Performative Vocal Synthesis for Foreign Language Intonation Practice

XIAO XIAO, Léonard de Vinci Pôle Universitaire, Research Center, France
BARBARA KUHNERT, Laboratoire de Phonétique et Phonologie UMR 7018, Sorbonne Nouvelle, France
NICOLAS AUDIBERT, Laboratoire de Phonétique et Phonologie UMR 7018, Sorbonne Nouvelle, France
GRÉGOIRE LOCQUEVILLE, Institut Jean Le Rond d'Alembert, Sorbonne Université, France
CLAIRE PILLOT-LOISEAU, Laboratoire de Phonétique et Phonologie UMR 7018, Sorbonne Nouvelle, France
HAOHAN ZHANG, Laboratoire de Phonétique et Phonologie UMR 7018, Sorbonne Nouvelle, France
CHRISTOPHE D'ALESSANDRO, Institut Jean Le Rond d'Alembert, Sorbonne Université, France

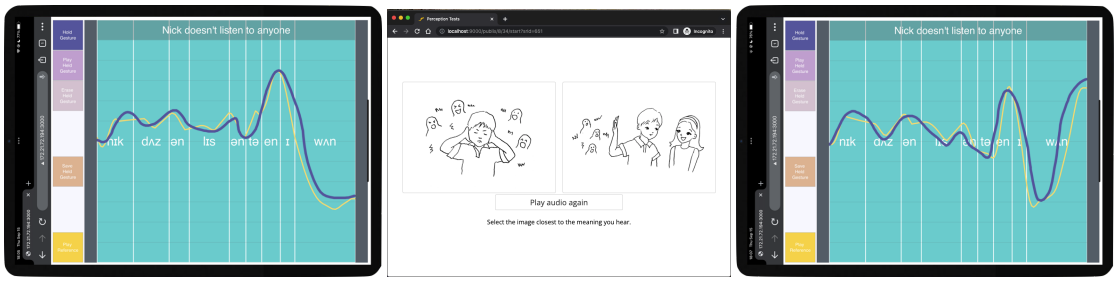

Fig. 1. (left, right) Gestural control interface, Gepeto, used in the experiment [44]. Subjects listened to a reference recording and had to imitate the intonation by drawing a curve on the touchscreen to control the production of a vocal synthesizer. Buttons on the left enable subjects to playback the reference, playback their gestural production, erase, or submit their gesture. User-drawn intonation curve are shown in purple. The yellow line is a visual guide depicting the frequency curve of the reference recording. Two different versions of the same phrase are shown. (center) Perceptual evaluation interface showing images that depict the two meanings.

Typical foreign language (L2) pronunciation training focuses mainly on individual sounds. Intonation, the patterns of pitch change across words or phrases is often neglected, despite its key role in word-level intelligibility and in the expression of attitudes and affect. This paper examines hand-controlled real-time vocal synthesis, known as Performative Vocal Synthesis (PVS), as an interaction technique for practicing L2 intonation in computer aided pronunciation training (CAPT).

We evaluate a tablet-based interface where users gesturally control the pitch of a pre-recorded utterance by drawing curves on the touchscreen. 24 subjects (12 French learners, 12 British controls) imitated English phrases with their voice and the interface. Results of an acoustic analysis and expert perceptive evaluation showed that learners' gestural imitations yielded more accurate results than vocal imitations of the fall-rise intonation pattern typically difficult for francophones, suggesting that PVS can help learners produce intonation patterns beyond the capabilities of their natural voice.

CCS Concepts: • **Human-centered computing** → **Empirical studies in interaction design**; **Gestural input**; Sound-based input / output; Touch screens.

Additional Key Words and Phrases: vocal synthesis, foreign language learning, intonation, gesture

Manuscript submitted to ACM

ACM Reference Format:
Xiao Xiao, Barbara Kuhnert, Nicolas Audibert, Grégoire Locqueville, Claire Pillot-Loiseau, Haohan Zhang, and Christophe D'Alessandro. 2018. Performative Vocal Synthesis for Foreign Language Intonation Practice. In *CHI 2023 submission.* ACM, New York, NY, USA, 12 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Foreign language (L2) pronunciation training focuses mainly on the component sounds of a language, both in the classroom and in Computer Aided Pronunciation Training (CAPT) [4, 31, 37]. CAPT systems typically analyze a user's production to identify mistakes and often employ vocal synthesis to present corrected pronunciations [2, 5, 24, 30]. A mistake is generally syllabic-level problem, either in individual sounds or stress, that causes a word to be misheard as another.

Our research focuses on the use of vocal synthesis in CAPT for the learning of intonation, the patterns of pitch change on words or longer sentences that give each language its distinct "music" [25, 31]. Though problems in intonation have been shown to affect intelligibility more than phonetic errors [29], intonation has received far less attention in L2 training than phonetics, both in classrooms and in interactive systems [31, 37].

