Speech Prosody 2024
2-5 July 2024, Leiden, The Netherlands

Acoustic-prosodic Analysis for Mandarin Disyllabic Words Conveying Vocal Emotions

Xuyi Wang[1,2], *Hongwei Ding*[1,2]

[1]Speech-Language-Hearing Center, School of Foreign Languages, Shanghai Jiao Tong University
[2]National Research Centre for Language and Well-being, Shanghai, China

wxy_evie@sjtu.edu.cn, hwding@sjtu.edu.cn

Abstract

This study conducted a comprehensive analysis of features using a validated audiometry corpus comprising 450 Mandarin Chinese disyllabic words across five emotional states: "Angry," "Sad," "Happy," "Fearful," and "Neutral," produced by both male and female speakers. Employing machine-learning tools, the research identified and elucidated crucial acoustic-prosodic features for emotional vocalization. Results revealed several key points: First, the models showed that fear was acoustically the most recognizable emotion, while joy presented most difficulties. Second, in the identification of Mandarin emotional prosody, the spectrum characteristics like formant energy ratios were of primary significance, followed by those F0-related parameters such as the 20th and 80th percentiles of F0. Third, data of formant energy ratios mainly indicated that fearful voices were more turbulent, and those of F0-related features suggested a general increase in pitch for emotional speech. Moreover, considerable cross-speaker variations in affective vocalization strategies were observed, reflected in distinct feature patterns that our speakers exploited for their emotional expressions. Despite the considerable audio samples gathered from each speaker, the current corpus remains limited by its two-speaker scale. Nonetheless, ongoing efforts involve expanding the corpus with additional speakers. The scalability and replicability of the paradigm can facilitate seamless transplantation for future investigations.

Index Terms: Mandarin emotional prosody, acoustic-prosodic analysis, machine learning, speech data

1. Introduction

In oral communication, we employ a wide range of vocal cues, together composing speech prosody, to better comprehend and convey intended messages. Through the modulation of multiple acoustic-prosodic features (e.g., F0, loudness, duration, speech rate, etc.), speech prosody encompasses various essential functions, conveying semantic, grammatic, pragmatic, and emotional information [1]. Thus, as a considerable part of our pragmatic and emotive communication capacity is dependent on the prosodic encoding and decoding of acoustic parameters [2], people with auditory deficits meet extra difficulty when they try to perceive and express their emotive feelings [3, 4].

To better present affective speech with hearing assistive devices, we need a better understanding of the mechanism underlying emotional prosody. Aiming to provide new research evidence, the Chinese Emotional Speech Auditory Project (CESAP), carried jointly by our team and a company of hearing care devices, attempts to investigate the Mandarin emotional prosody. As part of the CESAP project, the current study conducts an acoustic-prosodic feature analysis on a balanced, validated set of standard audiometry materials. The aim of this study is to examine and measure the correlation between the parameters and emotions in Mandarin Chinese, providing a foundation for future research and development.

Reviewing previous research on Mandarin emotional prosody, we find room for improvement. First, we meliorate the current literature with a standard feature set. Most studies on Mandarin emotional prosody are restricted to a few classic prosodic features, such as pitch, duration, and intensity [5–7]. Though a growing amount of recent research expands their examination scope to other source-filter features concerning phonation type, vowel space and articulatory clues [8–13], little literature include spectral-related features like mel-frequency cepstral coefficients (MFCCs), which nevertheless, are suggested to be grounded in human auditory processing [14, 15] and is playing an increasingly dominant role in machining-learning based speech emotion recognition (SER) task solutions [16]. To complete the depiction of Mandarin emotional prosody, it is necessary to include spectral features. In the present study, we choose the extended Geneva Acoustic Minimalistic Parameter Set (eGeMAPS) [17], which contains a wide range of frequency-, amplitude-, temporal- and spectral-related features efficient for affective computing [18], yet keeps a satisfiable light size. Second, we employ machine learning (ML) tools to assist our feature data analysis. Most previous studies of Chinese emotional prosody take basic descriptive and inferential methods. They presume a limited number of important features as priori, and then investigate the statistics one by one for each parameter in a half-manual style. Such a paradigm can hardly cope with the large amount of feature data obtained from an extended parameter set, and its restricted vision cuts off access to further inspection of the nuanced acoustic characteristics in speech data. Therefore, in addition to the traditional statistical analytical tools, we adopted the XGBoost (extreme gradient boosting) ML model, which has been widely used and proved to be useful in feature importance analysis for vocal speech [13, 19], to help us screen, measure, and sort the parameters.

