Pitch visualization for prosody acquisition

Online study with English-speaking learners of Italian

Catherine Scanlon
cscanlon@ucsb.edu

Antón de la Fuente
cadelafuente@ucsb.edu

Department of Linguistics
University of California, Santa Barbara
USA

Abstract: We have developed a simple online app for L2 prosody practice, using pitch traces to provide learners with a visualization of native speaker intonation as well as feedback on how their own productions compare. To test the effectiveness of this training method, we are conducting a randomized experiment investigating whether visual representations of the target intonation and the learner's own production provide additional benefits over simple auditory feedback. Learning is assessed at multiple time-points using a DTW-based similarity measure as well as a native speaker perception task. Pilot testing has confirmed that the pitch plotting system is robust and the training is enjoyable and engaging for participants. Data collection for the randomized experiment is not yet complete.

Keywords-prosody, language teaching, computer-assisted prosody training, Italian

I. INTRODUCTION

Although prosody varies substantially between languages and is important for successful communication, targeted prosody training is rare in second language instruction. To help fill this gap, we have developed a simple web application that plots the F_0 (pitch) contour of a target recording and then superimposes an aligned version of the learner's pitch contour, allowing them to visually inspect how their production compares to the model. This method of computer-assisted prosody training is language-independent and allows students to practice online. We seek to answer the question "Does F_0 visualization aid L2 acquisition of prosody?". To address this question, we are conducting a randomized experiment testing our web training with and without pitch visualization. Our hypothesis is that learners who see the visualization will demonstrate greater learning than those who do not.

For this study, we focus specifically on the acquisition of Italian prosody for polar questions versus statements by English-speaking learners of Italian. This contrast is particularly important because polar (yes/no) questions in Italian typically have the same syntax as statements and are distinguished by prosody alone [1, 2, 3]. In addition, the intonation patterns of statements and questions in Italian are substantially different from English.

This is also a topic for which pedagogical resources are not readily available. Materials aimed at English-speaking learners of Italian generally explain that yes/no questions are formed by changing the intonation, and they often state that the pitch goes up at the end for questions. For example, three of the top five results of a Google search for "how to ask a yes no question in Italian" instruct the learner to use rising intonation at the end of the sentence, without including any audio examples. However, such explanations are not very informative and may even be misleading for learners. Italian question intonation is substantially different from English, and polar questions often do not end in a

SCANLON, Catherine; LA FUENTE, Antón de. Pitch visualization for prosody acquisition: online study with english-speaking learners of italian. In: CONGRESSO BRASILEIRO DE PROSÓDIA, 2, 2022, Florianópolis. **Anais [...]** . Florianópolis: Lbass, 2023. v. 1, p. 23-33. Disponível em: http://www.periodicos.letras.ufmg.br/index.php/anais_coloquio/index.

rise. Moreover, even for those that do have a rise, it is often quite small, with other parts of the contour contributing more to the question intonation. Our method allows learners to see the pitch contour and visually compare their own production.

II. Literature Review

A. Previous research on F0 visualization for prosody learning

Visual feedback in the form of F0 traces have been found to be helpful in L2 intonation teaching [4, 5, 6]. Early work by De Bot (1983) using real-time F0 traces reported that visual feedback was more effective than simple auditory feedback for Dutch learners of English to imitate English intonation [4]. Hardison (2004) had English learners of French record their production and then as this recording was played back to them their pitch trace was drawn in real time [6]. The results of this study indicated that this type of visual feedback was effective at helping learners generalize to novel sentences in a post-test condition. Both Hardison (2004) and De Bot (1983) used native speaker raters to measure the efficacy of their trainings.

B. Italian statement and question intonation

Across regions of Italy, broad focus statements are consistently characterized by a falling nuclear pitch accent, transcribed as H+L* in the ToBI system. Narrow focus (semantic or contrastive) can greatly change the pitch contour, such that there is a large peak on the focused constituent [1, 2, 3]. For the sake of simplicity and consistency, and because narrow and/or contrastive focus would require additional context to set up, for this study we use broad focus statements.

