# INTEGRATION OF SPECTRAL CUES IN THE DEVELOPMENT OF MANDARIN TONE PRODUCTION

Nari Rhee[1], Aoju Chen[2], Jianjing Kuang[1]

University of Pennsylvania[1] Utrecht University[2]
nrhee@sas.upenn.edu, aoju.chen@uu.nl, kuangj@ling.upenn.edu

## ABSTRACT

Previous studies have shown that voice quality is integrated in Mandarin tone production and perception: Voice quality systematically co-varies with F0 in tone production, and the covarying voice quality cues (such as spectral cues) are sufficient to classify tonal categories both by human listeners and by automatic classifiers. Given the salient role of voice quality in Mandarin tonal contrast, this study investigates whether the integration of spectral cues in Mandarin tone production is acquired during the process of language development, or is purely determined by the physiological correlation between F0 and voice quality. By modeling the contrastivity of Mandarin tones produced by children of different age groups (4-11) as well as adults, we show that children increasingly become better at using spectral measures to contrast tonal categories as they grow older. Therefore, the integration of spectral cues in tone production is at least partially learned, rather than purely physiological.

**Keywords:** spectral cues, voice quality, tonal contrast, Mandarin, tonal acquisition

## 1. INTRODUCTION

Voice quality serves as an important cue in the production and perception of Mandarin tonal contrasts. It is well-known that, among the four lexical tones in Mandarin (Tone 1: high level, Tone 2: rising, Tone 3: low-dipping, Tone 4: falling), the low-dipping tone (Tone 3) is often produced with creaky voice [1], and this allophonic non-modal voice in turn can facilitate the perception of Tone 3 [1, 25]. However, the integration of voice quality cues is not unique to Tone 3; rather, all Mandarin tones are subject to the same phonetic mechanism [7], where the presence of allophonic non-modal voice is largely driven by extreme pitch targets: high pitch is naturally associated with tense voice, and low pitch is naturally associated with creaky voice [18, 17]. As demonstrated in [7], any tone can creak as long as it is produced in low-pitched prosodic conditions, such as the sentence final position [7, 8]. Moreover, [7] shows that, as a consequence of the systematic and continuous co-variation between voice quality production and F0 in pitch production, the acoustic correlates of voice quality (such as spectral cues) also systematically and continuously co-vary with F0 for a given speaker.

This strong co-variation suggests that spectral cues can be fairly informative in the tonal distinctions in Mandarin. Indeed, it has been reported that tonal categories were more successfully recognized using spectral information (mel-frequency cepstral coefficients) than using F0 by a deep-neutral network classifier [15]. Human listeners of Mandarin were also able to identify tonal categories fairly accurately in the absence of F0 information [12]. More generally, spectral cues are important part of pitch perception, since manipulating spectral cues can significantly shift the perception of pitch height [9].

Therefore, because of the co-variation between spectral cues and F0 cues in pitch production, pitch contrasts can be realized as either F0 and/or spectral cues. The question is then when the integration of spectral cues in tone production and perception is developed for Mandarin speakers? More specifically, is it gradually acquired during the process of language development, or is it purely determined by the physiological correlation between F0 and voice quality? If it is the former, we expect to observe a developmental path towards adult-like production in their use of spectral cues; if it is the latter, we expect children to exhibit patterns similar to adults at a very young age, as soon as they are able to produce the pitch contours.

The acquisition of Mandarin tonal contrast has been studied in both production and perception by a number of different studies. In production, studies have found that children are capable of contrasting tones before the age of 3 [10, 4]; however, their production of the pitch contours do not accurately match that of the adults even at the age of 5 [22]. In perception, while sensitivity to tones begin to develop before age 1 [19, 20], children only reach adult-like discrimination by the age of 5-6 [3].

Nevertheless, little is known about whether and how voice quality cues are produced and perceived in tonal contrasts by children. This study will serve as the first step to answer this question, by modeling the tonal spaces of children of different age groups and comparing them with the production of adults. Specifically, we test whether spectral cues significantly contribute to the dispersion of the tonal categories.