A number of teaching strategies for intonation have been proposed, which aim to draw attention to changes in pitch (i.e. fundamental frequency, F0) using visualizations [4, 6, 12, 20, 33] or gestures [7, 28, 35]. Visualizations show learners the target shape of native productions and the places in which learners' productions differ from native models. Gestures have been used to kinesthetically reinforce some key features of vocal production, such as beat gestures for syllabic rhythm [28] or rise and fall movements for prosody [7].

While the use of visualizations and gestures has shown promise in improving learners' perception and productions of pitch change [7, 12], studies have shown that learning intonation requires not only the perception and production of pitch change but also the establishment of new mental categories for different types of pitch change [8, 21, 22]. Though melodic primitives of many languages have been described, simply repeating a canonical pattern is not sufficient for learning an intonation system [8, 25]. An understanding of the range of variation within a given category is also needed, as well as the freedom to produce a range of expressive variations. Achieving this poses substantial challenges for learners whose pitch control of the natural voice is limited by the existing pitch patterns in their mother tongue [13, 21].

This paper examines the potential of hand-controlled, real-time vocal synthesis, known as Performative Vocal Synthesis (PVS) [18, 27, 44, 45] as an interaction technique for computer-aided intonation training. Originally developed as a new type of musical instrument [18, 45], PVS enables the real-time gestural control of synthesized speech through the modulation of voice pitch, syllable timing, vocal effort, and vocal quality [27]. PVS gives the learner a "prothetic" voice, whose gestures for control are explicit and externalized, contrary to the natural voice. Toward using PVS for intonation training, this paper investigates the following questions:

- To what extent can L2 subjects produce comprehensible categorical contrasts in English intonation with their voice and with gestures?
- How do vocal and gestural productions differ for L2 subjects?
- How do the productions of L2 subjects differ from native productions?

We conducted a study with 12 French L2 learners of English and 12 native British control subjects using the PVS interface from [44], where users control the intonation of pre-recorded sentences by tracing frequency curves on a touchscreen with the help of a visual guide (Figure 1). A corpus of sentences was designed around pairs of phrases

with the same words whose meaning differ based on intonation, focusing on the contrast between the fall and fall-rise intonation patterns [8]. An acoustic analysis using General Additive Mixed Modeling [36, 41, 43] and an expert perceptual evaluation were performed. Results indicate that French learners' productions were significantly better in the gestural condition than both free reading and vocal imitation for the fall-rise intonation pattern, known to be difficult for native French speakers. The same difference was not found for native subjects. By providing learners a means to bypass the limitations of their natural voices, PVS can enable learners to experiment with a wide range of expressive variation when learning new intonation patterns.

2 BACKGROUND

We first describe past research on performative vocal synthesis in light of its potential for controlling speech intonation, followed by an overview of English intonation and the point of difficulty targeted by our study corpus. For a comprehensive review of CAPT systems in general, see [5].

2.1 Performative Vocal Synthesis & Intonation

Early systems for real-time gestural-controlled vocal synthesis, such as GloveTalk, attempted to modulate both phonemic and suprasegmental elements at the same time [16, 17]. Proficiency with these systems required many months of training to achieve bare minimum intelligibility.

For use within CAPT applications, we look to the model of performative vocal synthesis explored by [10, 14, 15, 18, 27, 44, 45], where hand-controlled[1] interfaces modify the fundamental frequency (F0) and other elements (e.g. timing and vocal effort) of a pre-recorded speech sample. A number of hand-controlled interfaces have been used in PVS systems, including the graphic tablet [14, 15, 18], the theremin [45], and the mobile touchscreen [44]. Songs in French, English, German, as well as poetry in Mandarin Chinese have been tested using hand-controlled interfaces in artistic performances [1, 11]. Visual guidance showing the frequency contour of a melody or intonation pattern can be presented for the graphic tablet and the touch-screen [10, 44].

Scientific studies on the hand control of intonation have only used French corpuses [14, 15, 44]. One study compared vocal and gestural imitations of reference phrases with untrained native subjects using a graphic-tablet interface that controlled the F0 of pitch-flattened recorded samples [14]. Similarity measures between the F0 curves of subject productions and reference recordings were computed, and a perceptual evaluation found the highest-scoring gestural contours to be indistinguishable from vocal productions. Another study with expert users of the graphic-tablet interface controlling the F0 and speed of synthesized sentences to express different emotions found that chironomic productions of fear, joy, sadness, and surprise were correctly identified more than 75% in a perceptual study, significantly higher than expressive intonations generated by statistical modeling [15]. A recent pilot study with both native and non-native speakers controlling the F0 and timing on a mobile tablet interface showed that subjects performed significantly better when a visual guide was present [44]. No significant differences were found between native and non-native subjects, but the authors reported that the non-native subjects were already advanced and had little problems vocally producing the target intonations in the study corpus.