In conclusion, the purpose of our ongoing acoustic-prosodic analysis of Chinese emotional prosody is to draw a more comprehensive descriptive mapping between acoustic-prosodic features and perceptive vocal emotions with a research framework of commendable scalability. Specifically, we aim to answer which acoustic parameters are the most important and how they change along with different affective expressions.

2. Methods

2.1. Speech data

The materials used in the current study contained 15 lists from the CESAP corpus [20], each consisting of 30 Mandarin disyllabic words. All words were semantically neutral lexicons

10.21437/SpeechProsody.2024-193

frequently used in everyday contexts. Within each list, the segmental phonetic characteristics (the target number of vowels, consonants, and lexical tones) of all lexicons were rigorously balanced to exclude their effect on intonation [21]. All items were pronounced with five emotions (neutral, anger, fear, joy, and sadness) by a female and a male amateur voice actor with a 1B level certificate in the Chinese National Mandarin Proficiency Test, resulting in a total number of 4,500 (15×30×5×2) audio source files. The audio clips were recorded at a sampling rating of 44.1 kHz with 16-bit quantization. The signal-noise ratio (SNR) was controlled at approximately 50 dB and the speaking rate was manipulated at 932 ± 81 ms/disyllable (851 – 1013 ms/disyllable). All chosen items were proved by validation tests to be perceptually matched with their intended emotional expressions.

2.2. Feature extraction

We used the Python package OpenSMILE (version 2.4.1, in Python 3.9.12) [22, 23] to extract 88 eGeMAPS parameters from the materials. The eGeMAPS set was created to serve as a foundation for affective speech processing to facilitate replication and enhance parameter comparability. For detailed descriptions of the features, please see [17].

2.3. Feature analysis

Our data analysis comprised two subparts: an ML-based analysis with the XGBoost algorithm, and a follow-up t-test analysis. In the former step, we measured the importance of all 88 parameters and ranked them accordingly to narrow down later examination scope. In the second step, we further investigated the variation patterns of the top-ranked parameters that contribute substantially to the models.

2.3.1. ML-based analysis

Given that the size of our dataset is relatively limited, the XGBoost classifier was adopted to lower the possibility of overfitting [24]. This algorithm is a gradient boosting algorithm that utilizes decision trees, and has demonstrated high performance on datasets of limited size [13]. The hyperparameters of the XGBoost models were set to the values that yielded the highest performance after multiple trials in a 5-class speech emotion classification task, and were not further optimized. We randomly split the test and training sets with stratified sampling methods at a ratio of 2:8. Two XGBoost models were trained, one on the male feature data, and another on the female data.

In order to illustrate the contribution of each parameter to the model accuracy in recognition of emotions, we computed the importance scores, which are provided in the form of Gain in XGBoost [25]. The nine parameters that ranked top respectively in the importance lists of the male and female models were included for further investigation, as they produced the first turning points in the post-hoc model training. We employed the XGBoost [24] Python package to build the model.

2.3.2. Mean-value-based analysis

To examine variations in acoustic parameters across different emotions, we calculated z-scores for the top nine acoustic parameters in the four emotions using the mean score and standard deviation of the neutral recordings for each speaker. Follow-up t-tests were done between emotions with SciPy package.

3. Results

The results are presented in three parts. First, we report the overall performance of the male and female classification models. Then, we display an overview of the XGBoost feature importance gain of all 88 parameters in the models. Finally, we further examine the variation patterns of the top ranked features in the two models.

3.1. Model performance

Both of the two XGboost models achieved admirable performances on the classification task, with a slightly higher accuracy score of 97.78% in the male model compared to that of 97.56% in the female model. Other performance measurements of the male and female models are shown in Table 1. Noticeably, both models had the highest accuracy in distinguishing instances of the fearful emotion ($F1_m = 100\%$, $F1_f = 99.44\%$) and performed well in discerning sad voices ($F1_m = 99.45\%$, $F1_f = 97.78\%$), but had relatively more difficulty in recognizing joyful voices ($F1_m = 95.45\%$, $F1_f = 96.70\%$). The male model exhibited converging errors distinguishing between joy and anger, misclassifying joyful speech as furious. In contrast, the female model's inaccurate predictions were distributed across sad, neutral, joyous, and furious voices, with the greatest level of uncertainty occurring between neutral and joyful emotion states.

