega_i)\right|^2}{E(\omega_i)} \end{aligned} \quad (5)$$

The aperiodicity ratio $R(\omega_i)$ indicates, in each frequency band, what fraction of the total spectral power comes from the unvoiced part of the signal. Together, those three parameters allow for resynthesis of a high-quality signal [33].

3.2. WORLD analysis

WORLD is a collection of C language source programs, distributed under the permissive 3-clause BSD license. The original WORLD software [26] is organized in two main parts. The analysis part can be used to estimate WORLD parameters for a given audio file. The synthesis part computes the voice signal corresponding to the WORLD parameters. It is possible to modify the voice parameters between analysis and synthesis, by manipulation of the intermediate WORLD parameters.

WORLD follows and improves the principles of STRAIGHT/TANDEM, a successful high-quality vocoder based on short-term Fourier analysis/synthesis of speech [34]. The main features of these vocoders, in Equation 3, is that all the spectral information is

incorporated in the filter component (spectral envelope) of the model. The excitation component spectral envelope is flat, either a white noise or a flat pulse train. The filter spectral envelope is as smooth as possible.

The WORLD analysis module takes as input audio sample $s(\tau)_{\tau\in[0..T]}$, where T is the duration of the audio, and $s(\tau)$ is the audio sample at time τ. Assuming a source-filter model described in section 3.1, parameters of the model $F(\tau_i)$ are estimated at evenly spaced time points or *frame* $(\tau_i)_{i\in[0..N-1]}$ (where N is the total number of analysis points).

The entire set of parameters $(F(\tau_i))_{i\in[0..N-1]}$ for a given recording is called the *WORLD parameters*, whereas for a given time τ, $F(\tau)$ will be called the *WORLD frame* corresponding to τ. For each frame, three types of parameters are computed, the pitch component, the spectral envelope and the periodicity ratio. At time τ, a "WORLD frame" F_τ is then the triplet $(f, E(\omega_i), R(\omega_i))_{i\in[0..N_{\text{fft}}/2]}$, where N_{fft} is the size of the discrete Fourier transform used.

The pitch component $f(\tau)$ is obtained by the HARVEST method [35]. It represents the fundamental frequency in Hertz (0 for unvoiced speech) for each frame. The spectral envelope E is obtained using short-term Fourier analysis and the CHEAPTRICK method [36]. For each frame a vector of $N_{\text{fft}}/2+1$ points representing the spectral amplitudes only is computed. The periodicity ratio R is computed using the D4C method [37]. For each frame a vector of $N_{\text{fft}}/2 + 1$ points representing aperiodicity (between 0 and 1) is computed.

With a frame rate of 5 milliseconds, a sampling rate of 48 kHz, a FFT size N_{fft} of 2048 points, and double precision real numbers, Voks parameter rate is 3.28 Mbyte/s (compared to 96 Kbyte/s for the 16 bits monophonic audio signal). WORLD analysis is an automatic and robust process that is performed off-line prior to synthesis.

3.3. Real-time synthesis: M-WORLD

The synthesis process performs the reverse operation of the analysis process: it takes a stream of WORLD frames $F(\tau)$ as input, and outputs a synthesized audio signal $s(t)$. The re-synthesized audio from WORLD parameters is not mathematically equal to the input audio, but it is almost perceptually identical to it [33]. For Voks, the WORLD analysis parameters are exported as a Jitter/Max matrix. Using the Max SDK, new software was developed to package the original WORLD C functions as a real-time Max[38] external called M-WORLD. M-WORLD takes modified WORLD frames as input. The original pitch component is replaced by real-time pitch data coming from an input device. The spectral envelope and aperiodicity ratio vectors are also modified in real time based on input data.

The synthesis process is based on overlap-adding a train of filtered glottal pulses and filtered noise, based on Equation 1. The synthetic pitch contour is converted into a point process corresponding to $\text{Ш}_{T_0}(t)$. Each time point represents the position of a filtered glottal pulse, that is computed as the impulse response $h_p(t)$. Filtered noise $n * h_a p(t)$ is computed and added to the periodic component.

4. Real-time control and signal modifications

4.1. Time, tempo, and rhythm control

This section details the implementation of rhythmic control based on principles described in section 2.

Rhythm control in Voks amounts to specifying, at a given instant t during performance time, a numeric value $\tau(t)$ (in seconds) and the *time index*, corresponding to a temporal position in the original sample. Once $\tau(t)$ has been computed, the corresponding WORLD frame $F_{\tau(t)}$ is selected, undergoes some modifications (described in section 3), and is synthesized to deliver the audio signal.

Three different rhythm control modes are available in Voks: syllabic, scrub and speed. The syllabic mode is actually a rhythmic control mode: it is akin to tapping or hand clapping synchronously with syllables to create the rhythmic patterns. The Scrub mode corresponds to the direct control of the time index, and the speed mode to controlled variation of reading tempo of the samples. These three modes of control are likely to produce very different types of musical gestures.

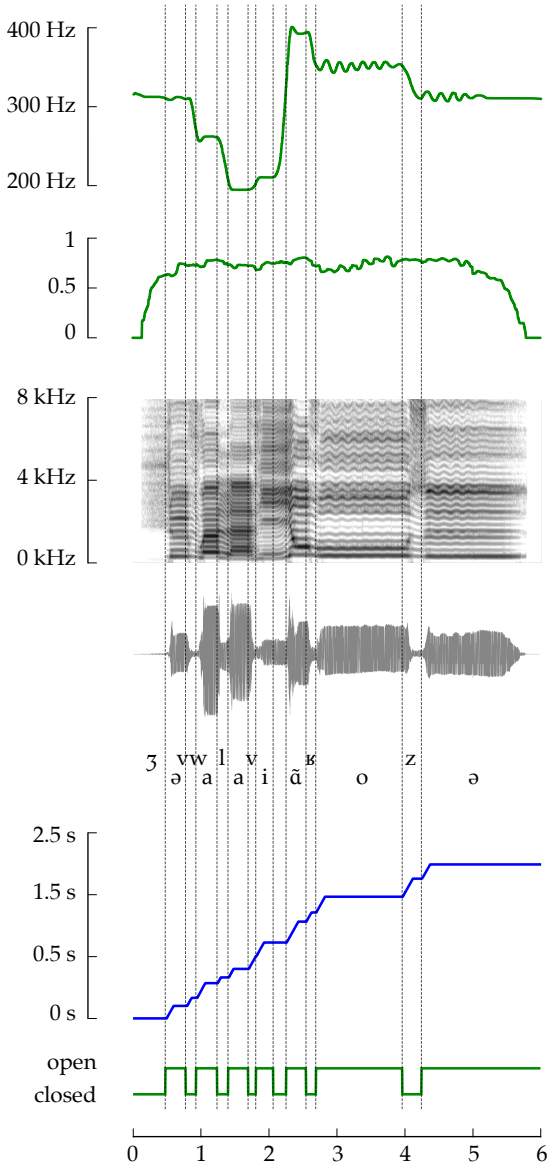

Figure 3: Performance using Voks in the syllabic control mode. French sentence "Je vois la vie en rose" (/ʒəvwalaviɑ̃ʁozə/, duration 2.5 s). Synthesis duration: 6s. Input control data in green, internal data in blue, output audio in greyscale. From top to bottom: pitch, normalized vocal effort, spectrogram, oscillogram, phonemic labels, internal time index τ (blue), binary rhythm control (green). Times at which the rhythm controller is pressed or released are marked with a vertical dotted line.

4.1.1. Syllabic control

The syllabic rhythm control mode is close to natural rhythm control in voice production. Rhythmic beats correspond to syllables. The opening and closing motions of the mandible for the natural voice corresponds to a two states button in performative control, as explained in section 2. When using syllabic rhythm control, the control parameter of the rhythm control device $\rho(t)$ can take on two values, 1 and 0. In addition to those two control states, the system itself can be in two states internally: *frozen state* and *running state*. The two internal states describe the ratio between the reading speed of input samples and synthesized samples. These states are represented in Figure3: the $\rho(t)$ parameter is the bottom green line, and the time index is the blue curve just above. The corresponding oscillogram and spectrogram, together with phoneme labels are above the blue line. The circled *2* indicates a zone where the system is in the frozen state, and the *3*, a zone where it is in the running state.

The time index $\tau(t)$ represents the actual position in time of the analysis frame used for synthesis. For this reason, $\tau(t)$ takes its values between time 0 (beginning of the utterance) and the utterance duration. When the system is in the frozen state, the time index $\tau(t)$ is constant, waiting for a change in $\rho(t)$. The synthesis parameters are "frozen," resulting in the repetition of the same spectral pattern. Note that intensity and pitch can still vary, giving life to the sound. As soon as a change in $\rho(t)$ from 0 to 1 (or from 1 to 0) is registered, the system switches to the running state. The time index $\rho(t)$ begins increasing with a constant, positive slope, called *articulation speed*, until the next vocalic (resp. transient) control point is reached. The system then switches back to the frozen state, and the time index again becomes constant; its value is now that of the control point in question. In other words, the articulation rate controls the reading speed of input sample. It is equivalent to the speed of articulation in natural speech.