[1]The term *Chironomy*, from the Greek *chiro*, meaning "of the hand" is employed by prior work to describe hand-controlled performative vocal synthesis

2.2 English Intonation & The Fall-Rise Tone

The present research investigates PVS through an English corpus that targets a known difficulty for the French L2 subjects. It adopts the descriptive system commonly known as the "British" school of English intonation analysis [8], widely favored for the teaching of intonation to L2 students in France [23, 29]. The British system divides the flow of speech into tone units, each comprising a nuclear syllable or tonic, on which major movements of pitch start. Directions of the pitch movements are called "tones," and the most frequently used are the fall, the rise, the fall-rise, and the rise-fall. Tone units most commonly correspond with clauses, a phrase including at least a subject and a verb, which may be a simple sentence or a part of a complex sentence [8]. In neutral conditions, there is a tendency for the tonic to be located toward the end of each tone unit, typically on the stressed syllable of the last lexical word. Pre-tonic tonal patterns are considered to be semantically less important. For example, a statement like "Ken is feeding the cat" would generally be pronounced as a single tone unit, and the tonic bearing the major pitch movement would be placed on "cat."

In British English, fall tones are commonly associated with finality and neutrality in declarative sentences. The fall-rise, on the other hand, is often used to signal some kind of implication (i.e. a "but..."), such as a contrast, reservation, warning or hesitation [40]. When used in a negative sentence, the fall-rise limits the scope of the negation so that it includes the nucleus but not the main verb [8]. In the example, "he didn't get $^{\backslash}$one credit," the simple statement with a falling contour means he got not a single one, i.e. none. Produced with a fall-rise, "he didn't get ^{V}one credit," the meaning ascribed by the listener is rather that he got not just one but many credits. Likewise, "he didn't feed the ^{V}cat" signals an implicit contrast. If made explicit the likely utterance would be, "he didn't feed the ^{V}cat but the $^{\backslash}$dog."

Attesting to the difficulty of acquiring the fall-rise tone, a study with advanced Finnish speakers of English showed that they did not use the fall-rise in ways that are assumed to be typical of spoken English [38]. A perception study of English intonation patterns with native Portuguese speakers showed that the subjects were unable to discern the implicit negation meaning of the fall-rise tone [9], which does not exist in Portuguese like in French. This result is consistent with the analysis of [32], which observes that an intonation pattern that exists in one language and not another is likely to pose problems for speakers of the latter language.

3 PRODUCTION EXPERIMENT

The experimental protocol was based on prior evaluations of PVS for intonation production [14, 44].

3.1 Corpus

4 pairs of phrases were used in the experiment, where each pair shared the same sequence of words but differed in whether the phrase final tonic was produced with a fall or fall-rise contour (Table 1). All phrases were pronounced by a native male English speaker from England, an instructor in English phonetics residing in Paris. The phrases were selected from a larger corpus of 12 phrase pairs based on an online perception test (See Figure 1, center). Selected phrases were ones where both meanings were perceived as intended by native British English speakers more than 75% of the time.

3.2 Subjects

24 subjects took part in the study. The 12 non-native subjects (10 female, 2 male, aged 18-30, average age 21.75) were L1 French speakers with intermediate level in English, recruited through their university. The 12 native subjects (5 female,

Table 1. The study corpus. Each phrase in the left column was recorded twice, with the fall and fall-rise intonation on the tonic, respectively. The tone-bearing part of each phrase (region of interest, ROI) is shown in bold. The ROI can be one or more syllables but excludes unvoiced consonants at the beginning at the end. The hints that appeared to subjects to prompt the target intonation in the free reading condition are shown in the center and right columns.

Phrase	Hint for Fall	Hint for fall-rise
Susan didn't get **one credi**t.	She got zero.	She got lots.
The food didn't taste **good**.	It tasted bad.	It tasted great.
Ken didn't feed the c**a**t.	He forgot to.	He fed the dog.
Nick doesn't listen to **anyone**.	He only follows his own ideas.	He only listens to his best friends.

6 males, 1 other, average age 28.9) were born and educated in England. They reside either in Paris or London and were recruited through word of mouth. Subjects were not paid for their participation.

3.3 Tasks

The Gepeto interface, presented in [44] was used in the PVS portion of the study. The interface consists of a mobile tablet (Figure 1), which controls the Voks vocal synthesizer running on a laptop computer. Drawing a curve in the control region of the tablet interface outputs a resynthesis of the current phrase from Voks. The horizontal axis determines the temporal position in the original sample to resynthesize, and the vertical position modulates the output pitch. It is evenly spaced on a semitones (ST) scale with a range of 24ST (2 octaves) calibrated around the study corpus (82Hz-330Hz). Phrases controlled by Gepeto are pitch-flattened text-to-speech (TTS) recordings of the corpus, generated using the OS X British English voice "Daniel." A visual guide appears for each phrase, showing the intonation curve of each reference phrase, scaled on a phoneme-level to the TTS sample.

The experiment comprised three stages. Subjects first recorded themselves reading each phrase based on their own interpretation. The text of the phrase was shown along with a drawing depicting the meaning and a text-based hint (Table 1). Next, subjects recorded vocal imitations of the reference phrases. Finally, subjects used the Gepeto interface, running on an iPad Air, to gesturally produce each phrase given the drawing, hint, an audio reference, and the intonation curve as a visual guide.