Table 1: *Model performance with male and female speech data*

Score (%)	Precision		Recall		F1-score	
	M	F	M	F	M	F
Fearful	100	100	100	98.89	100	99.44
Sad	98.90	97.78	100	97.78	99.45	97.78
Neutral	98.88	96.63	97.78	95.56	98.32	96.09
Joyful	97.67	95.65	93.33	97.78	95.45	96.70
Angry	93.62	97.78	97.78	97.78	95.65	97.78

3.2. Feature importance

The feature importance patterns displayed in Figure 1 have a similar overall shape. Both male and female models prioritized parameters related to frequency, such as *F0Pct20, F0Pct80*, and spectral amplitude, specifically *F3AmpStd* (see Figure 2).

Nevertheless, despite the similarity, there were considerable variances in the specific types of the most essential features employed for emotion identification between the male and female models. The male speaker's top-ranked nine features in the XGBoost model included one amplitude-related (*LoudStd*), two frequency-related (*F0Pct20, F0Pct80*), five spectral-related (*F3AmpStd, F1AmpStd, V0-500, UV0-500, V500-1500*), and one temporal-related parameter (*LoudPeak*), producing an impressive turning point accuracy of 96%. The female speaker's model depended less on temporal characteristics, with three frequency-related (*F1band, F0Pct20, F0Pct80*), five spectral-related (*F3AmpStd, F3AmpMean, FluxV, MFCC1V, MFCC2*), and one amplitude-related parameters (*LoudFallStd*) stated on the top, resulting in an admirable accuracy of 93.55% at the turning point. The post-hoc training and testing of the nine top features in each model were conducted using identical parameters and test-training sets as previously employed.

3.3. Feature variation patterns

To better demonstrate the feature variation patterns, we follow previous descriptions and classify the selected top parame-

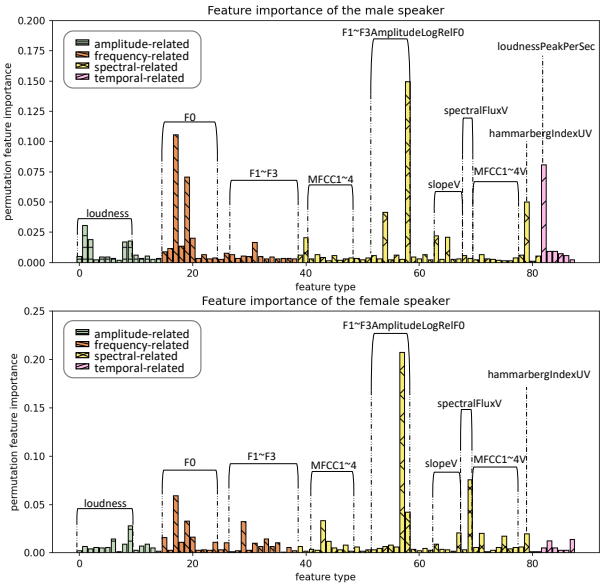

Figure 1: *An overview of the feature importance for male and female models. Parameters are divided into four categories; A) amplitude-related parameters, B) frequency-related parameters, C) spectral-related parameters, and D) temporal-related parameters.*

ters distinguished by the models into four groups: frequency-related, spectral-related, amplitude-related, and temporal-related features. The following subsections present the mean-value-based results for each group.

3.3.1. Frequency-related features

According to our results, when it comes to the classification of emotional expressions, the crucial information regarding frequency for both male and female individuals was not conveyed by the basic F0 values. Instead, it was better represented by the 20th and 80th percentiles of F0.