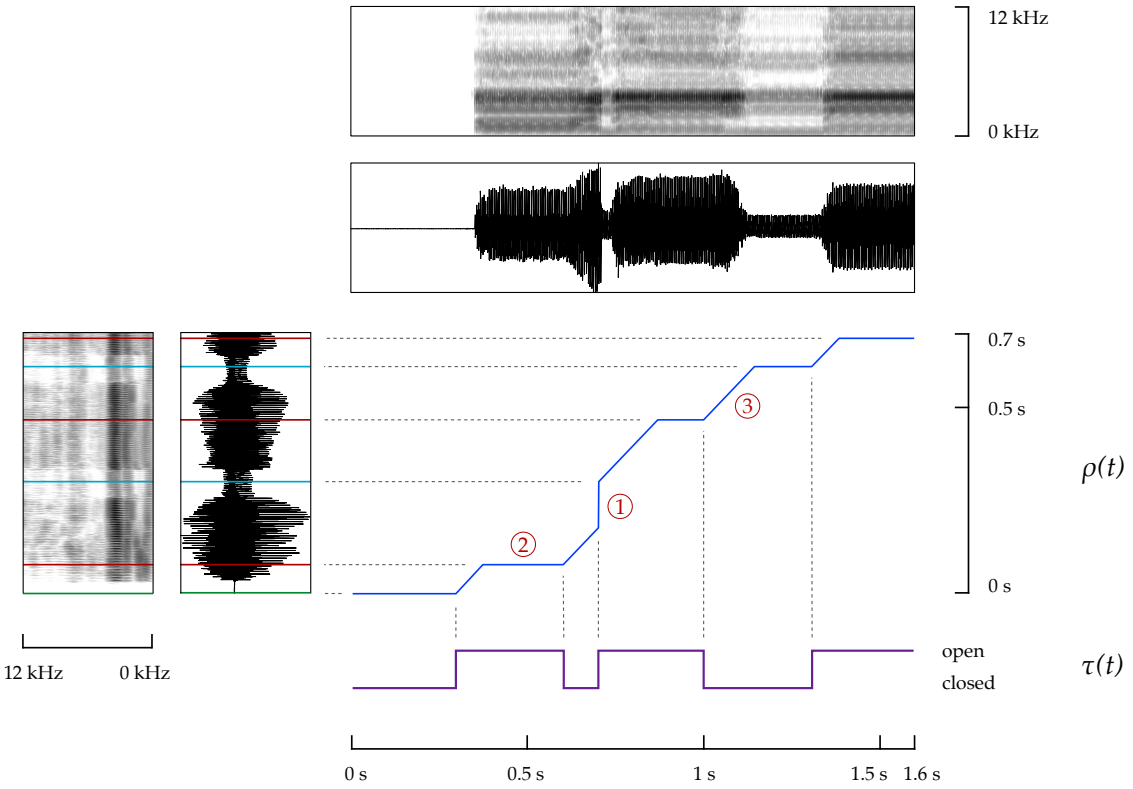

Figure 4: State of the control parameter (open or closed, in purple) and the associated temporal evolution of the time index (in blue). On the X-axis, performance time; on the Y-axis, recording time. On the left, spectrogram and waveform of the original sample, with the control points marked as on figure 2. On top, spectrogram and waveform of the generated sound.

Three possible relationships between the recorded samples and the performer's gestures timing are

possible, as illustrated in Figure4. When the performer's gesture is faster than the recorded samples, a change in $\rho(t)$ happens before the time index $\tau(t)$ has reached its target value. In this situation $\tau(t)$ jumps to the control point it was aiming for and immediately sets the next control point as its new target value. This discontinuity manifests by a vertical slope in the graph of $\tau(t)$, as can be seen in the portion of Figure4 marked with a circled 1. This discontinuity ensures that the delay of the time index relative to the performer's gesture is bounded. When the performer's gesture is slower than the recorded samples, the system enters the frozen state, in which $\tau(t)$ is constant. This manifests by a horizontal slope as seen on the portion of the signal marked with a circled 2. The third situation never happens in practice, when the recorded samples and performer's gestures are exactly with the same timing.

The articulation speed, i.e. the slope of the portions of the graph similar to the one marked with a circled 3 in Figure4, can be set using the graphical interface, or mapped to a control device. Increasing its value makes discontinuities in the time index less likely. The farther away from 1 it is, the more temporal distortion of some phonemes is noticeable, which can reduce the realism and intelligibility of the synthesized voice.

The syllabic control mode is needed when one wants to play a precise rhythm, as when singing a song or saying a text, where syllable nuclei coincide with note onsets or beats. This mode allows for accurate syllabic placement in time.

4.1.2. Scrub control mode

Scrub mode allows for replay by directly controlling the time index. Contrary to the syllabic rhythm control method, the rhythm produced has no precise anchoring in the phonetic content. The scrub mode relies on a continuous control parameter $\eta(t)$. We assume $\eta(t)$ only takes on values in the interval $[0, 1]$, O representing the beginning and 1 the end of the voice utterance ($\eta(t)$ is normalized by T, where T is the length of the pre-recorded audio). Scrub mode then consists in linearly mapping each value of the continuous parameter $\eta(t)$ to a value for the analysis time τ:

$$\tau(t) = T\eta(t) \tag{6}$$

By increasing η with a given speed, the performer can play the utterance faster or slower; by decreasing η, they can also play backwards, and by making it constant, play a steady sound. The effect of playing Voks in the scrub mode, and variation of η are displayed in the left half of Figure5.

The scrub mode is a direct *time* control mode. It is not well suited for accurate syllabic placement in time. It is rather suited for DJing-like effects and gestures, like the direct manipulation of a play-head on a turntable.

4.1.3. Speed control mode

Speed mode gives control on the sample replay speed, i.e. the *tempo*. This corresponds to the first derivative of the position of the time index τ in the recorded utterance:

$$\begin{aligned} \frac{d\tau}{dt}(t) &= \mathbb{1}_{v(t)\neq 0} g(\kappa(t)), \text{ or equivalently} \\ \tau(t) &= \int_0^t \mathbb{1}_{v(t)\neq 0} g(\kappa(u))du \mod T \end{aligned} \tag{7}$$

where:

- $\kappa(t)$ is the value of the speed control parameter.
- g is a function that maps values of κ (between 0 and 1) to speeds both negative and positive. It is continuous and increasing, antisymmetric around 0.5 (that is, $g(0.5 - \kappa) = -g(0.5 - \kappa)$, and features a strip around 0.5 where its value is zero.
- the $\mathbb{1}_{v(t)=0}$ factor denotes the function that equals 0 when the vocal effort v is zero, and 1 otherwise.

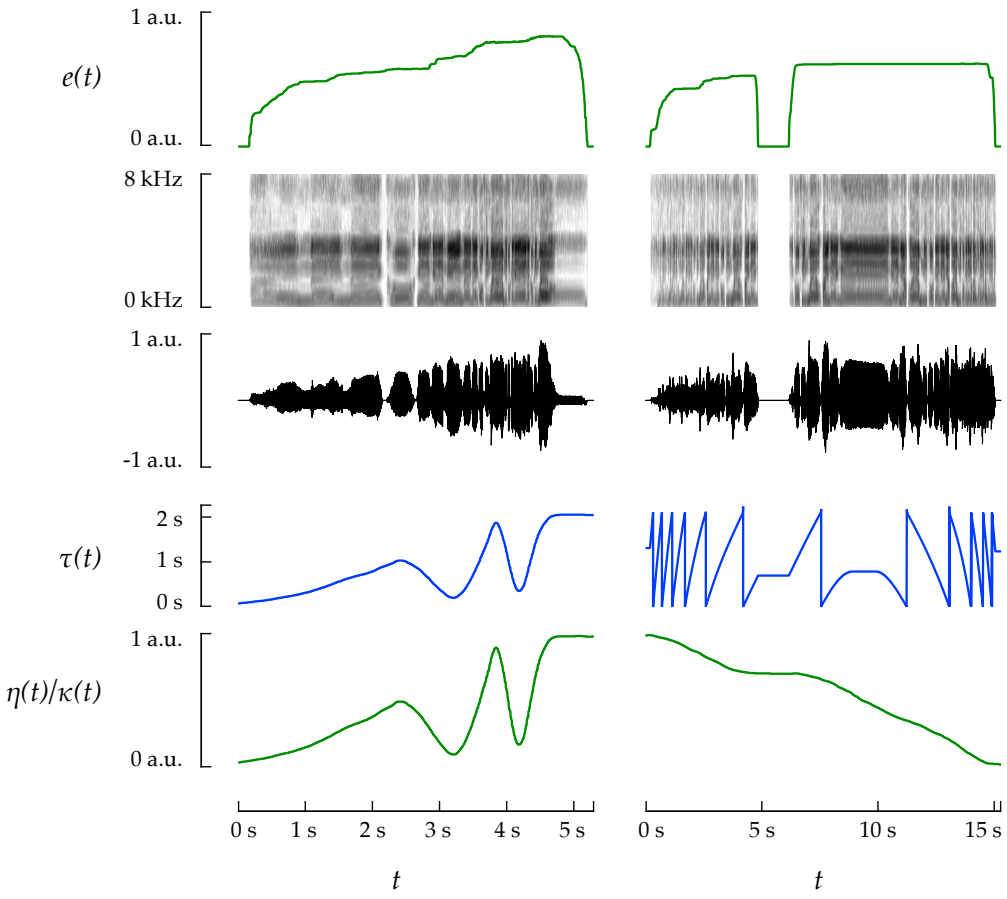

Figure 5: Short performance using Voks in scrub mode (left) and speed mode (right). Sentence "I may make all thing well" (duration 2 s). Synthesis duration: 10s (left) and 20s (right). From top to bottom: vocal effort, in arbitrary units, in green, spectrogram and waveform of the output sound, internal time index τ, in blue (see section 2), scrub/speed control parameter η (left) or κ (right), in arbitrary units, in green.

- $x \mod T$ refers to the remainder in the euclidean division by T, the length of the utterance.

The $\mathbb{1}_{v(t)=0}$ factor appears in the equation to keep the utterance from silently progressing when the vocal effort is zero. Using the remainder in the division by T makes the sample loop back to the beginning when its end is reached.

The right half of Figure5 illustrates various behaviours of speed mode: it begins with looping slower and slower with a positive speed; at around 5s, the effort drops to 0, stopping progression of the time index until it becomes strictly positive again. Around 8s, the control value $g(\kappa)$ becomes 0, stopping the progression again, before taking lower and lower negative values, which has the effect of playing the audio (reversed in time) and accelerating.

4.2. Melodic control

Pitch (melody) is controlled by the player's gesture. For synthesis, the fundamental frequency $\phi(t)$ is obtained by simple re-scaling of the gestural contour. This contour can be produced by various interfaces as detailed below. As the pitch is often expressed in the MIDI format, $\phi(t)$ in Hertz is obtained by:

$$\phi_{Hz}(t) = 440 \cdot exp\left(\frac{\phi_P(t) - 69}{12}\right)$$

with $\phi_P(t)$ the pitch in MIDI (assuming an equal temperament). The frequency $\phi(t)$ is then fed to the M-WORLD synthesis module.

4.3. Glottal source control

The spectral representation of WORLD allows for spectral modification of voice quality. Voice quality parameters are related to the time and spectral features of the voice source [39]. The main voice quality parameters are vocal effort, vocal tension and noise in the source. These parameters can be modified in the spectral domain, according to spectral modeling of the glottal source [40, 41]. Using this theory, WORLD spectral representation is well suited to voice quality transformations.