Four familiarization trials (2 per phrase) were given for the PVS task using a pair of phrases not in the experimental corpus. During the familiarization, the experimenter briefed subjects on the use of the interface, informed them of the difficulty in timing control (rhythm of the phrase), and instructed them to pay close attention to the melodic shape of sentence endings. Subjects were instructed that both the shape and speed of their gesture are important, that the visual guide is a visual representation of the reference recording, but that it is not necessary to follow it exactly to arrive at an acceptable gestural production.

Phrases appeared in random order in all parts of the study, with paired phrases next to each other. Following [14, 44] no limits were placed on the number of times references were played and the time subjects spent finding the pronunciation of each phrase. The entire study, including verbal instructions and the subject information survey took about 1 hour. Data was collected in Paris, in a sound isolated studio, and in London, in a quiet room. All audio was heard through monitor headphones. An AKG C414 XLS microphone was used in Paris, and a Blue Yeti microphone was used in London. A MacBook Air laptop computer was used for the study, which displayed the user interface for the experiment.

4 ACOUSTIC ANALYSIS

An acoustic analysis based on General Additive Mixed Modeling (GAMM) was performed on extracted F0 curves of subject productions to determine regions of significant difference based on phrase type, subject type, and condition [36, 41, 43]. GAMMs statistically model non-linear changes over time through the use of spline functions, while simultaneously accounting for subject and item-related variability. They identify not only whether a significance is present based on a parameter but also where differences occur within a plot of time-varying data. The mgcv and itsadug R libraries were used for our analysis [34, 39, 42]. Following suggestions from [36], a 2-part significance test was employed. First, a version of each model was created without the key parametric term and compared using ANOVA with the original model to find the overall significance of the parameter, with a significance level set to 0.05. For models that pass the overall significance test, plots were then made to show areas of significant difference.

4.1 Data Preparation

Analysis was conducted on the tone carrying phonemes at phrase final-positions of each stimuli, called Region of Interest (ROI), whose F0 curve determines the tone choice and meaning of the sentence (Table 1). The F0 curves of the reference and subjects' vocal recordings were extracted and manually corrected for octave jumps using Praat [3]. All vocal recordings and resynthesized .wavs of recorded gestures were phonemically segmented using WebMAUS [26], and the resulting TextGrids were manually verified and labeled for ROIs. The F0 of each reference recording in its ROI was extracted and resampled at 10-ms intervals. Based on their ROI boundaries, the F0 curves of subjects' vocal and gestural recordings were then aligned with their corresponding references, with time rescaled between 0 and 1.

4.2 Statistical Models

Three types of models were fitted using subsets of the data to answer the questions shown in Table 2. All models were fitted using Equation 1, which includes an ordered factor difference smooth to model a fixed-effect parameter, and random smooths to account for the variability across subjects and phrases. An additional term accounting for auto-correlation is not shown [36, 41].

$$F0 \sim s(time) + s(time, by = PARAMETER) + s(time, subject, bs = "fs", m = 1) + s(time, phrase, bs = "fs", m = 1) \quad (1)$$

Table 2. Summary of GAMMs created

Research Question	Data Used	Parameter
Where are significant differences between the fall & fall-rise curves?	Within a single condition and subject type	PhraseType (isFallrise)
Where are significant differences between natives & non-natives subjects?	Within a single condition and phrase type	IsNative
Where are significant differences between curves for vocal & gestural imitation?	within a single subject and phrase type	Condition (isGestural)

4.3 Results

4.3.1 Significant difference between fall & fall-rise curves. 6 models were made, one for each combination of Subject (Native or Non-native) and Condition (Free Reading, Vocal Imitation, Gestural Imitation). PhraseType ($isFallrise$ =

TRUE or *FALSE*) was the fixed parameter. An overall significant difference was found for the *IsFallrise* parameter across all conditions for native subjects (free reading: p=7.812e-04; vocal imitation: p=0.012; gestural imitation: 6.451e-09). For non-native subjects, an overall significance was not found for the free reading condition (p=0.165) but was found for vocal imitation (p=0.002) and gestural imitation (p=1.628e-07). See Figure 4 for plots of all six models.

In all three conditions of Native speaker productions, the expected divergence of the fall and fall-rise curves at the end of the ROI appears as a region of significant difference. The ROI-end difference is similar in length (from 0.8 to 1) for vocal and gestural imitation of native speakers and for gestural imitation of non-natives. It is much shorter in non-native vocal imitations and native free-readings.

