VoixTenue: Exploring Real-Time Gestural Control of Vocal Synthesis on a Mobile Phone

Adrien Scazzola
Adrien.scazzola@edu.devinci.fr
École Supérieure d'Ingénieur Léonard de Vinci
De Vinci Higher Education
Paris, France

Xiao Xiao
xiao.xiao@devinci.fr
De Vinci Research Center & Institute for Future Technologies
De Vinci Higher Educations
Paris, France
MIT Media Laboratory
Cambridge, Massachusetts, USA

Figure 1: The two mappings of VoixTenue, Draw mode (left) and Gyro mode (right). In Draw mode, the output frequency is modulated by the vertical position of the user's finger on the phone's touchscreen. In Gyro mode, the pitch-axis tilt of the phone controls the frequency while the roll-axis tilt controls the volume of the output sound.

Abstract

We present VoixTenue, an interface exploring how a mobile phone's on-board sensing can be used to control pitch, dynamics, and intonation in real time for expressive vocal synthesis. VoixTenue supports two main interaction modalities: one where users draw and replay fingertip-drawn intonation curves for real-time pitch control, and another where the phone's orientation controls pitch and dynamics through inertial sensing. Using Pink Trombone as its synthesis engine, VoixTenue supports two modes for phonetic content: a vowel mode, in which users select a sustained vowel, and a phrase mode, in which a short English text can be entered, whose intonation is controlled by the user. We describe the system architecture and gestural mappings, and discuss potential use cases, including expressive performance and language learning.

Keywords

gestural interaction, vocal synthesis, intonation control, vocal prosody, speech-music interaction, mobile music interfaces

1 Introduction & Related Work

The voice is among the earliest and most direct forms of musical expression and has been explored extensively within NIME [12]. One recurring direction is performative vocal synthesis [6, 10], in which expressive parameters of a synthesized voice are controlled in real time, often by hand gestures ("chironomy") [11].

Early chironomic voice synthesis systems such as Glove-Talk I and II sought to synthesize speech from the ground up, simultaneously controlling articulatory configuration, phonemic content, and prosodic parameters in real-time through hand shapes and movements in space [8, 9]. While technically ambitious, such systems were difficult to learn and offered limited expressivity.

Subsequent work shifted toward optimizing gestural expressivity by restricting real-time control to a smaller set of salient parameters, particularly melodic shape, volume, and timing [6, 10, 19]. Graphic tablets were used to shape frequency trajectories and vocal effort, often focusing on sustained vowel synthesis [10, 11].

Later projects extended this approach to real-time re-synthesis of recorded vocal material, enabling gestural control of pitch, vocal effort, and syllabic articulation via various controllers, including graphic tablets [6], theremins [19], or mobile touch screens [18]. These systems demonstrated the expressive potential of gesturally controlled singing instruments across varied musical contexts [4]. These systems have also been explored for language learning, where gestural control externalizes prosodic contours and bypasses the physiological constraints of the natural voice [18, 19, 21].

Most chironomic performative vocal synthesis systems rely on a laptop or desktop computer for part of the synthesis or parameter mapping process. In [18], the mobile device served as a gestural interface, while sound synthesis was executed on a computer. In [5], vocal synthesis was rendered on a mobile phone, but

NIME '26, June 23–26, 2026, London, UK

parameter mapping remained computer-based. In contrast, we explore performative vocal synthesis fully contained on a smartphone, integrating sensing, mapping, and sound generation on a single device. Mobile music-making is well established within NIME [7, 15, 16]. Implementing performative vocal synthesis in this context raises questions about how a phone's on-board sensing capabilities can be mapped to expressive musical output.

We present VoixTenue (French for "held voice"), a browser-based mobile application built on Pink Trombone[1], an open-source articulatory synthesizer running in the browser. Pink Trombone implements a source–filter model in which users can control articulatory parameters such as tongue position, tongue diameter, lip constriction, and velum configuration to produce different phonemes, as well as expressive parameters including fundamental frequency, intensity, and voicedness. Interaction is typically performed via mouse or trackpad through clicking, dragging, and releasing gestures. VoixTenue adapts this synthesis engine for mobile use, enabling real-time control of vocal intonation and dynamics via touchscreen and inertial input. We describe the gestural mappings and the technical implementation, followed by applications and limitations.

