Prosodic imitation of audiovisual and audio-only prompts in L2 English

Štefan Beňuš1,2

1Constantine the Philosopher University in Nitra, Slovakia
2Institute of Informatics, Slovak Academy of Sciences, Bratislava, Slovakia
sbenus@ukf.sk

Abstract

Speech entrainment, also alignment, convergence, or accommodation, is a pervasive, sub-conscious, and complex tendency of interlocutors to speak more similarly over time. It has been shown that audiovisual speech facilitates this tendency more than auditory-only speech, and that this applies to both segmental features such as VOT or vowel formants and prosodic features such as f0. The effect of audio-visual vs. audio-only prompts on the prosodic imitation of non-native speakers, as a special case of this tendency, has not been extensively studied but can add to our understanding of how the visual modality might be relevant in imitation and second-language distant learning. 90 L2 speakers produced triplets of words first on their own and then imitating a model speaker presented either as a video clip or as a sound only. The words varied in their f0 contours (fall, rise, fall-rise). The imitation of f0 and duration was assessed acoustically, not perceptually. Contrary to the expectation, the audiovisual modality did not facilitate greater imitation than the audio-only prompts for f0, and for duration the audio-only modality resulted in slightly better imitation scores. We discuss how the specific features of this dataset might have influenced the results.

Index Terms: prosodic imitation, audio-visual speech, non-native English

1. Introduction

Imitation is a core process in language acquisition and particularly important in non-native (L2) language acquisition. For example, imitation facilitates the recognition of single words [1]. Imitation also presents a special case of a broader tendency for speech production to share features of perceived speech variably termed entrainment, alignment, accommodation, or convergence. While in L1 acquisition imitation tasks tend to involve a more conscious cognitive process, the tendency for entrainment or alignment in semi-spontaneous dyadic spoken interactions is more general, typically occurs at a sub-conscious level, and contributes to better online predictions of what the interlocutors are going to say [2] also in adverse acoustic environments [3].

Multiple studies have shown that the modality of interaction, i.e. if people also see each other (audio-visual) or only hear each other (audio-only), has an effect on the degree and quality of speech imitation. For example, the tendency for alignment was enhanced in audio-visual modality over audio-only modality (e.g. [4], [5]). The intricate relationship between the auditory and visual modalities of speech has received extensive attention since the findings of the McGurk effect [6], mirror neurons and their implication for speech perception (e.g. [7]) or the observations showing that spoken communication might be a sole source of spontaneous behavioral coupling [8] even without seeing such behavior as was the case for the entrainment of postural sways between interlocutors who did not see each other [9].

This positive effect of visual modality is understandable for the imitation of various segmental features where the degree of lip protrusion, jaw lowering or dental placement of the tongue tip might provide important cues for the articulatory adjustments in the realizations of various segments. Additionally, various aspects of facial expressions provide relevant information for pragmatic and social meanings that are co-encoded in the prosodic characteristics of utterances.

However, prosody is a critical component of phonetic imitation even in relatively context-free setups. For example, [10] showed that the absence of f0 in the model stimulus (through high-pass filtering) decreases the degree of imitation in a word shadowing task. The results in [5] showed that in addition to the segmental aspects, non-criterial measures such as f0 and duration were imitated more in the audio-visual condition compared to the audio-only condition. Additionally, larger f0 excursions associated with local prosodic prominence might be coupled with stronger and longer supra-laryngeal gestures [11], and [12] showed that presenting prominence as facial gestures on an animated talking head is beneficial for speech intelligibility. Hence, seeing the speaker’s face likely provides additional cues regarding f0 characteristics.

As regards L2 speakers, some positive effects of visual modality on f0 training have also been established. For example, seeing the representation of the relevant aspects of f0 contour has a positive effect on intonation training in L2 (e.g. [13]) and is readily accessible with wide-spread tools such as Praat. However, the effect of seeing a speaker vs. just listening to a speaker in the prosodic imitation of L2 speakers has not been studied extensively. For collaborative dialogues between L1 speakers [14] found a tendency to enhance speech entrainment among interlocutors in the audio-only modality.

Understanding the potential benefits of seeing a video of a speaker on the screen on the degree of prosodic imitation for non-native speakers might be informative particularly in the more wide spread distant learning that has gained importance originally due to the COVID pandemic but that is likely to remain a common mode of language learning and teaching for the future as well.