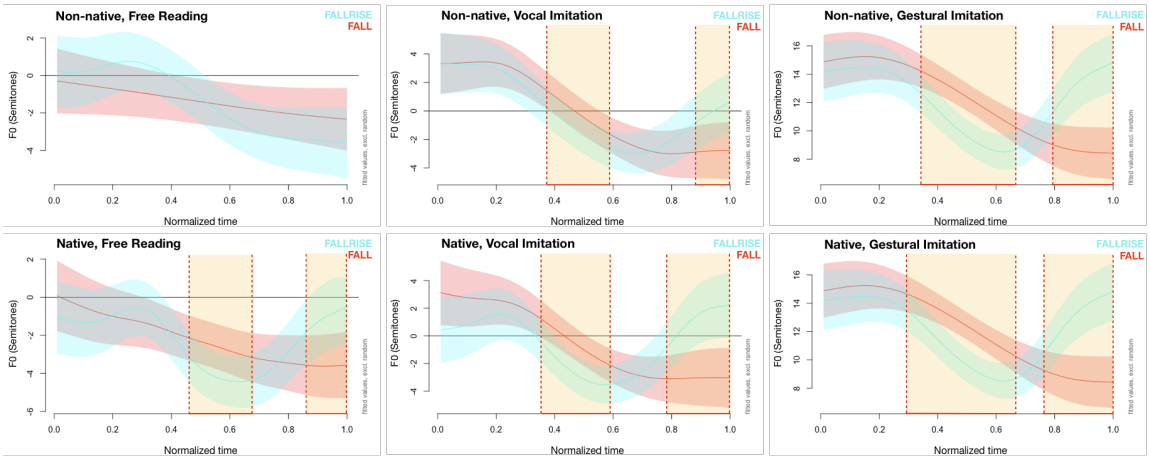

Fig. 2. GAMM smooth plots of fall and fall-rise F0 curves across the Region of Interest (ROI) with normalized time between 0 and 1. Areas of significant differences between two types of curves are highlighted in yellow. The IsFallrise parameter was not found to be significant for non-native free reading data.

4.3.2 Significant Difference between Natives & Non-Natives. A model was created for each combination of phrase type and condition, 6 in all, with Nativeness as the fixed parameter. For fall phrases, overall significance of the parameter was not found for free reading (p=0.497) and vocal imitation (p=0.860). While a significant difference was found for gestural imitation (p=0.026), inspection of a difference plot between the two curves showed that the maximum difference was around 1ST, lower than the threshold of perceptual salience for speech [46]. For fall-rise phrases, significant differences were found for free reading (p=0.001) and vocal imitation (p=3.689e-04), with differences of up to 4ST, but not for gestural imitation (p=0.137).

4.3.3 Significant Difference between Vocal & Gestural Imitation. Given the natural variation in free reading even for native speakers, we focused on vocal and gestural imitation for the modeling of cross-condition differences. 4 models were made for the fall and fall-rise data of the two types of subjects. Overall significance for the Condition parameter was found non-natives for both phrase types (fall-rise P=2.735e-14, fall p=0.003), but the maximum difference for fall phrases was less than 1.5ST, lower than perceptual salience [46]. For natives, no overall significance was found for either type of curve (fall-rise p=0.218, fall p=0.391).

5 EXPERT PERCEPTUAL EVALUATION

To determine to what extent acoustic differences influence categorical perception, a listening test was conducted with 5 expert evaluators, who used the forced-choice image selection interface (Figure 1, center) to select the meaning of the phrases they hear. Selection was based on listening to the entire phrase, not just the ROI. All raters are teachers of English phonetics at the university level and can be considered experts in English pronunciation.

Each evaluator listened to 576 stimuli collected from all subjects across all conditions. 384 stimuli were vocal productions (from the reading and imitation conditions), and 192 were resynthesized audio from the gestural condition. To prevent fatigue, stimuli were divided into 5 groups, 3 groups of 128 vocal stimuli, and 2 groups of 96 gestural stimuli. Stimuli in each group was presented in random order. Reviewers were instructed to listen to each group during a single sitting and to wait one day before listening to the next group. Vocal groups were presented first, and a notice was displayed before reviewers began listening to gestural groups indicating that the stimuli may sound unnatural.

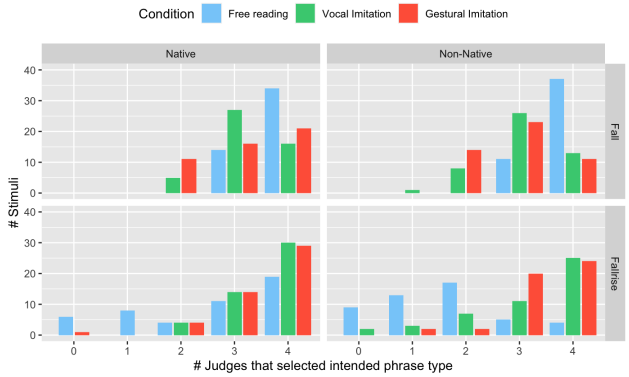

Fig. 3. Number of judges that selected the intended phrase type across subject and phrase types. For non-native speakers, gestural imitation resulted in higher numbers of correctly identified fall-rise productions than for free reading and vocal imitation.

5.1 Results

Fleiss Kappa for reviewer agreement was 0.414 ($p<0.0001$), indicating a moderate level of agreement [19]. Figure 3 shows the counts of stimuli by the number of judges that perceived it as intended. For fall phrases, natives and non-natives subjects performed similarly across conditions. Differences can be observed between natives and non-natives for fall-rise phrases. Non-natives had more free reading phrases not identifiable by a majority of judges than natives. They also had better identification for gestural imitation than vocal imitation productions, compared to natives who had similar results between the two conditions. Interestingly, natives also had a number of free reading phrases misidentified by most judges, indicating that natural fall-rise productions have some inherent ambiguity.