## 2. METHODS

### 2.1. Corpus

We used the corpus collected by Yang and Chen [24], consisting of 160 SVO Mandarin sentences spontaneously elicited in a question-answer dialogue setting in a picture-matching game [23, 2]. The sentences consisted of 4 verb tones, 4 object noun tones, 5 focus conditions, and 2 types of verbs and object nouns. The five focus conditions in the SVO sentences were: narrow focus on the verb (NF-m), narrow focus on the subject (NF-i), narrow focus on the object (NF-f), broad focus (BF), and contrastive focus on the verb (CF-m). The corpus had 10-12 native Mandarin speakers in each of the four age groups: 4-5, 7-8, 10-11, and a control-group of adult speakers. The 160 sentences were divided into two lists of 80 sentences, each elicited by half of the participants in two recording sessions. The different focus conditions in the corpus enable us to examine a range of variation of both F0 and spectral cues within the tonal categories. In this study, we focused on the contrastiveness of the tonal categories; the interaction between tone and focus will be discussed in more detail in a follow-up study.

The corpus was forced-aligned using the Mandarin forced-aligner [26]. F0 and spectral measures of the sentence-medial verb syllable were extracted using VoiceSauce [16] and time-normalized into 9 timepoints, of which we removed the first 3 timepoints to eliminate any influence of the onset consonants. We used STRAIGHT F0 [6], Cepstral Peak Prominence (CPP) for a measure of aperiodicity in the signal, and spectral measures of the relative amplitude difference of the lower and higher harmonics to capture information about the spectral slope (H1*-H2*, H2*-H4*, H1*-A1*, H1*-A2*, H1*-A3*, H4*-2K*, 2K*-5K*, corrected for the influence of formant frequencies and bandwidths on the harmonics [5]). All extracted measures were min-max normalized by speaker and recording session.

### 2.2. Computational modeling

To test whether all age groups were using spectral cues to distinguish the tonal categories, the tonal production of the different age groups were modelled using different feature sets: i) F0 (STRAIGHT), ii) only spectral cues (H1*-H2*, H2*-H4*, H1*-A1*, H1*-A2*, H1*-A3*, H4*-2K*, 2K*-5K*, and CPP) and iii) both spectral cues and F0.

First, non-metric multidimensional scaling(MDS) was used to calculate and visualize the dissimilarity of the tonal categories in different focus conditions, by the metaMDS function in the vegan package in R [14]. Euclidean distance was used to build the dissimilarity matrix.

Moreover, several machine-learning classification algorithms, namely Linear Discriminant Analysis (LDA)[21], Support Vector Machine (SVM; using radial basis function kernel)[13], and Random Forest (RF) [11], were used to cross-verify the findings from MDS. Average classification accuracy of tonal classification was calculated for each age group and feature set, from 100 trials of 10-fold cross-validation.