2 Design & Interaction

We explored two mobile sensor mappings for real-time intonation control: Draw and Gyro mode (Figure 1). In Draw mode, inspired by [6, 10, 11], vertical finger position controls fundamental frequency (F0), while the horizontal axis is unused. Informal testing indicated that users interpret horizontal movement as temporal progression, so restricting control to the vertical axis preserves this mental model.

In Gyro mode, inspired by [3], tilt along the pitch axis controls F0, while tilt along the roll axis controls amplitude. Shaking the device introduces vibrato. This mode supports two usage patterns. For novice users, the device can be held in one hand and tilted primarily along a single axis to shape melody. For more experienced users, coordinated control of both axes allows simultaneous modulation of pitch and intensity.

A chromatic piano visualizer spanning E2–E5 is displayed in both modes. Each semitone is assigned equal visual height, and the current pitch output is indicated by a synchronized marker on the corresponding key.

VoixTenue supports two phonetic output modes, selectable via a collapsible navigation menu (Figure 2). In vowel mode, users select among the five cardinal vowels. In phrase mode, users enter a phrase, which is converted into editable SAMPA.

In Draw mode, sound is produced while the finger contacts the screen. In Gyro mode, sound is activated by pressing and holding a button and stops upon release. Initial calibration maps the current tilt to 233 Hz, selected as a comfortable speaking frequency.

In phrase mode, playback is triggered by finger contact or button press and follows a fixed timing based on typical speech rate [20]. Releasing interrupts playback, and pressing again restarts from the beginning. Holding beyond completion causes the phrase to loop.

Draw mode includes an optional trace retention feature. When enabled, traces can be replayed or cleared. By default, successive traces are appended sequentially. In "multi" mode, traces are stored separately and replayed in parallel (Figure 3).

[1]https://imaginary.github.io/pink-trombone/

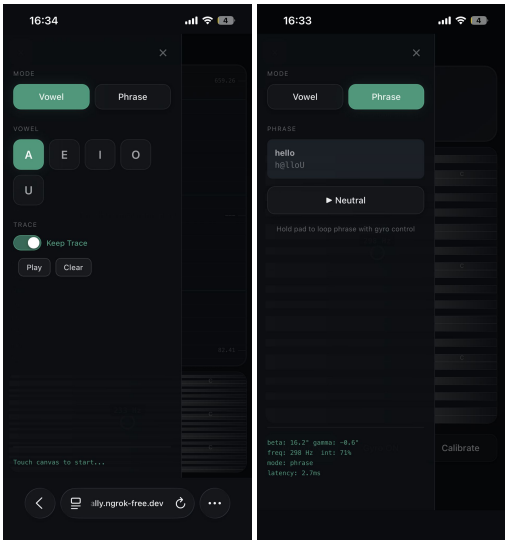

Figure 2: Expanded menu of control options for the Draw interface in vowel output mode (left) and Gyro interface in phrase output mode (right).

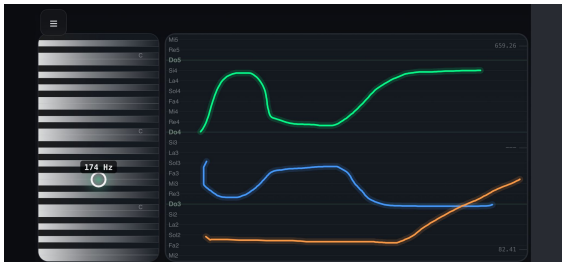

Figure 3: Multi mode of trace retention. These traces are played in parallel during playback.

3 Implementation

3.1 System Architecture

VoixTenue is implemented as a cross-platform browser-based application in JavaScript. The synthesis engine is a modified version of Pink Trombone [2] an articulatory synthesizer based on a source–filter model in which a glottal source excites a numerically simulated vocal tract. The DSP (glottal source and tract propagation) runs in an AudioWorklet on a high-priority audio thread, separate from the UI thread.

Pink Trombone exposes 19 parameters for articulator positions—e.g., tongue, lips—and source characteristics—e.g., intensity, voicedness (See Footnote 2 URL for full list of parameters and definitions). VoixTenue controls 8 of these. Intonation is controlled via F0 and intensity. Phonemic content is specified through parameters for the tongue and constriction of various parts of the vocal apparatus.