4.3.1. Glottal formant

Voice tension is an important voice quality parameter. In the spectral domain, the glottal pulse corresponds to a peak of the spectral envelope in the region of the first harmonics. Changing the voice tension results in a shift of this peak, called the *glottal formant* [41]. This makes the voice sound more tense, when the center frequency of glottal formant is raised, and more relaxed when the center frequency of glottal formant is lowered.

As the glottal formant center frequency is situated near the fundamental frequency, change in the relative level of the first and higher harmonics emulates the effects of a glottal formant shift. To manipulate the harmonics, a comb filter is applied to the spectral envelope. Separation of the harmonics is done by multiplying the whole power spectrum of the periodic part $|H_p(\omega)|$ by a function that is approximately equal to 1 for $\omega = 2\pi f$ and to 0 for $\omega \geq 2 \cdot 2\pi f$. A sigmoid function is used because a smooth function is required. The harmonics are weighted, according to the desired glottal formant modification using a γ parameter. This is equivalent to weighting the entire power spectrum with a function $g_\gamma(\omega)$. The modified power spectrum is:

$$\begin{aligned} |H'_p(\omega)|^2 &= \Big((1-\gamma)E_1(\omega) + \gamma E_{\text{sup}}(\omega)\Big)|H_p(\omega)|^2 \\ &= g_\gamma(\omega)|H_p(\omega)|^2 \end{aligned} \tag{8}$$

where E_1 and E_{sup}, the envelopes corresponding respectively to the first and higher harmonics, are sigmoids defined by

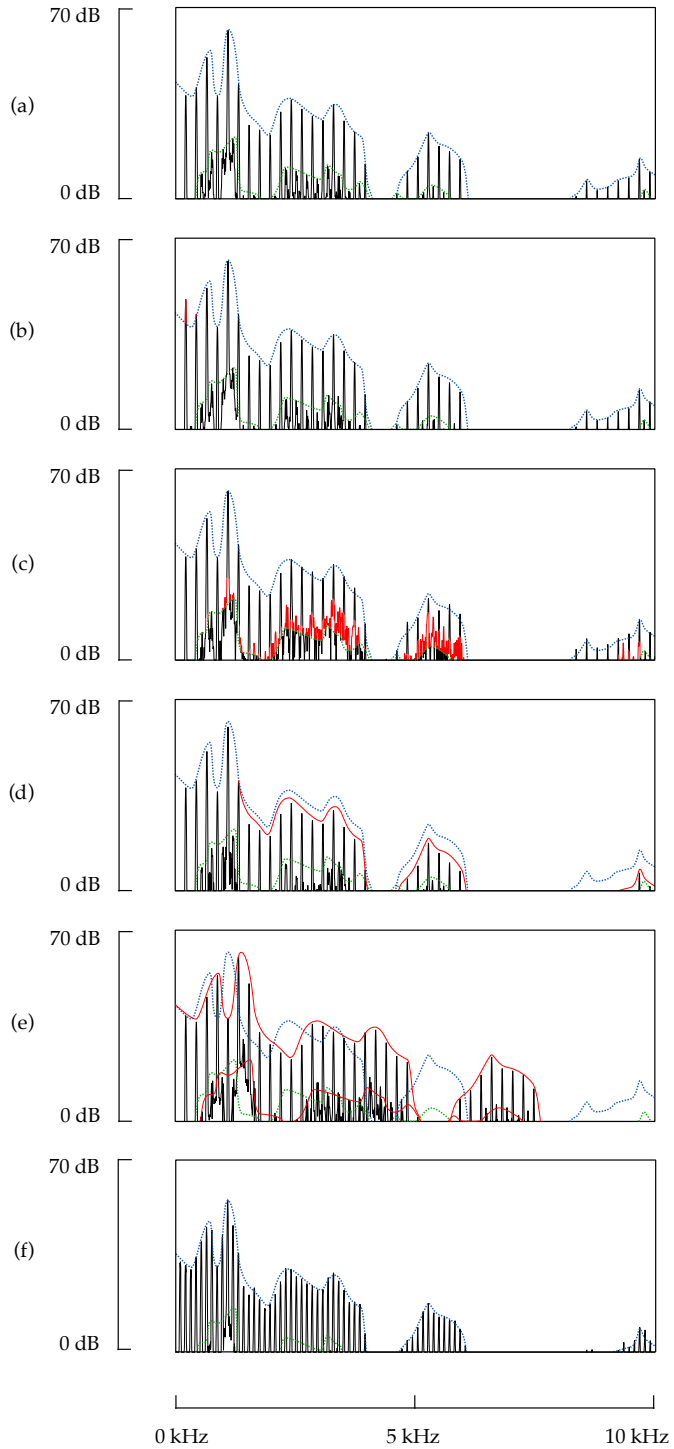

Figure 6: Power spectrum of the vowel /a/ played (a) with no timbre modifications (b) with a shifting of the glottal formant simulated by applying a gain of 10 on the first harmonic (c) with a gain of 10 applied on the aperiodic part (d) with the application of a spectral slope with a cutoff frequency of 1000 Hz (e) with a simulated vocal tract elongation with a factor of 0.8 (f) with fundamental frequency one octave higher. In (a) an envelope has been manually drawn around the peaks corresponding to the periodic (blue) and aperiodic (green) parts of the signal. These envelopes appear in (b), (c), (d), (e) and (f); in (b) and (c), deviations of those envelopes from the original ones are also marked in red. In (d) an updated envelope has been drawn as a red dotted line, also manually. In (e), a stretched version of the envelopes with a stretching factor of $\frac{1}{0.8}$ appear as a red dotted line. In (f), the original envelopes remain relevant.

$$
\begin{aligned}
E_1 &= 1 - \tanh\left(8\left(\frac{\omega}{2\pi F_0} - \frac{3}{2}\right)\right) \\
E_{\text{sup}} &= 1 + \tanh\left(8\left(\frac{\omega}{2\pi F_0} - \frac{3}{2}\right)\right)
\end{aligned}
\tag{9}
$$

with F_0 the fundamental frequency. In terms of the total power spectrum and the aperiodicity ratio, equation 8 becomes

$$
\begin{aligned}
E'(\omega) &= \left[\left(\left(1 - R(\omega)\right) g_\gamma(\omega) + R(\omega)\right)\right] E(\omega) \\
R'(\omega) &= \frac{R(\omega)E(\omega)}{E'(\omega)}
\end{aligned}
\tag{10}
$$

The weighting of the first harmonic can be seen in the second graph in Figure6 : here the first harmonic exceeds the envelope of the harmonics of the original sound.

4.3.2. Spectral slope

A higher vocal effort corresponds to higher intensity and higher spectral richness (lower spectral tilt). Conversely, a softer voice corresponds to lower intensity and the attenuation of higher harmonics (higher spectral tilt). In the spectral domain, an additional spectral slope, that is, a low-pass filter of order 1, with controllable cutoff frequency, is applied to the harmonic part:

$$
H'_p(\omega) = \frac{H_p(\omega)}{\sqrt{1 + \left(\frac{\omega}{\omega_c}\right)}}
\tag{11}
$$

where ω_c is the cutoff frequency, used as a control parameter.

The fourth graph in Figure6 shows the difference between the original harmonic envelope and the harmonic envelope of the same vowel with a spectral slope applied.

Equation 11 can only decrease vocal effort (produce a softer voice). Convincingly increasing vocal effort is more difficult, because higher order harmonics are often masked by noise. More sophisticated methods are needed (e.g. distortion [42]). In the present study, only lowering vocal effort is implemented.

4.3.3. Periodic-aperiodic ratio

A welcome feature of WORLD's parametric representation is its built-in separation of the periodic and aperiodic components of the signal. By varying the respective amounts of noise and voicing in the signal, one can achieve various interesting effects. The variation is achieved by simple amplitude scaling:

$$
\begin{aligned}
H'_{ap}(\omega) &= \alpha H_{ap}(\omega) \\
H'_p(\omega) &= \beta H_p(\omega)
\end{aligned}
\tag{12}
$$

where α and β are control parameters. Significantly decreasing β with respect to α produces a lax voice (joined to a lowering of vocal effort and/or tension). Setting it equal to zero produces a very convincing whispered voice.

The weighting of the aperiodic part independently of the periodic one can be seen in the third graph of Figure6: while the harmonic peaks are unchanged, the noise in between them is increased.

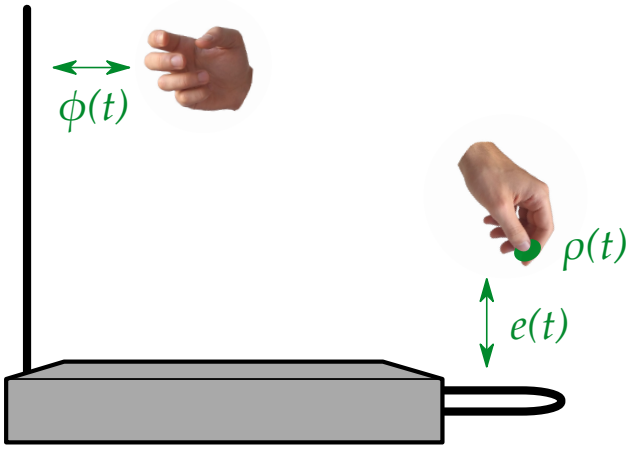

Figure 7: The T-Voks control interface. The antennae of a theremin respectively control pitch ($\phi(t)$) and vocal effort ($e(t)$), and a pressure sensor located in the hand controls vocal effort.

4.4. Vocal tract control

The size of the vocal tract has a dramatic effect on the spectrum. This effect is generally noticeable; the perceived size of the vocal tract is an important parameter, that allows listeners to instinctively estimate e.g. the perceived age and gender of a voice. Linear frequency warping is easily implemented thanks to the source/filter decomposition in WORLD:

$$\begin{aligned} H'_{ap}(\omega) &= H_{ap}(\lambda\omega) \\ H'_{p}(\omega) &= H_{p}(\lambda\omega) \end{aligned} \tag{13}$$

where λ is a warping factor.

For a uniform vocal tract (corresponding to the neutral vowel /ə/) the λ is a warping factor corresponding to a *vocal tract size factor*: the scaling factor of a fictitious "vocal tract" whose characteristics are described by H_{ap} and H_p. For other vowels, the situation is more complex and the linear warping factor λ does not correspond exactly to a vocal tract scaling factor. However, the perceptual effect of linear warping corresponds well to an apparent vocal tract size change, at least for voices. According to [43]: "The scaling of children's data from female data comes closer to a simple factor independent of vowel."