## 3. RESULTS

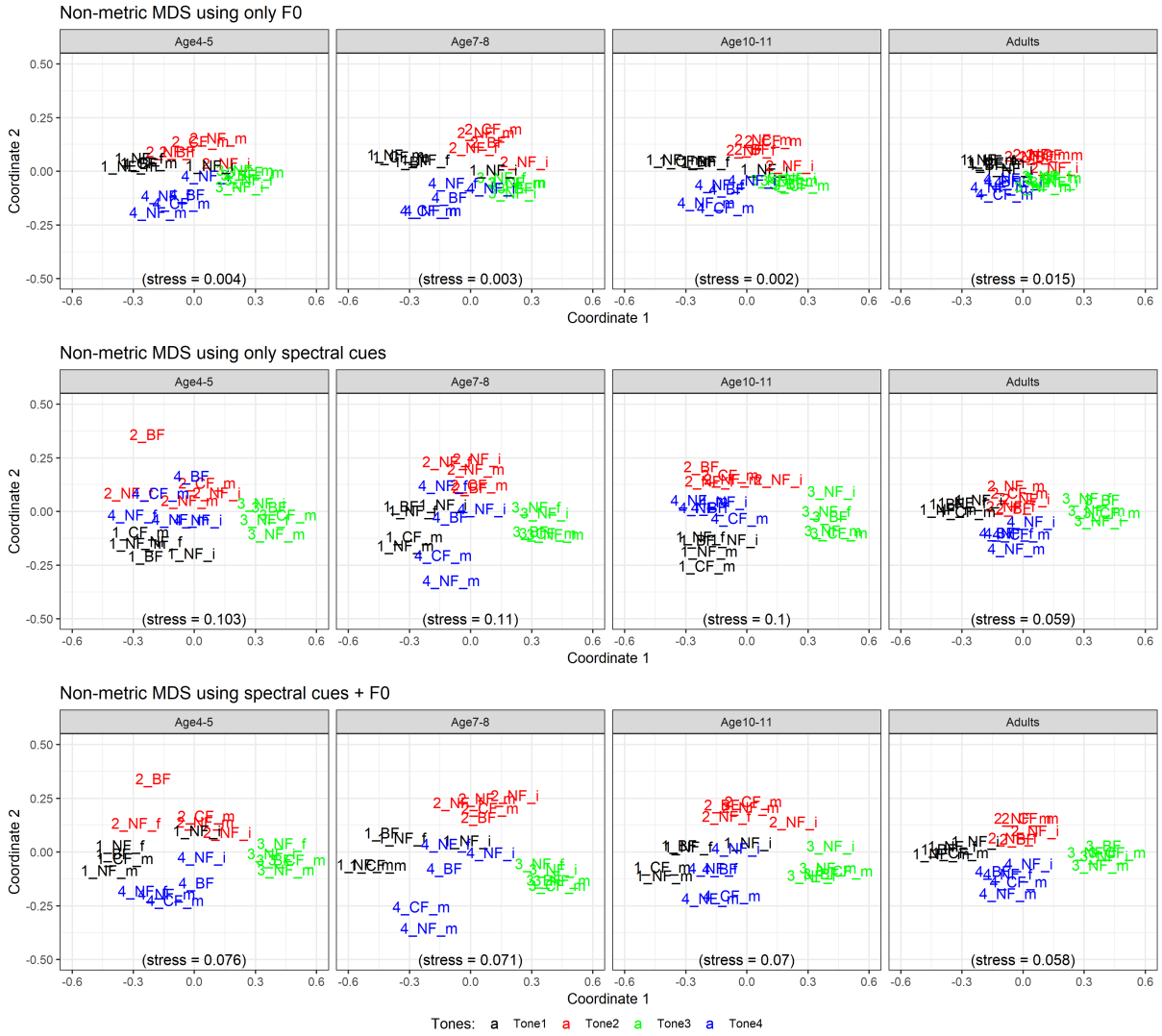
### 3.1. Multidimensional scaling

Results of MDS with $k = 2$ are shown in Figure 1. All plots have $stress \leq 0.11$, indicating that 2 dimensions were sufficient to achieve a good fit of the data. Overall, across all age groups, the MDS tonal spaces represented the distinction of high vs. low tone on the first dimension (x-axis), and the distinction of rising vs. falling contour on the second (y-axis).

As shown in the top panel of Figure 1, F0 cues alone achieved a decent separation of the tonal categories even for the youngest age group (Age 4-5), suggesting that children were able to make reliable tonal contrasts using F0 at the age of 4-5. Children of all age groups had distinct tonal clusters, older groups exhibiting a rather small improvement in using F0 cues to distinguish the tonal categories. The adults actually exhibited a tonal space that was less spaced-out than that of children, though each tonal category was more tightly clustered with no overlaps.

Shown in the middle panel of Figure 1, unlike the MDS spaces for F0, spaces based only on spectral cues exhibited a dramatic developmental pattern in which the overall separability of the tonal categories clearly improved with speaker age. At the youngest age (4-5), only Tone 3 was evidently dis-

**Figure 1:** Non-metric MDS plots ($k = 2$). The four tones are represented in different colors (Tone 1: black, Tone 2: red, Tone 3: green, Tone 4: blue). The five focus conditions in the SVO sentences were: narrow focus on the verb (NF-m), narrow focus on the subject (NF-i), narrow focus on the object (NF-f), broad focus (BF), and contrastive focus on the verb (CF-m).




tinguished from the other tones by the spectral cues, whereas the other tones were not as well-separated. Older children (Age 7-8 & 10-11) showed substantial improvement in the separation of all four tonal categories with just spectral cues, gradually forming tighter clusters that were more similar to those of the adults. For the adults, spectral cues alone were sufficient to well distinguish the tonal categories. As shown in the bottom panel of Figure 1), using both spectral and F0 cues showed minimal improvement in separability from using just spectral cues, particularly for the adult speakers. Tonal categories were well separated for all age groups, yet a clear developmental pattern could still be observed: as children grew older, within-category variation introduced by the different focus conditions gradually decreased, while cross-category variation gradually increased.

Overall, our results clearly suggested that spectral cues play more important roles in tonal contrasts for adults and children at older ages than for the very young children; and the integration of spectral cues in tone production gradually develops over time. Some tone-specific patterns were observed, such that Tone 3 was more clearly distinguished by the spectral cues even for the youngest children (Age 4-5), but the integration of spectral cues continue to

develop for other tones.