In sum, the main question in the current paper is if L2 learners imitate f0 and duration better when they have access to seeing the model speaker than when they only hear the audio in a context-free word imitation task. We expect that since the audio-visual modality is more ecological and natural, and shows benefits for speech intelligibility and imitation of the prosodic structure, the L2 subjects should imitate f0 better than in the audio-only modality. However, following the results of the inhibitory influence of visual modality on entrainment in
dyadic L1 interactions [14] one might also imagine that watching a model speaker in a non-interactional task does not provide additional benefits for f0 imitation or that it might even be distracting as the video does not include clear overt f0 cues for the subjects.

2. Methodology

2.1. Subjects

Data come from a corpus of recordings for the initial oral assessment of 90 college students (66 females, 24 males) starting a course of English phonetics at a bachelor level. The native language of the students was primarily Slovak (N = 71) but some were Hungarians living in Slovakia (13) and some were foreign students of the Russian or Ukrainian origins. They were randomly assigned to either the video (N = 46) or the audio (N = 44) condition.

2.2. Stimuli

The material consist of 24 short English words specifically selected to test the articulatory interference between L1 and L2 of the subjects at the segmental level. Hence, they included those aspects of English pronunciation most commonly associated with a so called L2 accent in their English due to the differences in inventories, sound realizations, phonotactic or allophonic patterns. Specifically, the words included aspirated plosives [pʰ, tʰ, kʰ], (inter-) dental fricatives [θ, ð], word-initial rhetic [z], word-final velar nasal [ŋ] and voiced obstruents commonly undergoing voice neutralization in L1, the contrast between the tense and lax high vowels [i, u] vs. [i, o], or the contrast between [e] and [æ] vowels. These 24 words were organized into 8 triplets of the form ‘word1’, ‘word2’ and ‘word3’; for example pick, tooth, and speak.

These triplets were produced and video recorded by a male model L2 speaker (the author) with the camera zoomed to a full facial profile of the speaker. The prompts were recorded using three basic intonational contours: fall (H*L-L%), rise (L*H-H%) and fall-rise (H*H-L%) in ToBI notation such that 3 triplets had falls, 3 had rises, and 2 had fall-rises. Note, however, that the contours vary only for the first two words of each triplet and the third word always had a fall. Hence, 14 words were produced with a fall, 6 with a rise, and 4 with a fall-rise. The tokens distribution for the mode, model curve and biological sex in our data is summarized in Table 1.

The video clips for all triplets were extracted from the recording using a standard video editing software and the sound from each triplet was then extracted and saved. This procedure then gave 8 video clips for the triplets and 8 corresponding audio files.

2.3. Procedure

Subjects came individually to a sound-treated studio and were seated in front of a screen with instructions and stimuli. A Rode NT3 condenser microphone was placed on the table roughly 30cm from the subject’s mouth and connected to a Zoom H6 digital recorder.

The oral assessment started with the instruction to say the following phrases in their natural comfortable style and speed. The 8 triplets were shown on the screen, always in the same order, and with transitions manually controlled by the experimenter. Then the subjects were asked to read a short text and after that, another instruction asked them to imitate the words and the way the phrases are said. In the same order as the first presentation of the triplets, subjects in the video condition saw a full screen clip of the model speaker producing the triplet after which the screen changed to the text of the triplet identical as in the first part of the assessment. Subjects in the audio condition saw a full screen image of a speaker icon and only listened to the sound of the model speaker.

2.4. Data processing

Although some studies assess the similarity of f0 contours via manual discrete ToBI-like annotations, e.g. [15][16], this approach is not feasible in the current study due to the expected high similarity of the broad contour shapes. We thus aimed to compare stylized contour shapes rather than their discrete categorizations. After extracting the mono track from the Zoom recorder, each recording was manually aligned using Praat such that the temporal onset and offset of each word was marked. For words with initial voiceless plosives the onset was placed 50ms

---

Figure 1: Illustration of data for one speaker after processing; the word and the curve type (‘r’ for rise, ‘f’ for fall, and ‘fr’ for fall-rise) are shown in the heading of each panel, the blue line shows the f0 of the model speaker, the black one of the subject’s first realization, and the red one subject’s realization after the exposure to the model speaker. The triplets are organized horizontally, so all words in the bottom row were falls since they concluded the triplet.
before the burst and the offset of the words with final plosives was marked immediately after the burst. The ProsodyPro script for Praat [17] was used to manually check each recording for spurious pulses and pitch points, to correct pitch halving, and to extract mean, maximum, minimum f0 values, time-normalized f0 (f0 in each word divided into 10 points), and durations for each word. Figure 1 shows time-normalized f0 curves in Hz as the output of this processing for a single female speaker.