In contrast to broad focus statements, the intonation of polar questions in Italian is much more variable. A common finding is that polar questions in Northern and Central varieties are characterized by a terminal rise (low nuclear accent followed by a rising phrase accent) whereas polar questions in Southern varieties have a rise on the nuclear accented syllable, followed by a fall [2]. Other studies question these findings, suggesting that the crucial factor is not geographical location but rather aspects of the communicative context and task such as whether the question is read or spontaneous [7]. Corpus studies have found that, even in Northern varieties, rising intonation may be present in only a minority of polar questions [8]. Crucially for our purposes, even when Italian questions end with a rise, the contour is different from that of the equivalent question in English. For this study, we use information-seeking polar questions, as confirmatory polar questions may have markedly different melodies [3].

Finally, some research suggests that analysis in terms of pitch accents and their placement may not capture all the relevant cues to statement versus question modality in Italian, or even the most important ones. Particularly in certain varieties such as Neapolitan, in which narrow focus statements and broad focus questions have similar contours, other features such as contour shape (concave versus convex) may be equally or more important [9]. The training we propose in this experiment does not depend on a particular theoretical analysis of Italian intonation in terms of pitch accents or any other particular feature.

III. METHODS

C. *Overall experimental design*

We employ a randomized design with a treatment (visualization) group and a control group; each participant is randomly assigned to a group as they sign up. All participants complete the training, but only those in the visualization group are shown the pitch plots. Learning is assessed at three time points: immediate imitation in the second training session, generalization at the end of the second training session, and delayed generalization a week later.

D. *Dependent variables and detailed hypotheses*

We use two methods to measure improvement: an automatic similarity measure between model and learner pitch contours using dynamic time warping (DTW), following Rilliard and colleagues (2011) [10], and a native speaker forced-choice listening task. For the DTW similarity measurement, we first register syllable boundaries and use these to align the contours using DTW. Then the Hermes similarity measure (the weighted correlation between two F_0 contours) is computed. If participants produce significantly more accurate contours (i.e., the similarity between the model and learner sentences is greater) after being exposed to visual feedback conditions, then we have some evidence to believe that the visualizations are useful for L2 prosody acquisition.

Although it is an easy and automated way of assessing intonation, DTW-similarity says nothing about how utterances are perceived by speakers of the target language. Thus, we also test whether native speakers of Italian can perceive an improvement due to the training. Native speakers are played pairs of recordings and have to indicate which sentence is from after the training. A higher proportion of correct classifications by the native speakers indicates greater learning. If the native speakers can discriminate which production is from the post-training condition more reliably for the participants in the visual condition over those in the auditory condition, then we have evidence that the visual condition contributes to the acquisition of communicative features of L2 prosody.

According to our hypothesis that pitch visualization aids prosody acquisition, we expect significantly higher proportions of correct classification for recordings from the visualization group versus the control group, for all time points. Likewise we expect the DTW analysis to show significantly higher similarity between the learner and model pitch contours for learners in the visualization group.

We also expect that the proportion of correct classifications by the listeners, as well as the DTW similarity scores, to be highest for the immediate imitation recordings, followed by the immediate generalization recordings, and finally the recordings made a week later.

E. Web app back end: automatic pitch plotting

We developed a system that automatically plots a learner's pitch against a model utterance, using Praat's audio processing methods [11] implemented in Python with Parselmouth [12].

Because we cannot control the learners' recording environment or equipment, we had to pack in quite a bit of pre-processing to each audio sample. These steps also help avoid plotting mouse clicks at the start of the recording and other such noise as part of the intonation contour. After recording, each audio sample is processed into a Praat Sound object. Following Arias and colleagues (2010) [13] a low pass filter with a cutoff frequency of 600Hz is applied, helping remove transient background noise, which is more likely to be detected as high F0 samples in the signal. After filtering, silent intervals are removed using Praat's silence detection tool, which evaluates the intensity of the signal and treats as a silence any interval within the silence threshold of -25 decibels that is longer than 0.2 seconds (to exclude plosive closures and terminal vowel intervals that are devoiced due to creak). Additionally, any interval that does not fall within the intensity silence threshold but is shorter than 0.1 seconds is also considered silent. The tool then removes all silent intervals and joins the rest of the signal together.

After silences are removed, a pitch object is created from the sound object, and instances of pitch doubling are halved. The signal is then smoothed using a bandwidth of 10 Hz. Lastly, a discrete array of pitch samples and an array of their corresponding timings are extracted, all leading and trailing zeroes are removed from this array, and the corresponding sample timing array is adjusted as necessary so that its first sample equals 0.