In the original interface, consonants are articulated manually. To reduce gestural load, VoixTenue uses an additive model: a consonantal constriction is temporarily superimposed on the current vowel configuration and automatically released after a fixed hold interval.

Real-time input (finger y-position or pitch/roll tilt) is mapped to pitch and intensity and transmitted with phonemic parameters

[2]https://github.com/zakaton/Pink-Trombone

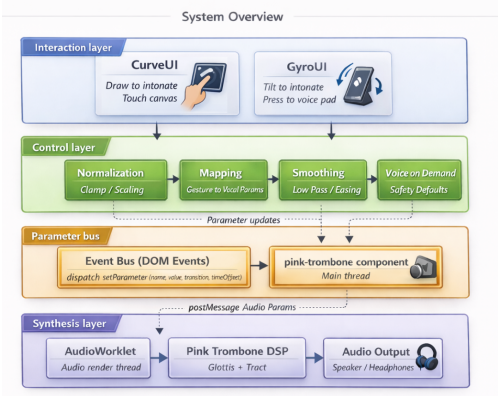

Figure 4: System overview of VoixTenue: CurveUI (draw) and GyroUI (tilt) map mobile gestures to Pink Trombone parameters via an event-driven bus, rendered in an AudioWorklet for real-time audio output

at 60 Hz (appx. 16 ms), a rate consistent with reported temporal sensitivities in auditory–motor interaction [13] and standard real-time system design.

3.2 Sensor Processing & Mapping

For Gyro mode, raw sensor signals require calibration and conditioning before being mapped to synthesis parameters.

Euler angles provided by the `DeviceOrientation` API (β for pitch, γ for roll) are first calibrated to compensate for the device's initial resting position. Upon activation of Gyro mode, neutral angles (β_0, γ_0) are recorded and relative angles computed as $\beta' = \beta - \beta_0$ and $\gamma' = \gamma - \gamma_0$.

$\beta' \in [-45°, +45°]$ is mapped to frequency in the range 82.41-659.26 Hz (E2–E5, exactly 3 octaves) using a logarithmic mapping. Let

$$n = \frac{\beta' + 45°}{90°}, \quad n \in [0, 1]. \tag{1}$$

The target frequency is defined as

$$F_{\text{target}} = F_{\min} \cdot 2^{n \cdot N_{\text{oct}}}, \tag{2}$$

where $F_{\min} = 82.41\,\text{Hz}$ and $N_{\text{oct}} = 3$. The neutral position ($\beta' = 0°$, $n = 0.5$) yields $f \approx 233\,\text{Hz}$ (A#3). This logarithmic mapping ensures perceptually uniform pitch spacing across the range, consistent with the chromatic piano visualizer used in the interface.

Similarly, $\gamma' \in [-30°, +30°]$ controls intensity via a bidirectional linear mapping centered at 0.7: left tilt ($\gamma' < 0$) increases intensity toward 1.0, while right tilt ($\gamma' \geq 0$) decreases it toward 0.2.

To reduce sensor jitter—inherent to MEMS-based inertial sensors while maintaining responsiveness, a first-order exponential moving average (EMA) is applied [2]. For frequency, smoothing is performed in the logarithmic domain to preserve musical intervals:

$$\log_2 f_t = (1 - \alpha)\, \log_2 f_{t-1} + \alpha\, \log_2 f_{\text{target}}, \tag{3}$$

with $\alpha = 0.25$ (i.e., smoothing factor $1 - \alpha = 0.75$). Linear interpolation for intensity uses the same coefficient. Vibrato is implemented using rotation rate data from the `DeviceMotion` API. The total angular speed is

$$\omega = \sqrt{\dot{\alpha}^2 + \dot{\beta}^2 + \dot{\gamma}^2}, \tag{4}$$

where $\dot{\alpha}$, $\dot{\beta}$, $\dot{\gamma}$ are the rotation rates (deg/s) about each axis. Over a 300 ms sliding window, when the peak angular speed $\omega_{\max}$ exceeds 200 deg/s, vibrato is applied:

$$f_{\text{mod}} = f\big(1 + A(t) \cdot \sin(2\pi f_v t)\big), \tag{5}$$

where the vibrato amplitude is

$$A(t) = d \cdot s(t), \quad d = 0.08, \quad s(t) = \min\left(\frac{\omega_{\max} - 200}{200},\ 1\right). \tag{6}$$

The vibrato frequency is $f_v = 6\,\text{Hz}$ (based on natural vocal vibrato rates documented by [14])). When rotation ceases, $s(t)$ decays geometrically by a factor of 0.94 per frame to prevent abrupt cutoff.