In general, emotional speech demonstrated an increase in pitch compared to neutral voices, regardless of whether it is measured by *F0Pct20* or *F0Pct80*. However, there was a distinction in the upward movement patterns between male and female speakers. In female utterances, the *F0Pct20* and *F0Pct80* values of different emotions showed a comparable level of rise, with the most prominent improvement observed in fear ($t_{20} = 86.99$, $p < .000$; $t_{80} = 80.71$, $p < .000$), followed by joy ($t_{20} = 37.46$, $p < .001$; $t_{80} = 58.71$, $p < .000$) and anger ($t_{20} = 48.28$, $p < .001$; $t_{80} = 48.59$, $p < .001$). Meanwhile, the increase in F0 for sadness was relatively smaller ($t_{20} = 18.60$, $p < .001$; $t_{80} = 4.36$, $p < .001$). For the male speaker, the increments for corresponding emotions in terms of *F0Pct20* values displayed a similar growth pattern to that of the female. But the rise in *F0Pct80* exhibited a discrepancy, as the most significant increase was observed in the emotion of joy ($t = 22.17$, $p < .001$) rather than fear ($t = 13.68$, $p < .001$), and the increase in *F0Pct80* for sad speech was statistically non-significant compared against neutral utterances ($t = 1.11$, $p = 0.267$). Furthermore, it is evident that the overall rise in pitch of male emotional speech was not as substantial as that of female speech (the highest z-score for value increase was above 5 in female's, but only slightly above 4 in male's). Nevertheless, it is suggested that the male compensated for this by reducing his F0 range between 20th and 80th percentiles to achieve comparable expressions of the emotions.

Furthermore, the model constructed using female speech also identified an extra parameter of F1 bandwidth. The *F1band* value of furious utterances exhibited a notable decrease ($t = -15.19$, $p < .001$), whereas the values for other emotions demonstrated a corresponding increase ($t_{joy} = 10.51$, $p_{joy} < .001$; $t_{sadness} = 17.21$, $p_{sadness} < .001$; $t_{fear} = 15.00$, $p_{fear} < .001$). This suggests that female furious voices were likely to be less yawning, while others were more so [26].

3.3.2. Spectral-related features

Among spectral-related parameters, the relative energy of F3 (*F3Amp*) was arguably the most predominant feature listed by both male and female models. Formant energy levels can serve as indicators of the level of expressive nasality or forceful glottalization in speech [26]. The male speech data analysis revealed that the *F3AmpStd* and *F1AmpStd* parameters were crucial for emotion identification, while the female model mostly depended on the *F3AmpMean* and *F3AmpStd* parameters. The male speaker's expressions of sadness were characterized by significantly lower *F3AmpStd* ($t = -21.40$, $p < .001$) and *F1AmpStd* ($t = -12.52$, $p < .001$) values, suggesting greater stability in terms of nasal and glottal efforts. Meanwhile, the male voices conveying fear had higher *F3AmpStd* ($t = 22.26$, $p < .001$) and *F1AmpStd* ($t = 36.64$, $p < .001$) values, indicating increased turbulence. In the female data, angry voices had higher *F3AmpMean* ($t = 11.31$, $p < .001$), which denoted a greater level of vocal energy. Conversely, terrified voices of the female speaker were characterized with lower *F3AmpMean* ($t = -49.42$, $p < .001$) and higher *F3AmpStd* ($t = 36.37$, $p < .001$), displaying weaker and more fluctuating forces.

In addition, the male speaker model placed significant emphasis on spectral-slope parameters, including *V0-500*, *UV0-500*, and *V500-1500*. Angry, sorrowful, and joyful voices showed a rise in *V0-500* and *UV0-500* values, whilst only the value of fear demonstrated a reduction. Nevertheless, fearful voices possessed the highest values in *V500-1500*, which signified a transfer of faster energy increase from low frequency to higher frequency with presence of fear.

On the other hand, the female speaker model underlined MFCC features and the spectral flux feature. All emotional speech had lower *MFCC2* values, with furious voices taking the lowest. Moreover, with the exception of utterances with sad emotion, all emotions showed reductions in *MFCC1V*; simultaneously, sadness also possessed the highest value for the *FluxV* parameter, whereas all others scored lower in that.

3.3.3. Amplitude-related and temporal-related features

The male showcased an amplitude-related and a temporal-related parameter, which were the standard deviation of loudness (*LoudStd*) and loudness peak per second (*LoudPeak*) respectively. The *LoudStd* values increased in angry voices, marking wobbly intensity, and decreased in sad and fearful utterances, producing relatively steadier speech volume. The *LoudPeak* parameter further revealed that in a longer time domain, expressions of fear failed to remain stable, as they had high *LoudPeak* values, whereas speech of sadness steadily possessed lower values.