Next, the system aligns the learner pitch contour to the reference contour. Both contours follow the same preprocessing steps above, and once these have been performed the learner pitch is time warped and normalized to align with the reference. Warping is done by taking the sample timing array (x-axis in the plot) and dividing each sample by the ratio of the duration of the learner utterance to the duration of the reference utterance. After the time warping, the mean pitches of both arrays are aligned (y-axis) for ease in visual comparison. The difference between the mean of the reference pitch array and the learner pitch array is subtracted from each pitch sample in the learner array. This makes it so that when they are plotted, the learner array is overlaid on the reference array. Mean pitch alignment, like the simple time warping, makes up for absolute differences in pitch range while preserving relative pitch relations in the contour. This should make it easier for untrained L2 learners exploring this tool to understand the differences between both traces, since the differences in the trace path are apparent.

F. Participants

Participants are recruited online in Facebook groups and forums dedicated to learning Italian. The requirements are to be fluent in English and to be learning Italian. Participants are told that the experiment aims to test a training method for Italian intonation, but they do not know that the key variable is visualization (i.e. those in the audio only group should not know they are missing anything). We aim to recruit 100 participants.

G. Experimental groups

Participants are divided into two groups. Those in the feedback group complete the training exactly as described below. Those in the control group complete the same training, except that they do not see any pitch visualizations. They do not see the representation of the native speaker sentences, and they do not receive visual feedback on their own productions. Other than that, the training is the same, and they spend the same amount of time on the training as the other group does.

H. Stimuli

We constructed four sets of seven sentences each, using only voiced phonemes. Within each set, all the sentences have the same syntactic structure and are matched word-by-word for number of syllables and stress placement. In addition, we avoided using a word more than once (with the exception of articles and copulas), and we tried to use words that would be familiar to lower intermediate students of Italian. The sentences were all checked by professors of Italian.

An additional 8 sentences to fill out the training were developed in the same way (using only sonorant sounds and matching the syntactic structures already used), except that they were not matched for the number of syllables and stress placement.

Within each of the four matched sets, the sentences are randomized across the seven stages of the experiment shown below. This is important so that the native Italian listeners in the evaluation phase of the experiment are not able to tell which stage of the experiment a recording came from based on its lexical content.

In order to provide us with easily comparable intonational melodies we had a single speaker record all the sentences. Italian professor Valentina Padula recorded the sentences in consultation with the researchers, who checked to make sure they have roughly similar prosody and decent sound quality.

I. Experiment structure

Participants complete three online sessions, one week apart, as follows:

Session 1:

1. Pre-training recordings (baseline): 4 pairs of sentences
2. Training: 8 new pairs of sentences, each repeated twice
3. Post-training recordings: 4 new pairs of sentences

Session 2:

4. Pre-training recordings: 4 new pairs of sentences
5. Training: 8 new pairs of sentences, each repeated twice
6. Post-training recordings: 4 new pairs of sentences

Session 3:

7. Follow-up recordings: 4 new pairs of sentences

In addition, Session 1 begins with a brief demographic questionnaire (age, gender, language background, Italian level), and Session 3 ends with an open-ended survey about learner experience. The experiment is hosted on our own website, https://prosody.delafuentealvarez.com/. Participants record directly in the browser. The surveys are conducted using Google Forms.

Within the training sections, the order of sentences is constant across participants, except for the within-set randomization explained above. The sentences in the non-training sections (the baseline and generalization recordings where they are simply reading without hearing a native speaker) are randomized, except that statement-question pairs are kept together.

Statement-question pairs are presented sequentially to promote attending to the statement-question intonational contrast. For each pair of sentences, the statement is presented first, followed by the question. For the non-training trials (pre-training, post-training, and follow-up), participants simply record themselves reading the sentence out loud. For the training trials, they first listen to a native speaker recording of the sentence. Next, the participant records themself saying the sentence. Finally the participant listens again to their own recording and the model recording. In the training section, there are two trials for each sentence. (So the order would be: S1 S1 Q1 Q1 S2 S2 Q2 Q2 …)

Participants in the visualization condition see the native speaker pitch plot during the first part of each trial (i.e. while listening to the native speaker recording and while recording themselves), as shown in Figure 1. In the second part of the trial (while re-listening to themselves and the native speaker) they see the combined pitch plot showing their own pitch contour superimposed on the native speaker contour, as shown in Figure 2.