The (e) graph in Figure6 shows the power spectrum of a vowel played with a vocal tract size factor of 0.8. While the space between harmonics remains the same as in the original signal, their envelope is a stretched version of the original envelope, with a stretching factor of 1.25, the inverse of the vocal tract scaling factor.

Values of λ close to 1 ($0.85 \lesssim \lambda \lesssim 1.2$) allow for turning a perceived male voice into a perceived female ($\lambda < 1$) or the other way around ($\lambda > 1$). More extreme values of λ lead to more extreme effects, such as the "chipmunk voice" for values of λ significantly lower than 1.

Note that the transformation $H(\omega) \mapsto H(\lambda\omega)$ is applied both to the harmonic part of the signal and the noise. Much of the noise comes from the motion of articulators (e.g. tongue, lips, etc.), which should not, in principle, be affected by a vocal tract change. However, some of the noise does directly come from the vibration of vocal folds. Decoupling that noise from the harmonic vibration by applying the transformation only to H_p and not to H_{ap} gives rise to a less natural voice, and in some cases even gives the impression of two voices being heard at the same time.

5. T-Voks: Theremin-controlled Voks

5.1. Instrument design

T-Voks is a theremin-controlled performative voice synthesizer based on Voks [3]. The theremin is particularly spectacular because it is played by free-hand motion, in the "ether" without any contact between hands and the instrument. Pitch is controlled using the right hand[2] and the theremin's vertical antenna, while volume is controlled with the other hand using the looped antenna.

5.1.1. Pitch control

Pitch control using the theremin is challenging. Like in fretless string instruments the player must find the pitch in a continuum, without pre-defined visual or tactile steps or marks. An additional difficulty comes from the free hand motion, without haptic feedback, such as the neck of a violin. Despite widely held views on the theremin's difficulty, many players manage to accurately play melodies within several months of practice. Theremin performance is a highly individual art, with no single standardized technique, though some educational resources do exist [44]. To reliably find notes and intervals, thereminists have developed specific hand and finger gestures, such as opening and closing the hand toward the antenna for the span of an octave and shaking the right hand to produce a vibrato. In addition, 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 modifies the electromagnetic fields detected by the antennas.

5.1.2. Vocal effort and syllabic control

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 (left or non-preferred) hand in relation to the horizontal antenna is used to cleanly delineate notes by removing unwanted glissandos in between. The volume control is associated with vocal effort in T-Voks. 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.

An additional pressure sensor for the volume-control hand allows for syllabic rhythm control. When the pressure value is lower than a threshold ("open hand"), the control parameter ρ (section 4.1) is equal to 1; when ρ is higher than the threshold, ρ is equal to 0.

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 syllabic control button positioned against the first knuckle of the index finger, held in place by a ring and pressed by the thumb. 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. Note that the addition of syllable advancement introduces a significant cognitive load to an instrument that already requires full concentration for playing.

[2] Right-hand (resp. left-hand) means here the preferred hand (resp. non preferred). It was actually the right hand in our theremin experiments, but it could be the left as well for another player.

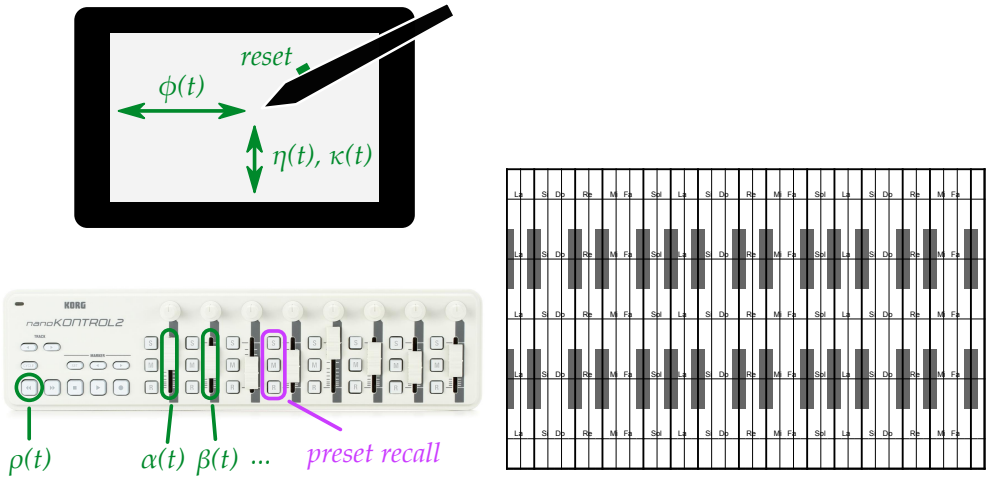

Figure 8: The C-Voks control interface, and mask affixed to graphic tablets for pitch accuracy

5.2. *Hardware and Computer Interfaces*

5.2.1. *Theremin*

The theremin used in T-Voks is the Etherwave Plus from Moog, which features control voltage (CV) analog outputs for pitch and volume. Pitch is associated to the vertical right-hand controlled antenna. The pitch CV ranges from -2.5v to 4.5v (with a change of 1 volt per octave of the theremin's pitch). Vocal effort is associated to the left-hand horizontal antenna, or volume antenna. The volume CV from our theremin measured from 0 to 5.5v. Voltage must be digitized before computer processing. This is achieved using an Arduino Uno. Each CV output is fed into an analog input pin of Arduino, with the CV ranges regulated to the 0 to 5 volt range of the Arduino.

5.2.2. *Syllabic control interface*

In a first experiment, 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[3], whose output vary from 0 to 5 volts between maximum pressure and no pressure. 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. 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. In a second experiment, a wireless interface (Bluetooth mouse controller) is used, giving more freedom to the performer, and direct input to the computer.

6. C-Voks: Tablet-controlled Voks

C-Voks (standing for Calligraphic *Voks*) is a voice synthesizer based on Voks controlled by a graphic tablet. The graphic tablet has been used as a control interface for singing instruments before [12]. In addition to the tablet, a MIDI controller is used for control of rhythm and other parameters, as well as preset management. C-Voks is a new implementation of the Vokinesis and Calliphony systems. The main differences between these systems are the vocoder used (RT-PSOLA for Vokinesis and Calliphony, WORLD for C-Voks) and the graphical user inferface. They are otherwise very close as far as sound quality, playing modes and musical functionality are concerned.

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

6.1. Instrument design

C-Voks is a bimanual voice synthesizer controlled by a pen on a graphic tablet (preferred hand) and an additional interface, which allows for the control of syllable sequencing, voice quality and vocal tract scaling. The instrument design relies on a different approach toward chironomic control than for the theremin. In the case of C-Voks, very accurate motions of the pen on a surface allow for multimodal reinforcements [45]. Taking advantage of the cooperation among visual, audio and kineaesthetic modalities, C-Voks is easily accessible to beginners, who can perform simple melodies almost at the first training session.

6.1.1. Pitch control

C-Voks pursues the lineage of a series of instruments based on the graphic tablet and writting/drawing gestures: Cantor Digitalis [12], Calliphony [25] and Vokinesis [1]. Playing C-Voks is reminiscent of playing Cantor Digitalis and other tablet-based melodic instruments.

The graphic tablet offers a three dimensional control: the two position coordinates of a stylus on the rectangular surface of the tablet, as well as a value for the vertical pressure of the stylus on the surface.

The graphic tablet as a musical instrument already has some precedent [46]. It is particularly well-suited to intonation pitch control in speech [21] and singing [47]. This is because on the one hand the gestures on the tablet re-use the manual skills acquired for hand writing and drawing, and on the other hand because the visual, kineaesthesic and auditory modalities collaborate in the task [48]. In C-Voks, horizontal position of the stylus on the surface of the tablet is mapped to pitch. A mask with lines locating the notes of a chromatic scale on the surface, shown in Figure8, is affixed to the tablet for visual assistance of the player.

6.1.2. Time and Rhythmic controls

C-Voks offers three different modes for rhythm and timing control: syllabic rhythm control, speech rate (speed) control and signal scrubbing. Depending on the timing control mode, different kinds of control parameters are needed: a continuous control for the speech rate and scrub modes, or a discrete, binary parameter for the syllabic control mode.

When playing in scrub or speed mode, the time parameter is continuous, and although any continuous control dimension can be used, one of the axes of a graphic tablet is especially well-suited to the task. In this case the stylus is used for two simultaneous tasks: melodic control and timing control. Gestures on the surface can create new sounds and new ways to play with the voice. Moving the time index τ with fast motions, using either scrub or speed mode, results in a rapid succession of syllables, either forward or backwards, too fast and chaotic to have been produced by a human, but still retaining most of the attributes of voice. When using speed mode, each time the time index τ reaches its maximum value, it loops back to 0. For sentences that last several seconds, this just results in the same text being repeated over. However, for shorter utterances, typically single syllables, fast repetition makes the audio less voice-like, granting it an acousmatic quality. Depending on speed, pitch, source sample used, and whether it is being played forward or backward, the resulting sound can evoke different textures, like e.g. babbling, bubbles, or a motor.

When playing in syllabic mode, the time parameter is discrete, and a two state button or keyboard needed, like for T-Voks. In this case the bimanual control is divided in melodic and vocal effort control (preferred hand) and syllabic rhythm control (non preferred hand).

6.1.3. Vocal effort and voice quality controls

The voice parametric representation in M-WORLD allows for many kinds of voice quality control in C-Voks : vocal effort, vocal tension, periodic-aperiodic ratio, apparent vocal tract size.

The most important voice quality parameter is vocal effort. This parameter combines volume and spectral variation of the vocal source. The stylus pressure is used to control vocal effort in C-Voks, whatever the timing control mode. Other parameters are controlled using a MIDI controller with buttons, knobs and sliders:

- One of its buttons is used as the temporal control when playing in syllabic mode.

- Sliders and knobs control various additional parameters such as vocal tract scaling factor, articulation speed, overall sound level and many more.
- Numerous buttons allow for easy selection of presets.

Vocal effort is varied by the combination of three signal modification:

- Glottal formant shift, as described in section 4.3, to increase/decrease tension.
- Application of a different gain to the periodic and the aperiodic part of the signal, as described in section 4.3.3: for a low vocal tension, the periodic part is attenuated and the aperiodic part enhanced.
- Variation of the cutoff frequency of the spectral slope, as described in section 4.3.2: when the vocal effort parameter is high, the cutoff frequency is set to a high value, allowing more higher harmonics to make it through than in the converse case.