So for each word we have three time-normalized f0 contours: a) the first realization (black), b) the second realization after hearing the model speaker (red), and c) the contour of the model speaker himself (blue). Following [18], we assessed the similarity of contours by calculating Pearson coefficients and RMSE for pairs of contours; FS corresponds to the first vs. the second realizations, FM to the first vs. the model speaker, and SM to the second vs. the model speaker. Since the f0 values were hand-corrected we did not employ more sophisticated methods of weighing by the intensity. We then assessed the imitation in f0 curves induced by the model speaker by calculating the difference between the first and second realizations with respect to the model speaker curve as follows:

\[ \text{pearson}_{\text{IMI}} = \text{pearson}_{\text{SM}} - \text{pearson}_{\text{FS}} \]  
\[ \text{rmse}_{\text{IMI}} = \text{rmse}_{\text{FM}} - \text{rmse}_{\text{SM}} \]  
\[ \text{range}_{\text{IMI}} = |\text{range}_{\text{FM}}| - |\text{range}_{\text{SM}}| \]

(Derived variable pearsonIMI increases with high similarity between the second and model curves and high dissimilarity (i.e. negative values) between the first and second curves.) Variables rmseIMI and rangeIMI are calculated in a similar way, only the order of FM and SM values is reversed since greater values here mean dissimilarity while greater values for pearson mean similarity. Hence, we gauge imitation as the difference between the second realization and the model speaker (small value means high similarity), which is deducted from the difference between the first realization and the model speaker (higher the value greater the difference). Range is calculated as the difference between the maximum and minimum f0 value for a given word. PearsonIMI can be considered the most conservative measure of imitation since the imitation of subject’s second realization after hearing the model is corrected by the (dis)similarity between the first and second realizations. For all derived variables, the higher the value the greater the imitation.

For example, the panel in the top row and second column of Figure 1, “soothing – F” shows robust imitation: the original realization (black) is very different from the model speaker (blue) while the realization after the exposure to the model speaker (red) is more similar to the blue curve than the original black curve. On the contrary, the panel in the top row and the fifth column, “scam – F” shows minimal imitation when the black and red curves are similar and both are different from the blue curve of the model speaker. PearsonIMI and rmseIMI for ‘soothing’ is 1.67 and 53.34 respectively while for ‘scam’ it is -1.50 and 13.74 respectively. Finally, we also assessed the imitation of durational variation with the formula in (4). The greater the difference between the first and model realizations and smaller the difference between the second and the model realizations, the greater the imitation and the value of durIMI.

\[ \text{dur}_{\text{IMI}} = |\text{dur}_{\text{FM}}| - |\text{dur}_{\text{SM}}| \]

Hence, similar to measures (1)-(3), the higher the value of durIMI the greater the durational imitation.

3. Results

The primary question of the study concerned the influence of the modality (audio-video vs. audio-only) on the imitation of three intonational curves (fall, fall-rise, rise) in terms of f0 and duration. We used two potentially related measures, pearsonIMI and rmseIMI, but in the former the difference between the subject’s first and second realizations were included directly while in the latter both first and second realizations were related to the model speaker’s realization. The correlation test showed a significant but only moderate level of relationship (r(2158) = 0.35, p < 0.001). Hence, we used both of them in examining our research questions. Since the model speaker was male, we also examine the effect of subject’s gender on imitation.

We employed linear mixed effect models with the four dependent variables pearsonIMI, rmseIMI, rangeIMI, and durIMI and random slopes for the subject and word variables (since models with random slopes usually did not converge). Statistical models were run in R’s lmerTest package [19] that calculates p-values using the Satterthwaite’s method.

Figure 2 shows the effect of modality and f0 curve on imitation captured with the pearsonIMI variable.

The linear mixed effects model showed that the mode of presentation had no significant effect on imitation and a very weak tendency for the contour type: F = 2.58, p = 0.099 arising from the difference between the fall and fall-rise contours (t = -1.87, p = 0.07). The interaction between the two factors was also not significant and biological sex of the subjects did not affect imitation significantly.

Finally, the trend showing the benefit of the video mode over the audio in imitating the fall-rise contour was not significant in the pooled data but did return a significant main interactional effect between mode and curve type in the corpus that included only 71 Slovak speakers (F = 3.94, p = 0.02), which arises from the interaction between audio-video mode and fall vs. fall-rise (t = -2.62, p = 0.009) and rise vs. fall-rise (t = -2.52, p = 0.012). The test for rmseIMI variable did not return any significant result for mode or contour but a tendency for males to imitate more than females (F = 3.34, p = 0.07).
The third f0 derived variable rangeeu showed a weak but significant interaction between the mode and the contour (F = 3.87, p = 0.021) shown in the left panel of Figure 3. This arises primarily from the difference between the fall and fall-rise contours: the former are slightly better imitated in the video condition whereas the latter in the audio-only condition (t = -2.51, p = 0.012). The corresponding difference between the falls and the rises reached a tendency (t = -1.79, p = 0.073); both post-hoc tests were run in the lsmeans R package using the tukey test.