### 3.2. Machine learning classification

Results from machine learning classification are summarized in Table 1.

**Table 1:** Machine learning classification accuracy of the four Mandarin tonal categories, averaged from 100 trials, 10-fold CV

| **Classification Using Only F0** | | | |
|---|---|---|---|
| **Age Group** | **LDA** | **SVM** | **RF** |
| Age 4-5 | 54.246% | 57.667% | 57.544% |
| Age 7-8 | 63.879% | 68.348% | 68.121% |
| Age 10-11 | 68.230% | 70.885% | 73.092% |
| Adults | 60.210% | 64.741% | 71.185% |
| **Classification Using Only Spectral Cues** | | | |
| **Age Group** | **LDA** | **SVM** | **RF** |
| Age 4-5 | 50.526% | 55.702% | 54.491% |
| Age 7-8 | 66.606% | 72.409% | 73.121% |
| Age 10-11 | 66.299% | 74.264% | 72.977% |
| Adults | 70.148% | 79.123% | 75.864% |
| **Classification Using Spectral Cues & F0** | | | |
| **Age Group** | **LDA** | **SVM** | **RF** |
| Age 4-5 | 59.228% | 62.614% | 62.070% |
| Age 7-8 | 77.803% | 80.455% | 80.348% |
| Age 10-11 | 77.908% | 81.678% | 81.034% |
| Adults | 78.556% | 84.247% | 82.407% |

Using just F0, the algorithms achieved higher accuracy for older children, with the highest of 73% for the group Age 10-11 by the Random Forest (RF) model. The classification accuracy for the Adults was in fact lower than Age 10-11(RF), or even Age 7-8(LDA&SVM). Since the highest accuracy was reached by the age group 10-11, we concluded that by the age of 10-11, speakers were able to make clear, even maximal tonal contrasts using F0.

When only spectral cues were used as features, the highest accuracy of 79% was achieved by the SVM model for the adult data. Results from using spectral cues again displayed a developmental pattern into adulthood, ranging from 56% for Age 4-5, 72% for Age 7-8, 74% for Age 10-11, and 79% for the Adults (SVM). The results suggested that the integration of spectral cues in producing the tonal contrasts was still not yet fully adult-like by the age of 10-11.

Finally, tone classification achieved the highest accuracy of 84% using both F0 and spectral cues, and for the Adults group. In general, classification using both F0 and spectral cues had the best accuracy for every age group. This suggested that at all ages, spectral cues were enhancing tonal contrasts to some extent. The largest jump in accuracy (17%, SVM) was observed between Age 4-5 and Age 7-8, indicating that spectral cues begin to play a more important role for tonal distinction from age 7-8 onward.

## 4. DISCUSSION AND CONCLUSION

This study investigated how spectral cues are integrated in the tonal production of children and adults. The results of the both MDS and machine learning classification corroborated the finding that even after children acquire the basic F0 contrasts, the contrastiveness of the tonal categories continues to improve, by integrating spectral information in the voice. As shown in Figure 1 and Table 1, for the adults, all tones were so clearly distinguished by the spectral cues that adding F0 cues had minimal effect in establishing maximal tonal contrast. This finding is consistent with the result in [15], which is based on a different corpus and recording set-up. For the children, the youngest group in our study (Age 4-5) produced tonal contrast based on F0 cues; however, by the age of 7-8, children were capable of producing a more efficient yet reliable tonal contrast, using a larger set of available cues. Furthermore, this development in the integration of spectral cues in their tonal production, which is incomplete at age 10-11 and has a sharp curve between ages 4-5 and 7-8, occurs independently of the development of the F0 cues in tonal contrastiveness, which shows a more gradual improvement between the ages 4 to 11, and does not continue into adulthood.

As tone production and perception involve rich information from the voice source, it is important to investigate cues beyond F0. In this study, we have shown that spectral cues become increasingly useful in manifesting tone contrastivity. This developmental pattern can be explained in two ways: (i) children actively learn to integrate spectral cues in their production of tones as part of their linguistic development, or (ii) pitch and voice quality are physiologically correlated differently for children and adults, due to differences in the physiological control over their voices. Though both explanations are possible, the finding that tonal contrastiveness using just F0 cues are no better for the adults than for the older children suggests that adults have learned to incorporate other useful cues in tone production, and thus do not need to rely solely on F0 to make the contrast. A perception study on the development of children's sensitivity to cues beyond F0 will be done to verify this explanation.