Figure 1: Screenshot from the first part of a training trial, visualization condition, for the statement *La balena é enorme?* ‘Is the whale huge?’

Figure 2: Screenshot from the second part of a training trial, visualization condition, for the question *La balena é enorme?* 'Is the whale huge?'

J. Native speaker listening task

The learner recordings from the four sets of matched sentences will be used as stimuli in a binary listening discrimination task to be completed by native speakers of Italian. The recordings will be presented in pairs, and the task for the listener is to indicate which recording is from after the prosody training.

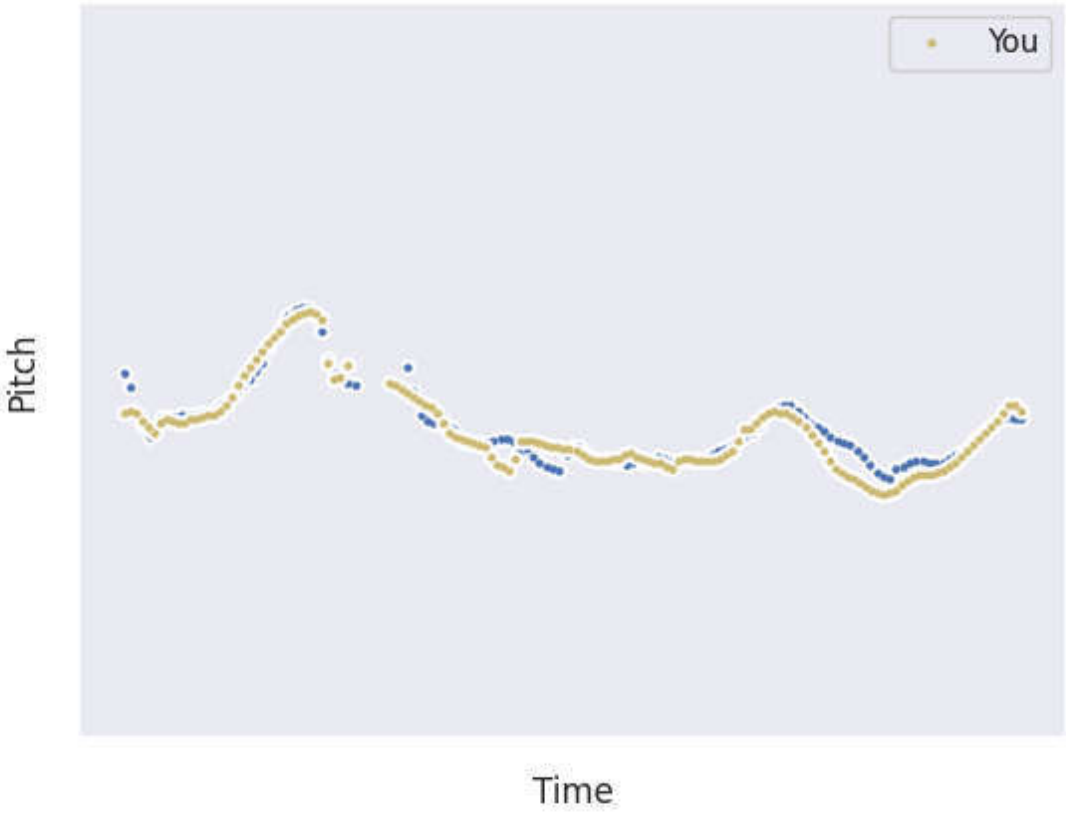
Each pre-test recordings from Session 1 is used as a "before" recording, to be compared with three "after" recordings from the same matched set: the second trial recording from Session 2 (produced immediately after listening to the native speaker recording a second time), the post-trial recording from Session 2, and the follow-up recording from Session 3. Recall that the order of the sentences within each set is randomized. Thus, although the listeners will be comparing pairs of sentences with different words, there will be no pattern as to which sentences are from "before" or "after" conditions. This randomization should also control for segmental and lexical effects.

Each learner produces one set of four "before" recordings and three sets of four "after" recordings, for a total of 12 comparisons per learner. Each comparison will be rated by five different listeners. Raters cannot listen to the same "before" recording more than once. Thus, to have each recording listened to 5 times, we will employ 15 raters, who each listen to four recording pairs per speaker. This should take about an hour and a half, which will be broken into several sessions to avoid fatigue. The order of the recordings will be randomized, both within and between comparison pairs.