Meanwhile, the female model sorted one amplitude-related feature and no temporal parameters. In the female speech data, the standard deviation values of loudness falling slope (*LoudFallStd*) were higher with expressions of anger and joy, and lower with sorrow or fear.

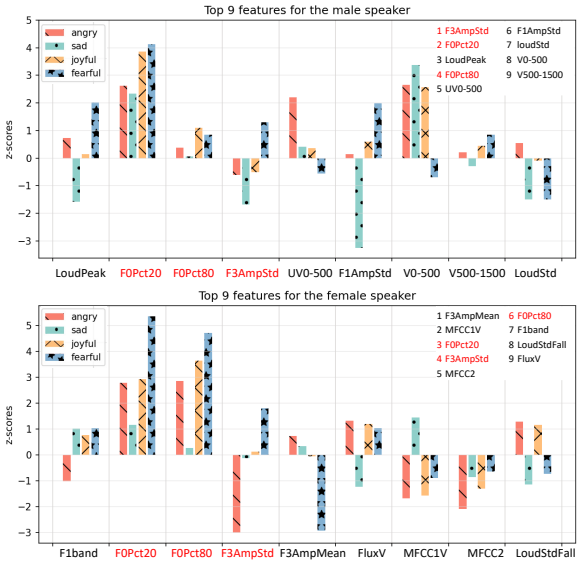

Figure 2: *Variation patterns and ranking lists of the top nine features for male and female models. LoudPeak = Rate of loudness peaks, F0Pct20 = Percentile 20th of F0, F0Pct80 = Percentile 80th of F0, F3AmpStd = Standard deviation of formant 3 relative energy, F3AmpMean = Mean value of formant 3 relative energy, F1AmpStd = Standard deviation of formant 1 relative energy, V0-500 = Spectral Slope V 0-500 Hz, UV0-500 = Spectral Slope UV 0-500 Hz, V500-1500 = Spectral Slope V 500-1500 Hz, LoudStd = Standard deviation of loudness, F1Band = Formant 1 bandwidth, FluxV = Spectral Flux V, LoudFallStd = Standard deviation of loudness falling slope.*

4. Discussion

This study analyzed Mandarin emotional prosody with a comprehensive parameter set covering feature data from multiple acoustic domains. To better extract in-depth information, we utilized the ML tool of XGBoost algorithm, which helped us assess and rank the features based on their contributions to the models. The results suggested intricate relationships between acoustic parameters and emotions, as the predictive capabilities of certain factors differ across moods and speaker genders.

Our work suggests that joy is the most challenging one for acoustic detection of vocal emotion, while fear is the most strongly predicted one. The XGBoost models built on male and female speech data both had the poorest classification performance on the prediction of joy, which is compatible with the common conclusion in SER research [27]. Previous perceptual validation experiments conducted in our project indicated that compared to other emotions, happiness can be recognized relatively poorly by Chinese Mandarin speakers, which also agrees with prior findings in typical Mandarin listeners [28] and CI Mandarin listeners [29]. Meanwhile, according to the feature analysis of the collected data in our study, comparable to previous observations in other languages [18], joy in Mandarin prosody also possessed moderate values across parameters, barely having any strongly distinguishable characteristics. Given that happiness is among the emotions best recognized by visual [30], a possible reason might be that people rely more on smiling faces to recognize joyful emotion, so this affection may not require strongly discernible acoustic characteristics.

The findings underscore the significance of spectral- and frequency- related parameters in identifying emotions. Regarding frequency-related features, the XGBoost classifiers suggested a heavy dependency on the parameters of *F0Pct20* and *F0Pct80*. Compatible with recent studies [18], our study also identified a considerable number of spectral-related features, including formant energy ratios, spectral slope, MFCCs, and spectral flux. Moreover, the results highlighted a primary reliance on the parameters concerning the relative energy of F3 (*F3Amp*). Noticeably, the spectral-related features gained more importance with their contributions to the classification accuracies of the models. This indicates that Mandarin emotional prosody may prioritize the utilization of spectrum features above frequency-related aspects. Because of the existence of lexical tones in Mandarin Chinese, the range of pitch change in emotional prosody is limited [31], so Mandarin speakers might lean on spectrum clues to express emotions in their speech.