We also modeled the ROI contours based solely on judges' majority perception (3 or more agreed), following Equation 1, with the parameter "heard fall-rise" set as TRUE or FALSE based on whether more than half of the judges heard the fall-rise interpretation of the phrase. Three models were made, one with data from subject productions, one with only the stimuli with 80% or more agreement from the judges, and one with only the stimuli where judges were split 2 to 3.

The graph with all productions shows modeled splines that resemble the modeled fall and fall-rise contours for non-native vocal imitation, suggesting that non-native vocal imitation can achieve 60% perceptual recognition. However,

inspecting the plot of only the data with 60% judge agreement, the curves are almost indistinguishable, with both contours showing a very slight final rise. This suggests that the slightest final rise can be interpreted as an implicit negation. A bigger final rise in the modeled fall-rise curve is seen in productions with 80% or more judge agreement, indicating that a more exaggerated fall-rise production results in a higher rate of correct perception.

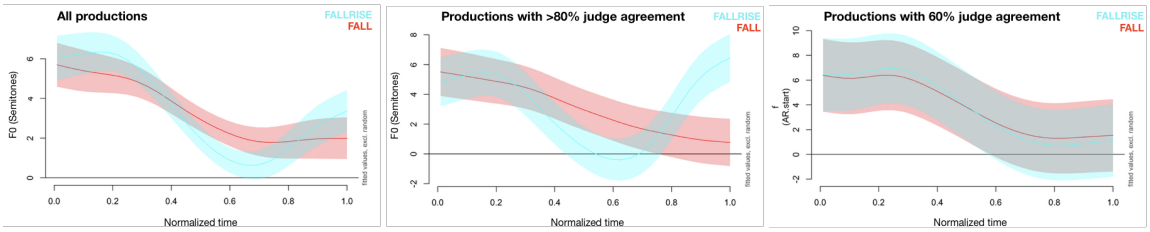

Fig. 4. Smooth plots based whether the majority of the judges heard the fall-rise intonation. (left) Model includes all productions, (center) Model includes only the productions with 80% or more judge agreement. (right) Model includes only the productions where 3 the 5 judges agreed.

6 DISCUSSION

Taken together, the acoustic and perceptual evaluations suggest that gestural imitation enables non-natives to produce more exaggerated versions of the fall-rise intonation than they can in vocal imitation, and that the more exaggerated intonation results in higher rates of perceptual identification. We also observe that non-native speakers do not seem to have an internalized sense of the fall-rise intonation with the meaning that it expresses. We now address implications and limitations of the study results.

6.1 Implications

Our results suggest that native British English speakers have a strong association between the F0 shape and its meaningful expression for the fall-rise tone, and they are freely able to dose the amount of expression by modulating the exaggeration of the fall-rise shape. In daily life, and in the free reading condition, the amount of fall-rise can be quite subtle, especially in the rising portion, which intentionally creates room for an ambiguity of interpretation. However, native speakers are capable of vocally imitating an exaggerated version of the tone with their natural voices.

In contrast, the French non-native subjects do not have an association between the fall-rise F0 shape and its expression, exemplified by the lack of rise in their free reading pronunciations. Their natural voice imitations also do not fully replicate the rising portion of the tone from the example. One reason for this could be their lack of their internal schema for the sound pattern, that causes them not to hear the full extent of the rise. Another reason may be that their natural voice lacks the dexterity to fully realize the target F0 shape. Guided gestural imitation addresses both possible sources of difficulty. The visual guide reinforces the auditory perception of the sound, and hand control of the synthesized voice allows more dexterity than the natural voice. With the help of performative vocal synthesis, the French subjects are able to use their ears to match the gestural production to the reference to an accuracy level on par with native subjects.

We imagine several ways that PVS can be integrated into a larger CAPT system to target the learning of intonation. With a gesture control interface as a modality to experiment with different intonation patterns, users could imitate examples from different speakers, to understand a range of possible productions. We could also imagine scenarios, where the user responds to virtual characters by producing the same phrase with different intonations. The system

can also include some sort of perceptual training where users must listen to and identify the intended meaning. In such examples, the user might still have an option to access the PVS interface to try out different ways of pronouncing the phrase. Finally, we can imagine different levels of difficulty, where the visual guide is initially provided but later removed.

6.2 Limitations

A potential critique of our study results is that non-native speakers were blindly tracing the visual guide in the gestural condition. While the visual guide was helpful and probably necessary for non-native speakers, the gestural pronunciation was influenced by both the position on the screen and the speed of movement. A prior study using the same interface [44], found that the speed of movement takes some practice to control, especially for large pitch movements that require changes in speed of gesture. The fact that no significant difference was found in the gestural fall-rise contours between native and non-native subjects, and that no significant difference was found between native gestural and vocal imitations of the fall-rise contour shows that non-native speakers were not just haphazardly following the visual guide. The similarity of non-native gestural fall-rise productions with native vocal imitations and the high rate of their identification on the perceptual evaluation suggests that the gestures were made purposefully, with the goal of replicating the reference.