In addition to the periodic/aperiodic ratio modification related to vocal effort changes, the respective gains of the periodic and aperiodic parts of the signal are also control parameters that can be set directly. Muting the periodic part then results in whispered voice, which the performer can then modulate to produce breathing, blowing, hissing sounds.

The vocal tract size can be changed by spectral morphing (section 4.4). Vocal tract size parameter expands or contracts the voice spectral envelope by a given factor. Shortening the vocal tract gives a female-sounding voice using a sample recorded by a male speaker, and vice versa (with a vocal tract scaling factor equal to about 0.8 and 1.22 respectively). More extreme changes give ”child” or ”giant” voices.

6.2. Hardware and Computer Interfaces

6.2.1. Graphic tablet

The graphic tablet used is a Wacom Intuos Pro. It is composed of a flat, rectangular surface, and a stylus. When the tip of the stylus is in contact with the surface, its position along both the X- and the Y-axis, as well as its pressure on the surface, are sent to Voks. The position along the X-axis is mapped to pitch by a linear relationship, the pressure is mapped to vocal effort, and the position along the Y-axis is mapped to the time control parameter, respectively $\eta(t)$ in scrub mode and $\kappa(t)$ in speed mode (see sections 4.1.2 and 4.1.3). In syllabic mode, position along the Y-axis is not used.

A push button on the stylus is also used for resetting the time index value τ at 0 when in syllabic and speed mode.

6.2.2. MIDI controller

In addition to the graphic tablet, a MIDI controller Korg Nanokontrol 2 has been used for concerts. It is a MIDI controller that features 8 groups of controls — each composed of one slider, one knob, and three push buttons — as well as 11 more push buttons on the left. The sliders and knobs are mapped to the continuous parameters that are not controlled by the tablet : vocal tract scaling factor λ, voiced and unvoiced factors α and β, etc. One of the 11 push buttons on the left is mapped to the syllabic parameter ρ when in syllabic control mode (section 4.1.1).

The push buttons from the 8 groups, originally intended for the soloing, muting, and recording of tracks, form an 8x3 matrix that is used in C-Voks for preset recalling. During rehearsals, the performer can define *presets*, that is, predefined values of:

- a sample to be recalled
- specific values for the parameters,
- a playing mode

which will be attached to one of the 24 buttons, and recalled at the time of the performance by simply pressing the corresponding button.

Figure 9: A performance [4] using one instance of T-Voks (left) and three instances of C-Voks

7. Comparative perceptual evaluation

7.1. Singing Synthesis Challenge Fill-in the Gap

Singing synthesis evaluation is an important but difficult task. This section reports on the Singing Synthesis Challenge held at SCA Interspeech 2016, in San Francisco, USA [49], where Voks has been evaluated in a comparative perceptual paradigm. This international challenge aimed at perceptual comparison of singing systems. This challenge was the third international singing synthesis evaluation, using common material shared by different research groups, following evaluation at Stockholm Musical Acoustic Conference in 1993 [50] and at ISCA Interspeech Conference in Antwerpen [51], in 2007. The Singing Synthesis Challenge "Fill-in the Gap (FiG)", was organized by one of the authors as a special session at the Interspeech 2016 conference in San Francisco, on september 10th 2016 (see a presentation of the challenge at [49]), in the framework of the ChaNTeR project. The following results have been presented at the closing ceremony of Interspeech 2016 in San Francisco, but have never been published. New statistical analyses of the results are given in the following sections.

7.1.1. Presentation of the challenge

The task was to synthesize well known jazz standards with new lyrics, written for the occasion. Two popular song where chosen: "Summertime" (song S) music by George Gershwin (1934), and "Autumn Leaves" (song A) music by Joseph Kosma, originally in French "les feuilles mortes" (1946). These songs were selected because they are international jazz standards, innumerable versions of these songs in many languages and various vocal styles (opera, pop, jazz ...) have been recorded. Original lyrics have been written for both songs in English and French for the FiG challenge. The scores are displayed in Figure . Participants were free to select their preferred version, or translate the lyrics into any language. A huge number of recordings and instrumental playback was available (e.g. on the web) and could be used for reference, acoustic analysis, machine learning, or comparison. The listening test was only performed for a capella (unaccompanied) versions of the songs, although the participants were also encouraged to produce accompanied versions for playing during the InterSpeech conference.

All aspects of singing synthesis and all methodologies were welcome, including both off-line (studio) singing synthesis systems, with no limits on time for producing the result, and performative (real-time) singing instruments. The Special Session FiG has been announced in December 2015. The musical material, score and lyrics, and the FiG Challenge rules and instructions were issued on January 21th, 2016, about two months before the Interspeech 2016 paper submission deadline, March 23th, 2016.

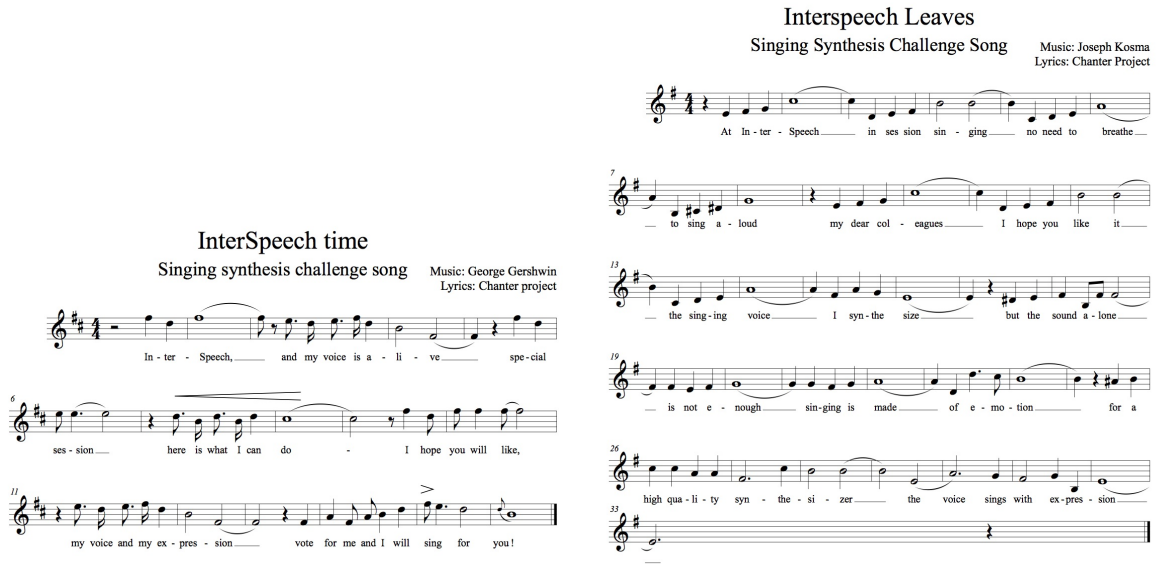

Figure 10: Songs for the Singing Synthesis Challenge Fill-in the Gap at Interspeech 2016. The lyrics have been especially written for the challenge.

# items	Lab	lang	songs	voice	style	method
2	WBHSM	English	S A	Male	jazz	concatenative
4	ISIS	French	S A	Male, Female	jazz	concatenative
2	C-Voks	French	S A	Male	jazz	performative
4	ACAPELA	French	S A	Male, Female	jazz	concatenative
1	Seraphim	Mandarin	A	Female/male	pop	concatenative
1	Bersokantari	Basque	A	Male	traditional	concatenative

Table 1: Participants to the Singing Synthesis Challenge Fill-in the Gap at Interspeech 2016

7.1.2. Participant to the challenge and Test methodology

A number of papers were submitted to the Special Session, and among them 6 research groups were selected and participated in the singing synthesis challenge. Table 7.1.2 summarizes the languages, numbers of submitted songs, voice genders and participating labs. For a detailed description of each system, the reader is referred to [49]: the WBHSM concatenative synthesizer (UPF, Barcelona) [16], ISIS, the Ircam Singing Synthesizer (Paris)) [52], the Seraphim system (A*STAR, Singapore) [53], the Bertsokantari system (UPV, Bilbao) [54], the ACAPELA singing synthesis system (Mons) [55], and Calliphony, an earlier implementation of C-Voks. For the sake of simplicity, the system is coined C-Voks. Overall, 14 samples of synthetic songs were used for subjective evaluation. The total duration of the 14 samples was 11mn30s. Therefore performing the test was relatively fast and easy. The samples were long enough, on average 49.3s, to elicit a true musical appreciation, encompassing sound quality, musical quality, singing style, musical interpretation, and so on.

All the participants except C-Voks developed an off-line singing synthesis system (or text-to-chant) systems. In such systems, the score and lyrics are written in a text file, and the sound is computed off-line (and not in real time) according to this input data. C-Voks was the only performative singing synthesis system. For this system, the song is played in real-time, and after a number of trials, the best version is selected. More or less knowledge is used depending on the system, but all systems are based, like C-Voks, on recorded voice samples. The synthesis paradigm used for all the text-to-chant systems is concatenation of diphones or other voice segments.

An Absolute Category Rating paradigm measuring the Mean Opinion Score seemed appropriate for the evaluation task. The subjects were asked to rate the quality of the synthesized songs on a 5-point quality scale, with 1 being the lowest perceived quality and 5 the highest perceived quality. It was an Absolute Category Rating test measuring the Mean Opinion Score. The subjects were advised to listen over headphones and to use the full judgement scale for reporting their appreciation of the songs. They could listen to the 14 samples sounds as many time as they wished, and in the order they wanted. All the samples were presented on the screen. The order of presentation of the samples on the screen was randomized and different for each presentation, in order to avoid a possible visual presentation effect. An internet-based international listening test was advertised on relevant speech, singing and music mailing lists, and launched for 12 days between August 29th and September 9th 2016.

7.1.3. Instrument used in the challenge

The instrument used in the challenge was an early version of C-Voks called Calliphony. Like C-Voks, the Calliphony system was controlled with a stylus on a Wacom graphic tablet, using the preferred hand. Rhythm was controlled by pressing / releasing the control button using the non-preferred hand. The main difference between the current version of C-Voks and Calliphony is the vocoder used. Sound processing in Calliphony was performed with the help of a real-time PSOLA vocoder. Sound quality of the RT-PSOLA and WORLD vocoders are equivalent, but as WORLD offers additional spectral controls, it is preferred in the current version of C-Voks.