IV. Results

We do not yet have data from the randomized experiment. However, we conducted a pilot study with 14 participants from the 6th quarter of Italian at UCSB, which allowed us to make substantial improvements to the website and the pitch plotting system. The limited data from the pilot also shows promising results for learners' ability to imitate native speaker prosody, as shown in Figure 3. However, there is too little data to draw conclusions about between-group differences or to assess participants' retention of their learning a week later.

Figure 3: Learner pitch trace from pilot study for the question *Luigi odia la neve?* 'Does Luigi hate snow?'



We are keenly looking forward to the results from the full study. We expect to have the learner data and the DTW results before the conference. The native speaker listening results may take longer, but it is possible we may have them before the conference as well.

V. Discussion

Our study experimentally tests the usefulness of a promising method of computer-assisted prosody training that can easily be implemented online. Our design allows us to investigate immediate imitation as well as generalization to novel sentences and delayed generalization a week later, letting us start to understand whether this type of training contributes to long-term learning.

Our project specifically addresses a gap in the teaching of Italian to speakers of English. However, the potential applications are much broader, as the methodology is language-independent.

Of particular interest is our use of both an automatic metric and native speaker perception. If the results of our computational metric and the native speaker task are both positive, we can reanalyze the results of the native speaker task to assess the effectiveness of automated DTW-similarity assessment. This would be as simple as replacing the predictor variable by the DTW scores of the corresponding recordings, and test whether or not an improved score is

predictive of which recording the native speaker picked. If this is a significant predictor, then we have provided further evidence that the system of assessment detailed by Arias and colleagues [13] is effective and could potentially be deployed together with the visualization app for better self-guided learning.

VI. Conclusion

We present a working web application for computer-assisted prosody training using F_0 visualization and detail an experiment to test its effectiveness. It is a promising system for self-guided prosody acquisition in an L2. Once we get the results of the experiment, we will have conclusions to share!

VI. References

[1] Avesani, Cinzia. 1990. A contribution to the synthesis of Italian intonation. First International Conference on Speech Language Processing (ICSLP 1990). Kobe, Japan, November 18-22, 1990.

[2] D'Imperio, Mariapaola. 2002. Italian intonation: an overview and some questions. Probus 14, 37-69.

[3] Grice, Martine, Mariapaola D'Imperio, Michelina Savino, & Cinzia Avesani. 2005. Strategies for Intonation Labelling across Varieties of Italian. in Sun-Ah, Jun (ed.) Prosodic Typology: The Phonology of Intonation and Phrasing. Oxford: Oxford University Press.

[4] De Bot, K. 1983. Visual Feedback of Intonation I: Effectiveness and Induced Practice Behavior. Language and Speech, 26 (4), 331–350.

[5] Anderson-Hsieh, J. 1992. Using electronic visual feedback to teach suprasegmentals. *System*, 20, 51-62.

[6] Hardison, Debra. 2004. Generalization of computer assisted prosody training: quantitative and qualitative findings. Language Learning & Technology. 8(1), 34-52.

[7] Savino, Michelina. 2012. The intonation of polar questions in Italian: Where is the rise? Journal of the International Phonetic Association, 42(1), 22-48.

[8] Rossano, Federico. 2010. Questioning and responding in Italian. Journal of Pragmatics, 42, 2756–2771.

[9] Cangemi, Francesco. 2014. *Prosodic detail in Neapolitan Italian.* (Studies in Laboratory Phonology 1). Berlin: Language Science Press.

[10] Rilliard, Albert, Allauzen, Alexandre & Boula de Mareüil, Philippe. (2011). Using Dynamic Time Warping to Compute Prosodic Similarity Measures.. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH.

[11] Boersma, Paul & Weenink, David. 2021. Praat: doing phonetics by computer. Version 6.1.40. http://www.praat.org/

[12] Jadoul, Yannick, Bill Thompson, & Bart de Boer. 2018. Introducing Parselmouth: A Python interface to Praat. Journal of Phonetics, 71, 1-15.

[13] Arias, Juan & Yoma, Nestor & Vivanco, Hiram. (2010). Automatic intonation assessment for computer aided language learning. Speech Communication. 52. 254-267. 10.1016/j.specom.2009.11.001.