One might also wonder whether aspects of the productions outside of the tone-carrying ROI could influence the perception of meaning, for instance, the tone and timing of the pre-nucleus portion of the phrase. This influence is possible, especially in the ambiguous cases identified by the perception test. However, the marked difference between the two contours in the GAMM model based only on majority judge rating for stimuli with clear consensus among judges, shows that a fully-realized fall-rise contour has a high importance in determining perception.

One limitation of our study is that we are not able to identify whether learners' PVS results were due more to the gestural control or to the visual guide. While researchers generally seek to isolate separate sources of influence, we may argue that it is precisely its multimodality that gives PVS its advantage. The gestural control gives learners extra dexterity and the ability to try out intonation patterns that are difficult with the voice. The visual guide shows learners an approximation, to reduce the frustration of blindly searching for the right pitch contour, and the challenge of timing control forces subjects to listen and compare their gesture with the reference. Taken together, this interaction engages auditory, visual, and gestural modalities, and forces the learner to listen carefully to differences between the reference and their own production.

7 CONCLUSIONS AND FUTURE WORK

This paper demonstrates that performative vocal synthesis enables non-native speakers to produce unfamiliar intonation patterns beyond the limitations in their natural perception and production. It opens new questions on how the long term use of PVS may improve the perception of intonation, and how productions with chironomy can transfer to improvements in the natural voice. Future work will explore the effects of long-term PVS usage and how PVS may be used for learning tonal languages, like Mandarin Chinese.

8 ACKNOWLEDGMENTS

Withheld for review.

REFERENCES

[1] 2022. Guthman Musical Instrument Competition, 2022 Competition. https://guthman.gatech.edu/2022-competition. Accessed: 2022-09-15.

[2] Pierre Badin, Atef Ben Youssef, Gérard Bailly, Frédéric Elisei, and Thomas Hueber. 2010. Visual articulatory feedback for phonetic correction in second language learning. In *L2SW, Workshop on" Second Language Studies: Acquisition, Learning, Education and Technology*. P1–10.

[3] Paul Boersma and David Weenink. 1992-2022. Praat: doing phonetics by computer [Computer program]. Version 6.1.08, retrieved 5 December 2019 from http://www.praat.org.

[4] Elena Boitsova, Evgeny Pyshkin, Yasuta Takako, Natalia Bogach, Iurii Lezhenin, Anton Lamtev, and Vadim Diachkov. 2018. StudyIntonation courseware kit for EFL prosody teaching. In *Proceedings of the 9th International Conference on Speech Prosody*. 413–417.

[5] Yaohua Bu, Tianyi Ma, Weijun Li, Hang Zhou, Jia Jia, Shengqi Chen, Kaiyuan Xu, Dachuan Shi, Haozhe Wu, Zhihan Yang, et al. 2021. PTeacher: a Computer-Aided Personalized Pronunciation Training System with Exaggerated Audio-Visual Corrective Feedback. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.

[6] Dorothy Chun. 1998. Signal analysis software for teaching discourse intonation. *Language Learning & Technology* 2, 1 (1998), 74–93.

[7] Cristina Crison, Daniel Romero, Joaquín Romero, and Rovira i Virgili. 2018. The practical application of hand gestures as a means of improving English intonation. In *Proc. ISAPh 2018 International Symposium on Applied Phonetics*. 45–50.

[8] Alan Cruttenden et al. 1997. *Intonation*. Cambridge University Press.

[9] Madalena Cruz-Ferreira. 1984. Perception and interpretation of non-native intonation patterns. In *Proceedings of the tenth International Congress of Phonetic Sciences*. De Gruyter Mouton, 565–569.

[10] Christophe d'Alessandro, Lionel Feugère, Sylvain Le Beux, Olivier Perrotin, and Albert Rilliard. 2014. Drawing melodies: Evaluation of chironomic singing synthesis. *The Journal of the Acoustical Society of America* 135, 6 (2014), 3601–3612.

[11] 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.

[12] Kees De Bot. 1983. Visual feedback of intonation I: Effectiveness and induced practice behavior. *Language and speech* 26, 4 (1983), 331–350.

[13] Tracey M Derwing and Marian J Rossiter. 2002. ESL learners' perceptions of their pronunciation needs and strategies. *System* 30, 2 (2002), 155–166.

[14] Christophe d'Alessandro, Albert Rilliard, and Sylvain Le Beux. 2011. Chironomic stylization of intonation. *The Journal of the Acoustical Society of America* 129, 3 (2011), 1594–1604.

[15] Marc Evrard, Samuel Delalez, Christophe d'Alessandro, and Albert Rilliard. 2015. Comparison of chironomic stylization versus statistical modeling of prosody for expressive speech synthesis. In *Sixteenth Annual Conference of the International Speech Communication Association*.

[16] 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.

[17] S. Sydney Fels and Geoffrey E. Hinton. 1998. Glove-Talk II-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

[18] 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.

[19] Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin* 76, 5 (1971), 378.