7.2. Results

7.2.1. Subjects

The listening test was launched worldwide and a grand total of 198 responses were received during the 12 days of test opening. Responses came from 18 different countries (France, Germany, Switzerland, United Kingdom, USA, Japan, Denmark, New Zealand, Spain, Austria, Belgium, Sweden, Canada, Poland , Australia, Brazil, Ireland, Marocco), with a noticeable bias towards Europa (with about 3/4 of responses) and France (with about 1/3 of responses). Among these 198 responses, only 80 complete responses, with scores for the 14 songs, were retained for further analysis. The other 119 responses were incomplete, probably just for curiosity, but they were not considered for analysis.

The total duration of the 14 sounds was 11mn30s. On the 80 retained full test, only 22 responses took longer than the full stimuli duration. 56 responses took longer than 6mn (half of the full sound stimuli duration). This means that subjects felt comfortable to take a decision on song quality before listening to the whole samples. The Number of subjects (y axis) having performed the test in less than a given time (x-axis, in mn) are plotted in Figure 11. As a basis for analyses, we selected the 56 subjects that took 6 minutes or more. This is because the results obtained in ranking scores are the same for this group and the smaller set of 22 subjects that took more than 11mn30, although the scores are slightly different.

7.2.2. MOS, ranks and groups

The Mean Opinion Scores obtained for the 56 subjects are displayed in Figure 12. The highest score is 4.21 MOS and the lowest 1.68. This indicates that the listeners used the whole scale for their judjements.

For further analysis of the Lab factor, a post-hoc Tukey's honestly significant difference (HSD) test was run on this factor. The analysis are reported in Table 7.2.2. The post-hoc test gives four statistically different groups A, B, C, D. The three Lab (L1, L2, L3, they are anonymized) in group D received statistically comparable scores. Both the MOS for all the songs for a same Lab and the best song for a given Lab are reported. Note that the same grouping result is obtained when all the 80 subjects are considered. The same ranking is obtained when taking the best Song in each Lab.

7.2.3. ANOVA

An ANOVA was run on the quality scores as a dependent variable, with the factors Song (song A or S), Voice (female or male voice) and Lab (the laboratory having produced the synthesis, 6 levels) as main factors, and the two way interactions between Lab * Voice and Lab * Song. Results of the analysis of variance

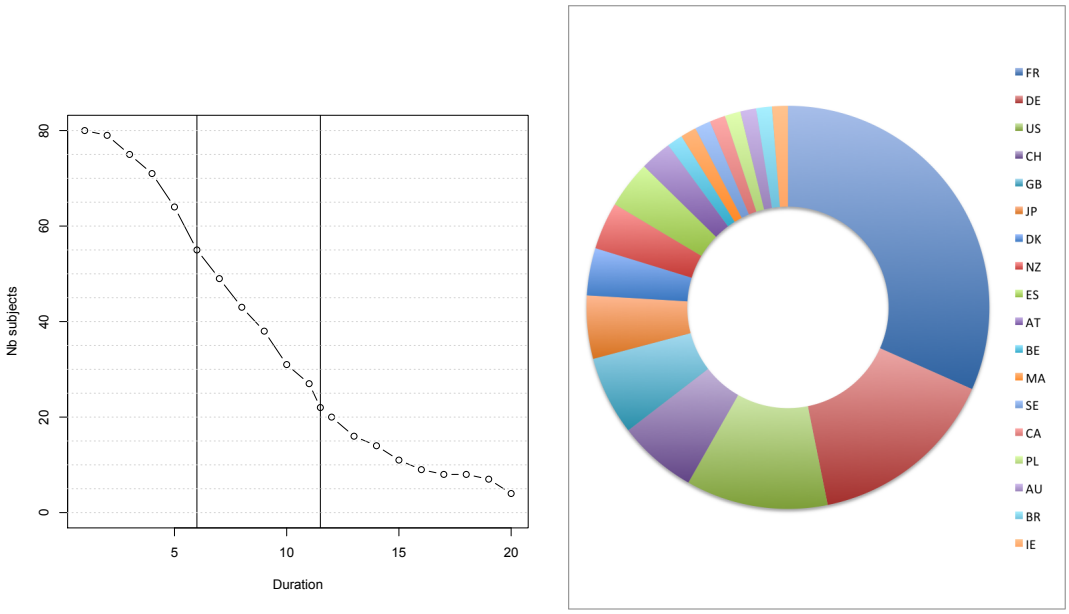

Figure 11: Left panel: number of subjects (y axis) having performed the test in less than a given time (x-axis, in mn). The two vertical lines indicates the full song duration (11mn30s) and more thanf halp song duration (6mn). Right panel: countries of participant.

Rank	Lab	Score	Group	best song
1	WBHSM	4.15	A	4.21
2	ISIS	3.20	B	3.67
3	C-Voks	2.57	C	2.83
4	L1	2.25	D	2.70
5	L2	2.22	D	2.22
6	L3	2.20	D	2.20

Table 2: Ranking and groups obtained by a post-hoc Tukey's HSD test.

are reported in table 7.2.3. All main factors have a significant effect on the result. The interaction between Lab and Song is significant: this means that for some systems the difference in appreciation between the two songs is statistically different. Male voices received on average a significantly higher score (3.1) than female voices (2.4). The A song received significantly higher scores (3.0) than the S song (2.7). As expected, the factor Lab has the strongest effect size (cf. the η^2 column in Table 7.2.3). Most of the variance in the result is explained by the system that produced the song.

7.3. Discussion

The top three systems of the Singing Synthesis Challenge Fill-in the Gap at Interspeech in 2016 were:

1. WBHSM [56].
2. ISIS [57]
3. C-Voks (audio examples 9 et 10, songs A and S with accompaniment)

C-Voks ranked third in the challenge, after two high-quality text-to-chant system. This demonstrates that the sound quality of performative (real-time) systems is lower than that of the best off-line concatenative systems. It obtained a MOS of 1.38 less than the best system. But it is higher than the three others off-line concatenative text-to-chant systems. Another comparative perceptual study, involving C-Voks (Calliphony) and ISIS, is reported in [15]. Comparable results are obtained with the same ranking and

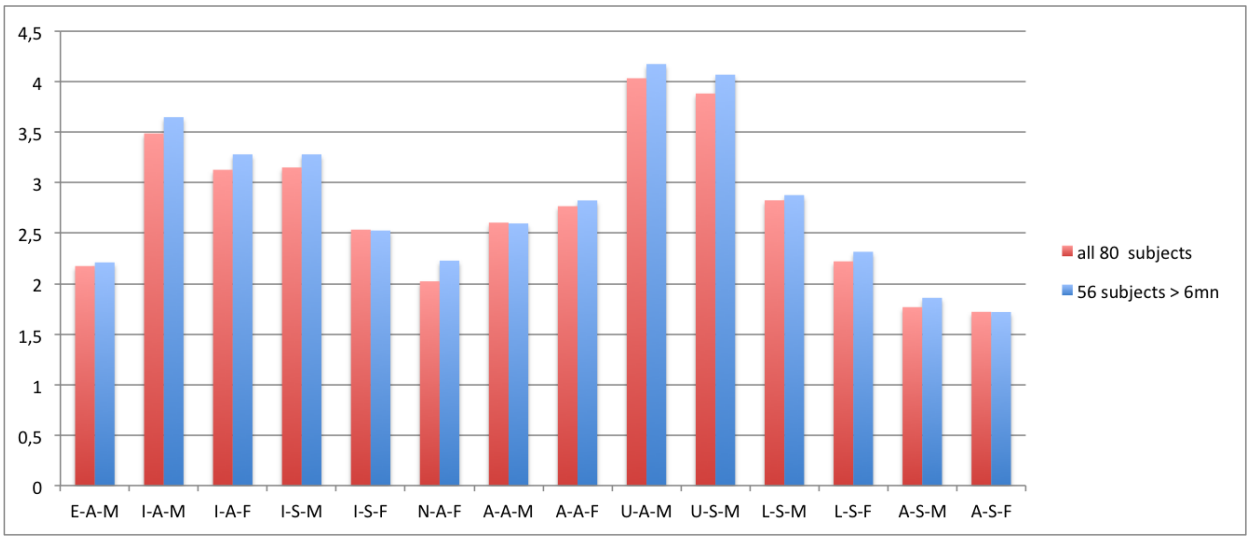

Figure 12: Lab: I=ISIS E=L1 N=L2 L=C-Voks U=WBHSM A=L3, Song: A and S. Voice: M=Male F=Female

	Sum Sq.	df	df error	F	p	η^2
Lab	341.1	5	758	73.3	<0.001	0.33
Gender	22.7	1	758	24.4	<0.001	0.03
Song	35.6	1	758	38.3	<0.001	0.05
Lab*Gender	3.1	2	758	1.7	0.19	0.00
Lab*Song	7.9	2	758	4.3	<0.001	0.01

Table 3: Effect of each factor on the dependent variable, as measured by an ANOVA

a similar difference of about 1 point in the scale MOS between off-line concatenative text-to-chant and performative singing synthesis systems. As the most advanced research groups worldwide working on Singing Synthesis at this time took part in the challenge, it can be regarded as an accurate picture of current achievements.

8. Discussion and conclusion

8.1. New vocal expressions

Performative voice synthesis is a new paradigm in the already long history of artificial voices. The voice is played like an instrument, allowing for speaking or singing with the borrowed voice of another. In Voks, voice samples, produced by a true vocal apparatus, are played by free-hand theremin-controlled gestures, and by writing gestures on a graphic tablet. The same types of sounds, controlled with other gestures, give rise to yet other musical instruments and expressive possibilities. The relationship of embodiment between the singer's gestures and the vocal sound produced is broken. A voice is speaking or singing, with realism, expressivity and musicality, but it is not the musician's own voice, and a vocal apparatus does not control it. This introduces a special relationship between the synthetic voice and the player, the voice being at the same time embodied (by the player gestures playing the instrument with her/his body) and externalized (because the instrument is not her/his own voice). In some performances, two different voices can be sung and played by the same person (Video example 2).