4 Applications

VoixTenue supports two primary application domains consistent with prior work: musical expression [6, 11, 19] and learning foreign language intonation [18, 21].

4.1 Musical expression

VoixTenue does not target conventional melodic repertoire, but emphasizes exploratory control of expressive contour. The characteristic synthetic timbre of Pink Trombone, along with the reduced parameter space, shifts the emphasis from technical precision to playful exploration of an externalized voice [1].

Draw mode enables explicit shaping of melodic contour through vertical F0 mapping. It does not currently support amplitude control, which limits dynamic nuance. Future iterations could introduce optional volume mapping to the horizontal finger position, allowing beginners to focus on pitch while more advanced users explore coordinated pitch and intensity control.

Gyro mode maps pitch and amplitude to two tilt dimensions. While this enables simultaneous control of melody and dynamics, accurate melodic rendering requires practice and remains less precise than drawing. The interface is thus better suited to continuous gestures such as glissandi and vibrato than to rhythmically exact melodic performance. In both modes, the piano keyboard displayed on the interface provides visual feedback of pitch.

4.2 Language learning

VoixTenue can be particularly suited to prosodic training. Prior work has shown that gestural control of F0 enables non-native speakers to produce perceptually distinct intonation contrasts beyond the constraints of their natural voice [18]. Drawing intonation contours externalizes mental representations of prosody and allows comparison with native models, supporting diagnosis of intonational troubles [21].

In Gyro mode, isotonic wrist tilt has been shown to be intuitive for novices and capable of producing recognizable stress patterns in speech resynthesis [3]. Compared to Draw mode, it places less emphasis on exact pitch matching and reduces reliance on visual attention.

Earlier classroom studies identified practical limitations when synthesis was computer-based [17]. A fully mobile implementation enables portable interaction and supports the possibility of longitudinal investigations into chironomic training for prosody acquisition. The availability of both vowel-only and

phrase modes further allows focused listening to the melodic structure of speech.

5 Conclusions & Future Work

We presented VoixTenue, a browser-based mobile system for performative vocal synthesis that runs entirely on a smartphone. The system implements two gestural mappings—touch drawing and inertial tilt—for real-time control of vocal intonation and dynamics using an articulatory synthesis engine. We outlined its design, implementation, applications, and limitations. This work demonstrates that performative vocal synthesis can be realized in a fully mobile form and supports further exploration of gesture-based prosody and voice interaction in portable settings. Future work will refine the interface for targeted applications, including expanded vowel inventories, improved phrase specification, and further optimization of consonant parameterization within Pink Trombone to enhance vocal quality.

6 Ethics Statement

This paper presents a technical prototype; testing was limited to the authors, no human-subject data were collected, and the application does not store or transmit personal data.

References

[1] Jeronimo Barbosa, Marcelo Mortensen Wanderley, and Stéphane Huot. 2017. Exploring playfulness in nime design: The case of live looping tools. In *NIME 2017-17th International Conference on New Interfaces for Musical Expression*. 87–92.

[2] Géry Casiez, Nicolas Roussel, and Daniel Vogel. 2012. 1€ Filter: A Simple Speed-based Low-pass Filter for Noisy Input in Interactive Systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. ACM, Austin, TX, USA, 2527–2530. https://doi.org/10.1145/2207676.2208639

[3] Delphine Charuau, Nathalie Henrich Bernardoni, Silvain Gerber, and Olivier Perrotin. 2025. Hand gesture realisation of contrastive focus in real-time whisper-to-speech synthesis: Investigating the transfer from implicit to explicit control of intonation. *Speech Communication* (2025), 103344.

