Going beyond F0: The acquisition of Mandarin tones

Online Supplement

In this supplement, we provide additional details to the Method section of the paper.

Computational modeling

Two computational modeling methods were taken in this study: multidimensional scaling (MDS) and automatic machine-learning classification.

Multidimensional scaling (MDS)

Multidimensional scaling (MDS) is a dimension-reduction technique that provides a visual representation of the pairwise (dis)similarity between data points. In our study, MDS was used to transform the multidimensional and highly correlated acoustic space into a more interpretable, low-dimensional ordination space. Specifically, we performed non-metric MDS, which utilizes ranks of the pairwise distances in the original multidimensional dataset to calculate the pairwise dissimiliarity in the low-dimensional space, and is thus generally more robust to possible missing datapoints and extreme outliers.

F0 and spectral measures of the sentence-medial verb syllable were extracted using VoiceSauce (Shue, Keating, Vicenik, & Yu, 2011) at 9 equidistant subsegments. Measurement from the first 3 subsegments were removed to eliminate the influence of the onset consonants. Measurement at remaining 6 subsegments were separately included for each cue, in order to ensure that the temporal information in the tonal production was preserved. For each age group (Age 4-5, Age 7-8, Age 10-11, and Adults), the following cue sets were used to model their tonal production.

Cue sets (by token):

(i) F0 cues: F0 values (STRAIGHT; Kawahara, Masuda-Katsuse, & De Cheveigne, 1999) at 6 subsegments for each token

(ii) Spectral cues: For each of the spectral measures (CPP, H1*-H2*, H2*-H4*, H1*-A1*, H1*-A2*, H1*-A3*, H4*-2K*, 2K*-5K*), values at 6 subsegments for each token (6 CPP values, 6 H1*-H2* values, …)

(iii) F0 and spectral cues: Both (i) and (ii) combined

MDS failed to converge when pairwise distance matrices were calculated using all tokens. In order to prevent convergence failures, the cue sets were averaged for every tone in every focus condition (4 tones x 5 focus conditions).

Cue sets (by tone and focus condition)

(iv) F0 cues: 6 F0 values at each subsegment, averaged for each tone in each focus condition

(v) Spectral cues: For each of the spectral measures (CPP, H1*-H2*, H2*-H4*, H1*-A1*, H1*-A2*, H1*-A3*, H4*-2K*, 2K*-5K*), 6 values at each subsegment, averaged for each tone in each focus condition

(vi) F0 and spectral cues: Both (iv) and (v) combined

Hence, the pairwise distance matrices were based on Euclidean distances between all tone-focus condition pairs, for each cue set (iv-vi).

The number of dimensions (k = 2) for the MDS was chosen based on stress evaluation as well as for visualization purposes. Stress values represent the goodness of fit: stress values at or below 0.2 are considered a fair fit, at or below 0.1 a good fit, and at or below 0.05 an excellent fit (Kruskal & Wish 1978). The stress values from MDS to two dimensions are shown for each age group in Figure 1, which are replicated in Table 2.

Table 2
Stress values from non-metric MDS, k=2

	Age 4-5	Age 7-8	Age 10-11	Adults
F0 cues	0.002	0.003	0.003	0.007
Spectral cues	0.091	0.108	0.113	0.071
Both F0 and spectral cues	0.095	0.062	0.059	0.063

All MDS plots yielded at least a fair fit ($stress \leq 0.2$) with just two dimensions; in fact, most except two yielded a good or excellent fit. Increasing the number of dimensions reduces the stress value, but we have chosen to represent our data with two dimensions because two dimensions are in fact sufficient to achieve fair to excellent fit, and the number of dimensions should be consistent across all age groups and cue sets for comparison.

Automatic machine-learning classification

Automatic tonal classification was performed to quantify and validate MDS results. For machine-learning classification, we used cue sets by token (i-iii above).

We interpret the classification accuracies as representing the informativeness of the cue sets in distinguishing tones. Comparing accuracies between age groups and cue sets reveals at what age and by how much each cue set becomes important. Furthermore, comparison between using only F0 or only spectral cues and both (F0 + spectral) cues allows us to estimate the level of additional information each cue set gives in the presence of the other. In other words, if an

automatic classification mechanism predicts the correct tonal category in X% of instances when only using cue A and in X+Y% of cases using cue B in addition, then Y quantifies the degree to which B is additionally informative. If either F0 or spectral cues contain unique, non-redundant information about the tonal categories, the classification accuracy of the combined (Both F0 and spectral) cue set should be higher than accuracy of the individual cue sets.

Three different machine learning algorithms were tested to cross-verify the results. Specifically, we chose Linear Discriminant Analysis (LDA, MASS package, Venables and Ripley 2002) to test for linear discriminability, Support Vector Machine (SVM; using radial basis function kernel, e1071 package, Meyer, Dimitriadou, Hornik, Weingessel, and Leisch 2018) to allow nonlinear Gaussian distribution of the tonal classes, and Random Forest (RF, randomforest package, Liaw and Wiener 2002, with parameters *ntree = 500, other parameters default*) for better interpretability of the classification splits. Average classification accuracy of tonal classification was calculated for each age group and cue set, from 100 trials of 10-fold cross-validation. Differences among the algorithms in our results are minor and thus are not discussed in the paper.

References in the Supplement

Kawahara, H., Masuda-Katsuse, I., & De Cheveigne, A. (1999). Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency based F0 extraction: Possible role of a repetitive structure in sounds. *Speech communication*, *27*(3-4), 187–207. doi:10.1016/S0167-6393(98)00085-5

Kruskal, J.B. & Wish, M. (1978). Multidimensional Scaling. *Sage University Paper Series on Quantitative Applications in the Social Sciences*, No. 07-011, Sage Publications, Newbury Park. http://dx.doi.org/10.4135/9781412985130

Liaw, A. & Wiener, M. (2002). Classification and regression by random forest. *R News*, *2*(3), 18–22. Retrieved from https://CRAN.R-project.org/doc/Rnews/

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2018). *E1071: Misc functions of the department of statistics, probability theory group (formerly: E1071), tu wien*. R package version 1.7-0. Retrieved from https://CRAN.R-project.org/package=e1071

Shue, Y.-L., Keating, P. A., Vicenik, C., & Yu, K. (2011). Voicesauce: A program for voice analysis. In *Proceedings of the ICPhS XVII*, 1846-1849.

Venables, W. N. & Ripley, B. D. (2002). *Modern applied statistics with S* (Fourth). New York: Springer. doi:10.1007/978-0-387-21706-2