Performative voice synthesis opens new expressive possibilities:

- Voice deconstruction sounding like computer music or electroacoustic voice. Parametric representation and modeling of the voice allows for extreme variations. Specific features of the voice can

Video example 1	excerpt from La Vie en rose, Edith Piaf & Louiguy (T-Voks)
Video example 2	excerpt from Brirouch, video performance (C-Voks)
Video example 3	excerpt from My Funny Valentine, Richard Rodgers and Lorenz Hart (T-Voks)
Video example 4	excerpt from Pierrot Lunaire, Arnold Schönberg (T-Voks)
Video example 5	Chun Xiao, Meng Haoran (T-Voks)
Video example 6	excerpt from All Shall Be Well, Christophe d'Alessandro & Julian of Norwich (T-Voks ar
Audio example 1 (still video)	Singing Synthesis Challenge Song A (C-Voks)
Audio example 2 (still video)	Singing Synthesis Challenge Song S (C-Voks)

Table 4: Accompanying audio and video examples.

be emphasized (formants, pitch, voice quality, vocal tract size, roughness), and a rich sonic material based on the voice can be worked out in real time.

- Voice imitation, on the contrary, favors proximity between natural and synthetic voice. How close to natural voice can a synthetic voice be ? In some situations, a realistic voice is desirable. It is at the (possibly interesting) risk of an "uncanny valley" effect.
- Voice extension in between deconstruction and imitation, the augmented voice is a realistic-sounding voice with augmented (naturally impossible) features: for instance, a voice with a very large register, a male/female voice, a very slow, very fast pronunciation, small and large vocal tracts. Another aspect of voice augmentation is the specific vocal gestures allowed by the control interfaces: here the theremin and graphic tablet.

These expressive features are useful for musical and poetic purposes, as well as for speech and voice communication applications (discussed below). Examples of performances are given in the next section.

8.2. Performance examples in various languages

T-Voks and C-Voks have been used on stage for several musical and poetic performances [4, 58]. Four languages have been used so far: French, English, Mandarin and German.

8.2.1. Examples in French

Video example 1 is a musical example featuring the French song "La vie en rose" (Edith Piaf, Louiguy), played by T-Voks. Samples are from spoken utterances of a French female speaker with no musical background or voice training. Syllabic rhythm control works well for this example, as French is a syllable-timed language [59]. The theremin is able to replicate key features of Piaf's signature vocal style, including dramatic vibratos and small glissandi at the start and end of phrases.

Video example 2 is a poetic example featuring the fairy tale "Histoire de Brirouch", played by C-Voks (Vokinesis). In this example, the speed timing control mode is demonstrated. The stylus is only used for melodic, rhythmic and dynamic control. Samples are from spoken utterances of a professional French male singer.

8.2.2. Examples in English

Video example 3 is a musical example featuring the jazz standard "My Funny Valentine", popularized by Chet Baker. Samples are from spoken utterances of a non-professional female American English speaker, resampled by a factor of 1.22 to yield a male voice. A specific setting of vocal effort lends "breathiness" to the synthesized voice, inspired by Chet Baker's singing style.

Unlike French, English is a stress-timed language [59]. Syllable control requires paying more attention to stress timing and inter-syllable transitions. Controlling English, a language where diphtongs, i.e. vocalic changes inside a syllabic nucleus, are common, using the biphasic syllable control, was found challenging by performers.

8.2.3. Examples in German

Between speech and singing, video example 4 demonstrate an example of *Sprechstimme* in German. *Sprechstimme* is a vocal technique where singing imitates the continuous pitch contours of speech. In the score of his *Pierrot Lunaire* [60], Arnold Schoenberg gives the following direction for achieving *Sprechstimme*: to "be well aware of the difference between speaking tone and singing tone", by singing the written pitches, but altering them right after, all the while singing the rhythm as written.

Samples are from an excerpt of *Pierrot Lunaire* recorded by a native French male speaker, resampled by a factor of 0.8 to yield a female voice. As a stress-timed language, but with a strong syllabic structure, German shares the same considerations for syllable advancement as English, but the smaller distinction between stressed and unstressed syllables [29] makes it closer to French. For a convincing *Sprechstimme*, pitch slides and their volume curves must also correspond to the correct stress pattern. Pitch slides are achieved with T-Voks by small displacements of the fingers or by pivoting around the wrist.

8.2.4. Examples in Mandarin

Video example 5 is a poetic example in Mandarin. 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 [61]. Classical poetry is typically recited with exaggerated tone enunciation.

A well-known Tang dynasty short poem was recorded by a Mandarin speaker pronouncing each syllable in monotone. The poem was then "recited" using T-Voks, with each tone shaped entirely by the theremin. Syllabic rhythm control is used.

Tones were created mostly with the preferred hand, with the non-preferred 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 changes 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.

8.3. Future work

Three points must be mentioned to conclude this article. The first point is a practical one. Data preparation is a time-consuming preliminary task for playing with Voks. Possible solutions are discussed for faster preparation of the sound and linguistic data. The second point concerns forthcoming application of Voks for education and reeducation. Finally, the general question of performative voice synthesis is discussed.

8.3.1. Sample production and text-to-speech synthesis

The task of labelling a speech sample is currently a manual and tedious one: one must input the location of each one of the control points by hand. An automated procedure for labelling samples, based on an automated phoneme segmentation and rules to convert such a segmentation into a labelling, is a possible solution, although it is prone to errors. Another option would be to make the recorded speaker generate the labelling during the time of the recording, by using a gesture similar to the one used by the performer, such as the pressing of a button simultaneous with the uttered syllables.

Input samples can be generated by text-to-speech systems. Those systems aim to emulate ordinary speaking voices, which differs from samples recorded by humans specifically for Voks, in which speakers make an effort to articulate and detach syllables. Thus coarticulation in synthesis based on text-to-speech-generated samples features altered phonemes. Generating samples with the help of text-to-speech systems also has some benefits: it eliminates the need to record a sample prior to the performance, and it makes automatic segmentation easier, which is an important step of automatic biphasic labelling.

8.3.2. Applications to education and reeducation

In a forthcoming project, the use of manual gestures, mediated by new Human Machine Interfaces like Voks, will be investigated for designing innovative tools and methods for intonation education (training) and re-education (re-training). The control of a synthesized voice through hand gestures is a new research

paradigm in the field of human-machine interaction. Previous studies have demonstrated that chironomic intonation using handwriting gestures on a graphical tablet can be even more precise and accurate than the natural voice in imitation tasks [21, 47]. The high performance in chironomy for performative voice synthesis can be attributed to its intrinsic multimodal integration (vision, kineaesthesis and audition [48], as well as to the existing dexterity of the handwriting movements (as used for writing and drawing purposes), which were repurposed for a new task.

It appears that performative voice synthesis could also foster new important applications in language acquisition and vocal substitution. A first foreseen application is to develop an educational program based on chironomy and to test it in language classes. The second foreseen application is to develop tools based on chironomy for vocal impairment assistance. In the case of phonatory function impairment, gestural control can improve expressive intonation in an augmented reality paradigm: phonation is controlled or enhanced by chironomy and articulation is controlled by the true vocal tract. An extreme case is that of vocal substitution. In the case of laryngectomy inducing a voice loss, the gestural control of intonation must enable the restoration of both linguistic and expressive intonation.

8.3.3. Next steps in performative voice synthesis

Voice instruments, or performative voice synthesizers, are still facing a compromise between sound quality and free selection of sound material. Some systems allow [6] for free sound material (i.e. any sequence of speech can be produced, like in text-to-speech systems), but with poor quality. Other systems, like Voks, delivers high quality sound, but are limited to pre-recorded sound material (or pre-synthesized sound material).

Voks only allows for linear resequencing of prepared samples. The main difficulty for free, real-time speech control is the large combinatorial complexity of the possible syllables and the difficulty to specify or select them in real-time. In text-to-speech systems, the linguistic material is presented as a text: either typed on a keyboard or copied from a file. In Cantor Digitalis, the speech material is free, but limited to vowels. Selecting full text (sequences of vowels and consonants) on the fly, i.e. free text selection, for a performative voice synthesizer, has no straightforward solution to date. A possible strategy would be to present the perfomer with a number of possible subsequent syllables, computed in real-time based on the previous ones and the statistic distribution of syllables in the considered language.

Acknowledgments

Part of this work has been done in the framework of the SMAC (FEDER IF0011085) project funded by the European Union and the Région Île de France, and the Agence Nationale de le Recherche ChaNTeR Project (ANR-13-CORD-0011, 2014-2017). The authors are indebted to Dr. Albert Rilliard who helped much in statistical analyses for the evaluation test.

List of Figures

References

[1] S. Delalez, C. d'Alessandro, Vokinesis: syllabic control points for performative singing synthesis, in: Proc. of New Interfaces for Musical Expression, 2017, pp. 198–203.
[2] S. Delalez, C. d'Alessandro, Adjusting the Frame: Biphasic Performative Control of Speech Rhythm, in: Proceedings of Interspeech 2017, Stockholm, Sweden, 2017, pp. 864–868.
[3] X. Xiao, G. Locqueville, C. D'Alessandro, B. Doval, T-Voks: the singing and speaking theremin, in: M. Queiroz, A. X. Sedó (Eds.), NIME 2019 International Conference on New Interfaces for Musical Expression, Proceedings of the International Conference on New Interfaces for Musical Expression, UFRGS, Porto Alegre, Brazil, 2019, pp. 110–115.
[4] C. D'Alessandro, X. Xiao, G. Locqueville, B. Doval, Borrowed voices, in: International Conference on New Interfaces for Musical Expression NIME'19, NIME'19 Proceedings of the International Conference on New Interfaces for Musical Expression, Porto Alegre, Brazil, 2019, pp. 2.2–2.4.
[5] S. S. Fels, G. E. Hinton, Glove-talk: A neural network interface between a data-glove and a speech synthesizer, Neural Networks, IEEE Transactions on 4 (1) (1993) 2–8.
[6] S. S. Fels, 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) (1998) 205–212. doi:10.1109/72.655042.
[7] P. R. Cook, Spasm, a real-time vocal tract physical model controller; and singer, the companion software synthesis system, Computer Music Journal 17 (1) (1993) 30–44.