[4] Christophe d'Alessandro, Xiao Xiao, Grégoire Locqueville, and Boris Doval. 2019. Borrowed voices. In *International Conference on New Interfaces for Musical Expression NIME'19*. 2–2.

[5] Nicolas d'Alessandro, Robert Pritchard, Johnty Wang, and Sidney Fels. 2011. Ubiquitous voice synthesis: interactive manipulation of speech and singing on mobile distributed platforms. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*. 335–340.

[6] Samuel Delalez and Christophe d'Alessandro. 2017. Vokinesis: syllabic control points for performative singing synthesis. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. 198–203.

[7] Georg Essl and Sang Won Lee. 2017. Mobile devices as musical instruments-state of the art and future prospects. In *International Symposium on Computer Music Multidisciplinary Research*. Springer, 525–539.

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

[9] S. S. Fels and G. E. Hinton. 1998. Glove-TalkII-a neural-network interface which maps gestures to parallel formant speech synthesizer controls. *IEEE Trans.on Neural Networks* 9, 1 (Jan 1998), 205–212. https://doi.org/10.1109/72.655042

[10] Lionel Feugère, Christophe d'Alessandro, and Boris Doval. 2013. Performative voice synthesis for edutainment in acoustic phonetics and singing: A case study using the "Cantor Digitalis". In *International Conference on Intelligent Technologies for Interactive Entertainment*. Springer, 169–178.

[11] Lionel Feugère, Christophe d'Alessandro, Boris Doval, and Olivier Perrotin. 2017. Cantor Digitalis: chironomic parametric synthesis of singing. *EURASIP Journal on Audio, Speech, and Music Processing* 2017, 1 (2017), 2.

[12] Rébecca Kleinberger, Nikhil Singh, Xiao Xiao, and Akito van Troyer. 2022. Voice at NIME: a Taxonomy of New Interfaces for Vocal Musical Expression. In *NIME 2022*. PubPub.

[13] Bruno H Repp and Yi-Huang Su. 2013. Sensorimotor synchronization: a review of recent research (2006–2012). *Psychonomic bulletin & review* 20, 3 (2013), 403–452.

[14] Johan Sundberg. 1987. *The Science of the Singing Voice*. Northern Illinois University Press, DeKalb, IL.

[15] Ge Wang, Georg Essl, Jeff Smith, Spencer Salazar, Perry R Cook, Rob Hamilton, Rebecca Fiebrink, Jonathan Berger, David Zhu, Mattias Ljungstrom, et al. 2009. Smule= Sonic Media: An Intersection of the Mobile, Musical, and Social.. In *ICMC*.

[16] Gil Weinberg, Andrew Beck, and Mark Godfrey. 2009. ZooZBeat: a gesture-based mobile music studio. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. 312–315.

[17] Xiao Xiao, Corinne Bonnet, Haohan Zhang, Nicolas Audibert, Barbara Kühnert, and Claire Pillot-Loiseau. 2024. Enseignement de l'intonation du français par une synthèse vocale contrôlée par le geste: étude de faisabilité. In *35èmes Journées d'Études sur la Parole (JEP 2024) 31ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2024) 26ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL 2024)*. ATALA & AFPC, 342–350.

[18] Xiao Xiao, Barbara Kuhnert, Nicolas Audibert, Grégoire Locqueville, Claire Pillot-Loiseau, Haohan Zhang, and Christophe d'Alessandro. 2023. Performative Vocal Synthesis for Foreign Language Intonation Practice. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–9.

[19] Xiao Xiao, Grégoire Locqueville, Christophe d'Alessandro, and Boris Doval. 2019. T-Voks: the singing and speaking theremin. In *NIME 2019 International Conference on New Interfaces for Musical Expression*. 110–115.

[20] Jiahong Yuan, Mark Liberman, and Christopher Cieri. 2006. Towards an integrated understanding of speaking rate in conversation.. In *Interspeech*. Pittsburgh, PA.

[21] H. Zhang, Y. Wu, X. Xiao, and C. Pillot-Loiseau. 2025. Gestural-Based Production of Ambiguous Mandarin Tones: A Multimodal Comparison of Native and Non-native Speakers. In *Proceedings of Speech Prosody 2026*. Accepted for publication.