[20] Atsushi Fujimori, Noriko Yoshimura, and Noriko Yamane. 2015. The development of visual CALL materials for learning L2 English prosody. In *Conference proceedings. ICT for language learning*. libreriauniversitaria. it Edizioni, 249.

[21] Abbas Pourhossein Gilakjani and Mohammad Reza Ahmadi. 2011. Why Is Pronunciation So Difficult to Learn?. *English language teaching* 4, 3 (2011), 74–83.

[22] Pierre A Hallé, Yueh-Chin Chang, and Catherine T Best. 2004. Identification and discrimination of Mandarin Chinese tones by Mandarin Chinese vs. French listeners. *Journal of phonetics* 32, 3 (2004), 395–421.

[23] Sophie Herment and Anne Tortel. 2021. The intonation contour of non-finality revisited: implications for EFL teaching. In *English pronunciation instruction: Research-based Insights, Applied Linguisics series*, Anastazija Kirkova-Naskova, Alice Henderson, and Jonás Fouz-González (Eds.). John Benjamins, Amsterdam, Netherlands, 175–195.

[24] Rebecca Hincks. 2003. Speech technologies for pronunciation feedback and evaluation. *ReCALL* 15, 1 (2003), 3–20.

[25] Daniel Hirst and Albert Di Cristo. 1998. Intonation systems. *A survey of Twenty Languages* (1998).

[26] Thomas Kisler, Uwe Reichel, and Florian Schiel. 2017. Multilingual processing of speech via web services. *Computer Speech & Language* 45 (2017), 326–347.

[27] Grégoire Locqueville, Christophe d'Alessandro, Samuel Delalez, Boris Doval, and Xiao Xiao. 2020. Voks: Digital instruments for chironomic control of voice samples. *Speech Communication* 125 (2020), 97–113.

[28] Steven G McCafferty. 2006. Gesture and the materialization of second language prosody. (2006).

[29] Murray J Munro and Tracey M Derwing. 1999. Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language learning* 49 (1999), 285–310.

[30] Yishuang Ning, Zhiyong Wu, Jia Jia, Fanbo Meng, Helen Meng, and Lianhong Cai. 2015. HMM-based emphatic speech synthesis for corrective feedback in computer-aided pronunciation training. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4934–4938.

[31] Martha C Pennington. 1999. Computer-aided pronunciation pedagogy: Promise, limitations, directions. *Computer assisted language learning* 12, 5 (1999), 427–440.

[32] Brechtje Post, Mariapaola d'Imperio, and Carlos Gussenhoven. 2007. Fine phonetic detail and intonational meaning. In *International Congress of Phonetic Science (ICPhS)*. 191–196.

[33] Evgeny Pyshkin, John Blake, Anton Lamtev, Iurii Lezhenin, Artyom Zhuikov, and Natalia Bogach. 2019. Prosody training mobile application: Early design assessment and lessons learned. In *2019 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, Vol. 2. IEEE, 735–740.

[34] R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

[35] Tetyana Smotrova. 2017. Making pronunciation visible: Gesture in teaching pronunciation. *Tesol Quarterly* 51, 1 (2017), 59–89.

[36] Marton Soskuthy. 2021. Evaluating generalised additive mixed modelling strategies for dynamic speech analysis. *Journal of Phonetics* 84 (2021), 101017.

[37] Ron I Thomson and Tracey M Derwing. 2015. The effectiveness of L2 pronunciation instruction: A narrative review. *Applied Linguistics* 36, 3 (2015), 326–344.

[38] Juhani Toivanen. 2007. Fall-rise intonation usage in Finnish English second language discourse. *Proceedings of Fonetik 2007, TMH-QPSR, 50 (1)* (2007), 85–88.

[39] Jacolien van Rij, Martijn Wieling, R. Harald Baayen, and Hedderik van Rijn. 2020. itsadug: Interpreting Time Series and Autocorrelated Data Using GAMMs. R package version 2.4.

[40] Gregory Ward and Julia Hirschberg. 1985. Implicating uncertainty: The pragmatics of fall-rise intonation. *Language* (1985), 747–776.

[41] Martijn Wieling. 2018. Analyzing dynamic phonetic data using generalized additive mixed modeling: A tutorial focusing on articulatory differences between L1 and L2 speakers of English. *Journal of Phonetics* 70 (2018), 86–116.

[42] S. N. Wood. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)* 73, 1 (2011), 3–36.

[43] S. N. Wood. 2017. *Generalized Additive Models: An Introduction with R* (2 ed.). Chapman and Hall/CRC.

[44] Xiao Xiao, Nicolas Audibert, Grégoire Locqueville, Christophe d'Alessandro, Barbara Kuhnert, and Claire Pillot-Loiseau. 2021. Prosodic Disambiguation Using Chironomic Stylization of Intonation with Native and Non-Native Speakers. In *Interspeech 2021*. ISCA, 516–520.

[45] 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.

[46] Johan 't Hart. 1981. Differential sensitivity to pitch distance, particularly in speech. *The Journal of the Acoustical Society of America* 69, 3 (1981), 811–821.