[8] G. Berndtsson, The KTH rule system for singing synthesis, Computer Music Journal 20 (1) (1996) 76–91.
[9] N. D'Alessandro, T. Dutoit, Handsketch bi-manual controller: investigation on expressive control issues of an augmented tablet, in: Proceedings of the 7th international conference on New interfaces for musical expression, ACM, New York, USA, 2007, pp. 78–81.
[10] N. d'Alessandro, C. d'Alessandro, S. Le Beux, B. Doval, Real-time calm synthesizer: new approaches in hands-controlled voice synthesis, in: Proceedings of the 6th International Conference on New Interfaces for Musical Expression (NIME'06), Paris, France, 2006, pp. 266–271.
[11] N. d'Alessandro, P. Woodruff, Y. Fabre, T. Dutoit, S. Le Beux, B. Doval, C. d'Alessandro, Real time and accurate musical control of expression in singing synthesis, Journal on Multimodal User Interfaces 1 (1) (2007) 31–39.
[12] L. Feugère, C. d'Alessandro, B. Doval, O. Perrotin, Cantor digitalis: chironomic parametric synthesis of singing, EURASIP Journal on Audio, Speech, and Music Processing 2017 (1) (2017) 2. doi:10.1186/s13636-016-0098-5.
[13] H. Kenmochi, H. Ohshita, Vocaloid-commercial singing synthesizer based on sample concatenation., in: INTERSPEECH, Vol. 2007, 2007, pp. 4009–4010.
[14] M. Umbert, J. Bonada, M. Goto, T. Nakano, J. Sundberg, Expression control in singing voice synthesis: Features, approaches, evaluation, and challenges, IEEE Signal Processing Magazine 32 (6) (2015) 55–73.
[15] L. Feugère, C. d'Alessandro, S. Delalez, L. Ardaillon, A. Roebel, Evaluation of singing synthesis: Methodology and case study with concatenative and performative systems, in: Interspeech 2016, 2016, pp. 1245–1249. doi:10.21437/Interspeech.2016-1248.
URL http://dx.doi.org/10.21437/Interspeech.2016-1248
[16] J. Bonada, M. Umbert, M. Blaauw, Expressive singing synthesis based on unit selection for the singing synthesis challenge 2016, in: Interspeech 2016, 2016, pp. 1230–1234. doi:10.21437/Interspeech.2016-872.
URL http://dx.doi.org/10.21437/Interspeech.2016-872
[17] M. Astrinaki, N. d'Alessandro, T. Dutoit, Mage-a platform for tangible speech synthesis, in: Proceedings of the International Conference on New Interfaces for Musical Expression, 2012, pp. 353–356.
[18] M. Astrinaki, Performative statistical parametric speech synthesis applied to interactive designs, Ph.D. thesis, University of Mons (2014).
[19] M. Blaauw, J. Bonada, A neural parametric singing synthesizer, in: Proc. Interspeech 2017, 2017, pp. 4001–4005. doi:10.21437/Interspeech.2017-1420.
URL http://dx.doi.org/10.21437/Interspeech.2017-1420
[20] C. d'Alessandro, N. d'Alessandro, S. Le Beux, J. Simko, F. Çetin, H. Pirker, The speech conductor: Gestural control of speech synthesis, Tech. Rep. Final Project Report #6, eNTERFACE'05, Mons, Belgium (July 18th – August 12th 2005).
[21] C. D'Alessandro, A. Rilliard, S. Le Beux, Chironomic stylization of intonation, Journal of the Acoustical Society of America 129 (3) (2011) 1594–1604.
[22] The MIDI Manufacturers Association, Los Angeles, CA, MIDI Polyphonic Expression, 1st Edition (March 2018).
[23] L. Haken, R. Abdullah, M. Smart, The continuum: a continuous music keyboard, in: Proceedings of the International Computer Music Conference, International Computer Music Association, 1992, pp. 81–81.
[24] R. Lamb, A. Robertson, Seaboard : a new piano keyboard-related interface combining discrete and continuous control, in: Proceedings of the International Conference on New Interfaces for Musical Expression, Oslo, Norway, 2011, pp. 503–506.
[25] S. Le Beux, C. D'Alessandro, A. Rilliard, Calliphony : a tool for real-time gestural modification and analysis of intonation and rythm, in: International Conference on Speech Prosody (SP 2010), Chicago, USA, Unknown Region, 2010, p. 4p.
[26] M. Morise, F. Yokomori, K. Ozawa, World: A vocoder-based high-quality speech synthesis system for real-time applications, IEICE Transactions 99-D (2016) 1877–1884.
[27] S. Le Beux, B. Doval, C. d'Alessandro, Issues and solutions related to real-time td-psola implementation, in: Audio Engineering Society Convention 128, Audio Engineering Society, 2010, pp. 1–6.
[28] P. F. MacNeilage, The frame/content theory of evolution of speech production, Behavioral and Brain Sciences 21 (4) (1998) 499–511. doi:10.1017/S0140525X98001265.
[29] P. Wagner, The rhythm of language and speech: Constraining factors, models, metrics and applications, Ph.D. thesis, Habilitationsschrift, University of Bonn (2008).
[30] P. Barbosa, G. Bailly, Characterisation of rhythmic patterns for text-to-speech synthesis, Speech Communication 15 (1) (1994) 127–137.
[31] S. Delalez, Vokinesis : an instrument for suprasegmental control of voice synthesis, Theses, Université Paris-Saclay (Nov. 2017).
[32] G. Fant, Acoustic theory of speech production, Mouton, 1970.
[33] M. Morise, Y. Watanabe, Sound quality comparison among high-quality vocoders by using re-synthesized speech, Acoustical Science and Technology 39 (3) (2018) 263–265. doi:10.1250/ast.39.263.
[34] H. Kawahara, M. Morise, Technical foundations of tandem-straight, a speech analysis, modification and synthesis framework, Sadhana 36 (5) (2011) 713–727. doi:10.1007/s12046-011-0043-3.
[35] M. Morise, Harvest: A high-performance fundamental frequency estimator from speech signals, in: Proc. Interspeech 2017, 2017, pp. 2321–2325. doi:10.21437/Interspeech.2017-68.
[36] M. Morise, Cheaptrick, a spectral envelope estimator for high-quality speech synthesis, Speech Communication 67 (2015) 1 – 7. doi:https://doi.org/10.1016/j.specom.2014.09.003.
[37] M. Morise, D4c, a band-aperiodicity estimator for high-quality speech synthesis, Speech Communication 84 (2016) 57 – 65. doi:https://doi.org/10.1016/j.specom.2016.09.001.
[38] M. Puckette, Max at seventeen, Computer Music Journal 26 (4) (2002) 31–43.
[39] C. d'Alessandro, Voice source parameters and prosodic analysis, in: S. S. et al. (Ed.), Method in Empirical Prosody Research, Walter de Gruyter, Berlin, New York, 2006, pp. 63–87.
[40] C. d'Alessandro, B. Doval, Voice quality modification using periodic-aperiodic decomposition and spectral processing of the

voice source signal, in: Proceedings of the 3rd ESCA International Workshop on Speech Synthesis, Jenolan Caves, Australia, 1998, pp. 277–282.
[41] B. Doval, C. d'Alessandro, N. Henrich Bernardoni, The spectrum of glottal flow models, Acta Acustica united with Acustica 92 (2006) 1026–1046.
[42] O. Perrotin, C. d'Alessandro, Vocal effort modification for singing synthesis, in: Annual Conference of the International Speech Communication Association (INTERSPEECH 2016), San Francisco, United States, 2016, pp. 1235–1239. doi:10.21437/Interspeech.2016-1096.
[43] G. Fant, A note on vocal tract size factors and non-uniform f-pattern scalings, STL-QPSR 7 (4) (1966) 22–30.
[44] Theremin world: Learn to play the theremin, http://www.thereminworld.com/Learn-to-Play, accessed: 2019-01-25.
[45] O. Perrotin, Singing with hands : chironomic interfaces for digital musical instruments, Theses, Université Paris Sud - Paris XI (Sep. 2015).
[46] M. Zbyszynski, M. Wright, A. Momeni, D. Cullen, Ten years of tablet musical interfaces at cnmat, in: Proceedings of the 7th Conference on New Interfaces for Musical Expression (NIME'07), New York, USA, 2007, pp. 100–105.
[47] C. d'Alessandro, L. Feugere, S. Le Beux, O. Perrotin, A. Rilliard, Drawing melodies: Evaluation of chironomic singing synthesis, JASA 135 (6) (2014) 3601–3612.
[48] O. Perrotin, C. d'Alessandro, Target acquisition vs. expressive motion: Dynamic pitch warping for intonation correction, ACM Transactions on Computer-Human Interactions 23 (3).
[49] [online, cited 2020-06-12][link].
[50] Session synthesis of singing, in: proceedings of the Stockholm Music Acoustics Conference (SMAC 1993), 1993, pp. 279–294.
[51] Synthesis of singing challenge, special session at interspeech 2007,, in: 8th Annual Conference of the International Speech Communication Association (Interspeech ISCA), 2007.
[52] L. Ardaillon, C. Chabot-Canet, A. Roebel, Expressive control of singing voice synthesis using musical contexts and a parametric f0 model, in: Interspeech 2016, 2016, pp. 1250–1254. doi:10.21437/Interspeech.2016-1317.
URL http://dx.doi.org/10.21437/Interspeech.2016-1317
[53] P. Y. Chan, M. Dong, G. X. H. Ho, H. Li, Seraphim: A wavetable synthesis system with 3d lip animation for real-time speech and singing applications on mobile platforms, in: Interspeech 2016, 2016, pp. 1225–1229. doi:10.21437/Interspeech.2016-484.
URL http://dx.doi.org/10.21437/Interspeech.2016-484
[54] E. del Blanco, I. Hernaez, E. Navas, X. Sarasola, D. Erro, Bertsokantari: a tts based singing synthesis system, in: Interspeech 2016, 2016, pp. 1240–1244. doi:10.21437/Interspeech.2016-1123.
URL http://dx.doi.org/10.21437/Interspeech.2016-1123
[55] M. Cotescu, Optimal unit stitching in a unit selection singing synthesis system, in: Interspeech 2016, 2016, pp. 1255–1259. doi:10.21437/Interspeech.2016-1390.
URL http://dx.doi.org/10.21437/Interspeech.2016-1390
[56] [online, cited 2020-06-12][link].
[57] [online, cited 2020-06-12][link].
[58] C. D'Alessandro, B. Doval, S. Delalez, W. Victor, R. Expert, Jouer avec les doubles artificiels de la voix: Cantor digitalis et Vokinesis.Conférence-concert, in: La voix à double tranchant, Voix et psychanalyse 2017, Solipsy, 2018, pp. 185–203.
URL https://hal.archives-ouvertes.fr/hal-02009009
[59] D. Abercrombie, Elements of General Phonetics, Edinburgh University Press, 1984.
[60] A. Schoenberg, Dreimal sieben gedichte aus albert girauds "pierrot lunaire" (1912).
[61] P. Hallé, Evidence for tone-specific activity of the sternohyoid muscle in modern standard chinese, Language and Speech 73 (1994) 103–124–1043. doi:10.1121/1.1531176.