## 5. REFERENCES


[1] Belotel-Grenié, A., Grenié, M. 1994. Phonation types analysis in standard chinese. *Third International Conference on Spoken Language Processing*.

[2] Chen, A. 2011. Tuning information packaging: intonational realization of topic and focus in child dutch. *Journal of Child Language* 38(5), 1055–1083.

[3] Chen, F., Yan, N., Wang, L., Yang, T., Wu, J., Zhao, H., Peng, G. 2015. The development of categorical perception of lexical tones in mandarin-speaking preschoolers. *Sixteenth Annual Conference of the International Speech Communication Association*.

[4] Hua, Z., Dodd, B. 2000. The phonological acquisition of putonghua (modern standard chinese). *Journal of child language* 27(1), 3–42.

[5] Iseli, M., Shue, Y.-L., Alwan, A. 2007. Age, sex, and vowel dependencies of acoustic measures related to the voice source. *The Journal of the Acoustical Society of America* 121(4), 2283–2295.

[6] 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 sounds1. *Speech communication* 27(3-4), 187–207.

[7] Kuang, J. 2017. Covariation between voice quality and pitch: Revisiting the case of mandarin creaky voice. *The Journal of the Acoustical Society of America* 142(3), 1693–1706.

[8] Kuang, J. 2018. The influence of tonal categories and prosodic boundaries on the creakiness in mandarin. *The Journal of the Acoustical Society of America* 143(6), EL509–EL515.

[9] Kuang, J., Liberman, M. 2018. Integrating voice quality cues in the pitch perception of speech and non-speech utterances. *Frontiers in Psychology* 9, 2147.

[10] Li, C. N., Thompson, S. A. 1977. The acquisition of tone in mandarin-speaking children. *Journal of Child Language* 4(2), 185â199.

[11] Liaw, A., Wiener, M. 2002. Classification and regression by randomforest. *R News* 2(3), 18–22.

[12] Liu, S., Samuel, A. G. 2004. Perception of mandarin lexical tones when f0 information is neutralized. *Language and speech* 47(2), 109–138.

[13] 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.

[14] Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Szoecs, E., Wagner, H. 2018. *vegan: Community Ecology Package*. R package version 2.5-2.

[15] Ryant, N., Yuan, J., Liberman, M. 2014. Mandarin tone classification without pitch tracking. *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE 4868–4872.

[16] Shue, Y.-L., Keating, P. A., Vicenik, C., Yu, K. 2011. Voicesauce: A program for voice analysis. *ICPhS*.

[17] Sundberg, J. 1994. Vocal fold vibration patterns and phonatory modes. *STL-QPSR* 35, 69–80.

[18] Sundberg, J., Titze, I., Scherer, R. 1993. Phonatory control in male singing: a study of the effects of subglottal pressure, fundamental frequency, and mode of phonation on the voice source. *Journal of Voice* 7(1), 15–29.

[19] Tsao, F.-M. 2008. The effect of acoustical similarity on lexical-tone perception of one-year-old mandarin-learning infants. *Chin. J. Psychol.* 50(2), 111–124.

[20] Tsao, F.-M. 2017. Perceptual improvement of lexical tones in infants: effects of tone language experience. *Frontiers in psychology* 8, 558.

[21] Venables, W. N., Ripley, B. D. 2002. *Modern Applied Statistics with S*. New York: Springer fourth edition. ISBN 0-387-95457-0.

[22] Wong, P. 2013. Perceptual evidence for protracted development in monosyllabic mandarin lexical tone production in preschool children in taiwan. *The Journal of the Acoustical Society of America* 133(1), 434–443.

[23] Yang, A., Chen, A. 2014. Prosodic focus-marking in chinese four-and eight-year-olds. *Speech Prosody 2014* 713–717.

[24] Yang, A., Chen, A. 2018. The developmental path to adult-like prosodic focus-marking in mandarin chinese-speaking children. *First Language* 38(1), 26–46.

[25] Yang, R.-X. 2011. The phonation factor in the categorical perception of mandarin tones. *Proceedings of ICPhS XVII* 2204–2207.

[26] Yuan, J., Ryant, N., Liberman, M. 2014. Automatic phonetic segmentation in mandarin chinese: boundary models, glottal features and tone. *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE 2539–2543.