Production and perception of tonal coarticulation: Evidence from computational simulation of communication

Huang, Po-Hsuan

Graduate Institute of Linguistics, National Taiwan University

benson32169@gmail.com

Introduction. Lexical tones have been found to coarticulate with the preceding and following tones (e.g., Xu 1994; Wang 2002). For instance, Xu (1994) has found that tones after high tone offsets (i.e., 55 and 35 tones) were raised, and those after low tone offsets (i.e., 51 and 21 tones) were lowered in Beijing Mandarin (BM). Such tonal coarticulation (TC) has also been found in languages including Southern Min (TSM; e.g., Wang 2002), Taiwan Mandarin (TM; e.g., Huang 2023), etc. Variations induced by such coarticulation have also been found to affect listeners' perceptions and cause a target tone to be perceived as other lexical tones (Xu 1994; Wang 2002). This therefore leads to the question of how tone language speakers keep faithful perceptions of tones under TC. In Xu and Zhang et al. (2022), BM speakers have been found to cope with such variations with normalization, where tones after high offsets were perceived as lower, and those after low offsets as higher. While similar normalizing effects have been found in other languages including TSM and TM (Wang 2002; Huang 2023), it may not be the only mechanism used to cope with tone variations induced by TC. Specifically, it has been proposed by Huang (2023) that linguistic differences may exist with regard to three aspects: 1) the magnitude of TC, 2) the magnitude of normalization for TC, and 3) the ranges of tone acceptance. While Mandarin may allow for TC with the listeners being able to retrieve the target tones back through normalization, it might be less viable for languages with a larger tone inventory, such as TSM. In TSM, the recoverability of the target tones through normalization is lower due to the multiple possible tones that may surface as the same coarticulated tones in the same position. TSM users might alternatively reduce such variations by avoiding the same magnitude of TC as Mandarin, or by maintaining narrower tone acceptance ranges to keep out coarticulated tones. In Huang, the latter was found. While both TM and TSM had similar degrees of TC, TM demonstrated stronger normalization for TC than TSM. On the flip side, TSM maintained narrower tone acceptance ranges as compared with TM. It is argued by Huang that such linguistic differences resulted from the different tone distributions of TM and TSM. However, since the experiments were behavioral experiments conducted on human subjects, multiple aspects need to be factored in, and a direct relation between the linguistic differences and tone distributions could not be easily drawn. Computational simulation of real-world communication may shed light on such an issue. Past studies have proved the ability of communication simulation to capture important linguistic features through the interaction of the speaker and listener agents. In Ren et al. (2020), compositionality emerged through the training of two neural agents simulating the speaker and the listener. Likewise, in Carlsson et al. (2023), it has been found that the joint combination of communication simulation and iterated learning could result in efficient color naming systems similar to those found in human languages. In this study, the author uses a speaker neural agent and a listener neural agent, modulated by the three aspects proposed by Huang (2023), to simulate real-world tone communication under TC, and seeks to provide a more direct observation of the relation between tone distributions and the linguistic differences in the production and perception of tonal coarticulation.

Methods. To simulate tone communication, tone contours were represented as tone onsets and offsets ranging from 1-5, based on the five-level tone marks. In TM, there were four tones: (5, 5), (3, 5), (2, 1), and (5, 1). In TSM, an additional (3, 3) was added, leading to a larger inventory. The data were generated based on these tones, and a TM model and a TSM model were trained. To simulate real-world variance, for each tone generation, the tone onset and offset were randomly sampled on normal distributions with the means being the standard values. To simulate TC between the preceding and target tones, for each possible combination, 2048 tokens were generated. Among them, 80% were taken as training data, and 20% were taken as validation data. To train the listener agent to recognize the canonical tone, an additional 2048 tokens were generated for each tone. Two neural agents were constructed with multilayer perceptrons (MLP) and trainable parameters to represent the speaker and the listener in communication. During each epoch, the listener was first trained with single tones for four sub-epochs (Phase A) to recognize the canonical forms of the tones, and then joined with the speaker and trained with tone pairs for another four sub-epochs (Phase B). There was a total of 256 epochs. In the speaker agent, an MLP was used to simulate coarticulation. A value from 0-1 was produced by the MLP, and used as the degree of coarticulation. A value of 1 meant complete coarticulation with the preceding tone, while 0 meant no coarticulation at all. The coarticulated tone pairs would then be taken as the input for the listener (in Phase B). In the listener agent, two trainable parameters and two MLPs were used. The trainable parameters represented the tone acceptance ranges as normal distributions for each lexical tone, with one being the mean value of the acceptance range, and the other being the standard deviation. For the two MLPs, one was used to represent phonological perception, where the continuous tone onsets and offsets of the target tone were taken in, and an initial guess of which tones this token might be was produced. If the training was in Phase A, the guess would be directly used as the final prediction and evaluated for backpropagation. If it was Phase B, the guess would be joined with the continuous tone onsets and offsets

of the preceding tone as input for the other MLP, which was used to allow for normalization of the listener and the final prediction of the target tone would be produced.