The right panel of Figure 3 shows the main effect of biological sex (F = 3.95, p = 0.05) in that males mimiced the model speaker’s f0 range more than females, and this difference is primarily attributed to the increased facilitatory effect of the video mode for males since the interaction of sex and mode was also significant (F = 4.06, p = 0.047).

Finally, the model testing the imitation of duration showed a similar pattern to that of range, but the differences were not significant.

![Figure 3: Imitation of f0 range for audio vs. video and three f0 contours (left) and biological sex (right)](image)

4. Discussion & Conclusion

We set out to investigate if the modality of the prompt (audio vs. video) affects the imitation of f0 contour, f0 range, and duration for L2 speakers shadowing short words arranged into triplets. The results suggest overall that the modality does not have a systematic effect on the degree of imitation. The main results show the absence of the effect for the f0 contour imitation. The observed partial effects had different directions: for biological sex or the curve type in f0 range video facilitated imitation for males and in the fall-rise contours, and audio-only facilitated imitation in duration and the f0 range of the (fall-rise) contour. Hence, for some measures, speakers, and curves the additional video was facilitatory and for some inhibitory. The core finding is thus a somewhat unexpected absence of the systematic facilitatory effect of the video modality on the degree of imitation in our dataset.

Our results are thus in line with [14] where the comparison of audiovisual and audio-only modalities in L1 collaborative dialogues returned a very complex pattern with a weak tendency for facilitation in the audio-only condition. We observed weak tendencies for this in our data for the f0 range in the (fall-rise) contours (and duration).

The observed main effect of biological sex on imitating f0 range where males imitated the range more than females might be attributed to the matched vs. mis-matched subject-model pairs. Since the male model speaker produced prompts in a somewhat hyper-articulated fashion and exaggerated pitch range, the difference between the model and first realizations for the males was greater than that for the females. Hence, the potential for imitation for males was greater than that for females.

Despite the absence of the interaction between the mode and the contour type, in all four models, the fall-rise contour showed greater difference between the audio and video modes than the other two contours, and significantly so for the Slovak-only speakers. In f0 shape imitation, as seen also in Figure 2, the video mode slightly helped with the imitation of this contour while in duration, the audio-only mode lead to slightly better imitation. These observations, however, should be treated with great caution since this contour was represented by the smallest number of tokens, and it is also not as common and natural for the L1s of the subjects as the other two contours. On the other hand

Although the observations regarding relative differences in imitation due to the examined factors above are valid, we note that the degree of imitation is quite low and inconsistent across the data as seen by the median values around zero in the most direct and conservative measure of Figure 2 (the other three measures yielded positive medians). Probably the nature of the instruction might have also lead to the low imitation levels and consequently the absence of systematic facilitation of the video mode on the imitation. Despite the rather explicit form of the instruction to “imitate the words and the way the phrases are said”, the subjects might have reasonably focused more on the segmental aspects and sub-consciously overlooked the prosodic factors in their efforts to imitate the model speaker’s segmental characteristics.

Another limitation of the study is the use of a single male model speaker and the absence of holistic perceptual measures of imitation in addition to the acoustics-based measures. As discussed in [20] imitation tends to be subtle, variable and inconsistent if all these aspects are experimentally controlled. It is plausible that with a fuller range of control parameters, the effect of the video modality on prosodic imitation might be completely neutralized.

The experimental setting is also a potentially influencing factor. On the one hand, the task is not interactional, which allows for tighter control but at the same time lowers the ecological validity of the resulting speech data. Additionally, psycho-social effects such as the identity of the model speaker and his perceived status as the course instructor and his role of the experimenter at the same time might have also influenced the motivation or stress levels and thus the imitation strategies of the subjects [21].

Despite these limitations, the dataset involves sufficient variability in the degree of prosodic imitation among speakers to allow for future investigation of how the modality of the prompt presentation affects the relationship between the imitation of the segmental features (e.g. VOT duration for aspirated plosives or vowel quality) and the imitation of the prosodic characteristics.

5. Acknowledgements

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 859588 and from the Slovak Granting Agency grant VEGA 2/0165/21.

6. References