The three aspects proposed by Huang were evaluated as follows. The mean degree of coarticulation of the speaker during validation was taken as the magnitude of coarticulation. For normalization for TC, following Zhang et al. (2022), a series of target tones simulating the continuum from the low tone (21) to the falling tone (51) following the different preceding tones were predicted by the TM and TSM models. If normalization was at work, a preceding tone with high offsets (e.g., 55 and 35) would lead the target tone to be perceived as lower, and it would have to be very close to a canonical 51 to be perceived as a falling tone, and vice versa. Finally, tone acceptance ranges were assessed by both the standard deviations of the tone acceptance ranges and the differences of the tone acceptance means with their original values at the onset (e.g., if the mean of the tone acceptance range for 51 is 4.5 at the onset, then the difference is 0.5) during validation. Both a smaller deviation and a smaller difference would indicate a narrower tone acceptance range (cf. Huang).

Results. The accuracies of the TM model and TSM models were 0.62 and 0.54. This relatively low performance was understandable since variances were intentionally introduced to mimic real-world speech. The mean degrees of coarticulation in the TM and TSM models were 0.52 and 0.46, suggesting a higher magnitude of TC in the TM model. The magnitudes of normalization for TC in the two models are shown in **Figure 1**. As can be seen, compared with TSM, the TM model was more subject to the offset height of the preceding tones, as indicated by the interval (1.80) between the orange line (51 as the preceding tone) and the green/blue lines (35 and 55 as the preceding tones) at the 0.5 midpoint, which is much larger than the one in the TSM model (1.20).

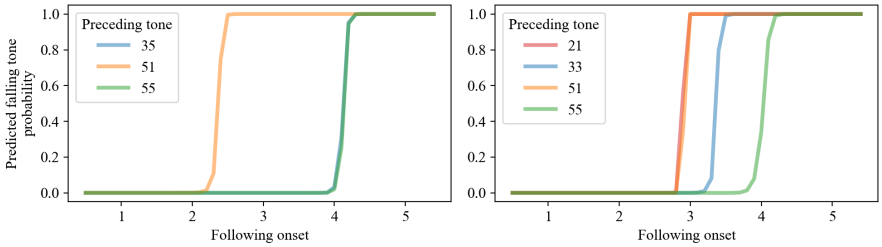

Figure 1:*Normalization of the listener neural agent for different preceding tones on a low-to-falling tone continuum (left: TM; right: TSM).*

Finally, the mean standard deviations of the TM and TSM models were 2.17 and 2.28; the mean differences of the tone acceptance mean with the original values were 0.27 and 0.20. These two metrics thus suggest opposite interpretations for the acceptance ranges in the two models. Judging by the standard deviation, TM had narrower acceptance ranges. However, based on the mean difference between tone acceptance means with the original values, TSM had stricter acceptance ranges.

Discussion. The results of the simulation in general supported the hypothesis that different tone distributions could lead to different strategies in dealing with the tone variations induced by TC. Specifically, by simulating the tone inventories of TM and TSM, the current study demonstrated that a language with more complicated tone inventories could opt to reduce the degree of coarticulation. More importantly, the behavior of the listener agent in the two models largely replicated the findings in Huang (2023) on human subjects. Normalization for tonal coarticulation has emerged, and a difference in magnitude existed between the TM and TSM models. The tone acceptance ranges, however, showed mixed results. Based on standard deviations, TM had narrower acceptance ranges, while looking at the mean difference between tone acceptance means with the original values, it was TSM that had narrower ranges. This was different from the findings in Huang (2023), where TSM was found to have narrower acceptance ranges. This could be due to the fact that in this study, TSM already had a smaller magnitude of TC, and consequently did not have to rely as much on perceptual mechanisms to resolve the tone variations, which was not the case in Huang's production experiments, where both TSM was found to have similar degrees of TC as TM. This might suggest that there exist certain biomechanical constraints that also need to be taken into account by the simulation. In general, this study demonstrates the possibility of simulating real-world communication and its ability to allow for more direct explanations of production and perception behaviors through the interaction of the speaker and listener agents.

References

Carlsson, E., Dubhashi, D.P., & Regier, T. (2023). Iterated learning and communication jointly explain efficient color naming systems. *ArXiv*, abs/2305.10154.

Huang, P.H. (2023). Perception and Production of Coarticulated Tones in Taiwan Mandarin and Taiwan Southern Min [Master's thesis]. National Taiwan University, Taipei.

Ren, Y., Guo, S., Labeau, M., Cohen S.B., & Kirby, S. (2020). Compositional languages emerge in a neural iterated learning model. International Conference on Learning Representations, online.

Wang, H. (2002). The prosodic effects on Taiwan Min tones. *Language and Linguistics*, *3*, 839-852.

Xu, Y. (1994). Production and perception of coarticulated tones. *Journal of the Acoustical Society of America*, *95*(4), 2240-2253.

Zhang, H., Ding, H., & Lee, W.-S. (2022). The influence of preceding speech and non-speech contexts on Mandarin tone identification. *Journal of Phonetics*, *93*, 101154