Regarding the patterns of variance, our research uncovers a few prevalent shifting trends. Both male and female speakers exhibited statistically significant pitch rise in their pronunciation of practically all emotions, except for a non-significant increase observed in the presentation of sorrow in male speech. Meanwhile, stability of nasal and glottal energy measured by the standard deviation of formant energy ratio was found to be prominent features that helped differentiating fearful emotion, which was characterized with turbulent vocal force. Nonetheless, we should notice that emotions can be verbalized through a number of different acoustic cues, the tactics employed by different speakers could exhibit considerable variations [32]. This accounts for the conflicting findings in previous studies as well as the diversity in the observed features and their variation patterns between speakers in the current study. In other words, the affective vocalization is a multidimensional process, in which one may rely less on one feature, and instead enhance another as a compensation. For example, in contrast to the female speaker, the male speaker did not employ F0 alleviation as the primary characteristic for expressing dread. Instead, he reduced the range of F0 between *F0Pct20* and *F0Pct80* as a form of compensation. Future study may focus on the multimodal aspects of prosodic emotionalization and further our comprehension of the synergistic functions of multiple auditory cues.

The limitations of this project can be addressed from two aspects. First of all, the two-speaker corpus employed in our research is quite limited in size. The patterns we observed in our research could be inevitably affected by individual articulatory characteristics. Therefore, we are extending the corpus by incorporating more speakers for further examination. Second, due to limited space, we only present statistics of the top nine features for each speaker. There are parameters we fail to cover, but we shall further elaborate on them in following studies.

5. Conclusions

This paper undertakes an extensive acoustic-prosodic analysis to advance our understanding of emotive vocalization in Mandarin speech. Our work enriches the phonetic knowledge of Mandarin emotional prosody by describing a comprehensive range of features and, at the same time, by employing a research framework that introduces ML tools. It provides a groundwork for future resolutions to the specific issue of emotion enhancement in hearing devices.

6. Acknowledgements

This study was supported by the major programs of the National Social Science Foundation of China [18ZDA293]. The corresponding author is Hongwei Ding (hwding@sjtu.edu.cn).

References

[1] H. Ding and Y. Zhang, "Speech prosody in mental disorders," *Annu. Rev. Linguist.*, vol. 9, no. 1, pp. 335–355, 2023.

[2] V. Lucarini, M. Grice, F. Cangemi, J. Zimmermann, C. Marchesi, K. Vogeley, and M. Tonna, "Speech prosody as a bridge between psychopathology and linguistics: The case of the schizophrenia spectrum," *Front. Psychiatry*, vol. 11, p. 531863, 09 2020.

[3] Y.-S. Lin, C. Wu, C. Limb, H. Lu, I. Feng, S.-C. Peng, M. Deroche, and M. Chatterjee, "Voice emotion recognition by Mandarin-speaking pediatric cochlear implant users in taiwan," *Laryngoscope Investig. Otolaryngol.*, vol. 7, pp. 1–9, 01 2022.

[4] M. Karimi-Boroujeni, H. R. Dajani, and C. Giguère, "Perception of prosody in hearing-impaired individuals and users of hearing assistive devices: An overview of recent advances," *J. Speech Lang. Hear. Res.*, vol. 66, no. 2, pp. 775–789, 2023.

[5] H.-Y. Lin and J. Fon, "Prosodic and acoustic features of emotional speech in Taiwan Mandarin," in *Proc. Speech Prosody 2012*, 2012, pp. 450–453.

[6] H. Wang, A. Li, and Q. Fang, "F0 contour of prosodic word in happy speech of Mandarin," in *Proc. 9th Int. Conf. Affect. Comput. Intel. Interaction*, 2005, pp. 433–440.

[7] S. B. Zhang, P. C. Ching, and F. Kong, "Acoustic analysis of emotional speech in Mandarin Chinese," in *Int. Symp. Chinese Spoken Language Processing*, 2006, pp. 57–66.

[8] D. Erickson, C. Zhu, S. Kawahara, and A. Suemitsu, "Articulation, acoustics and perception of Mandarin Chinese emotional speech," *Open Linguist.*, vol. 2, no. 1, 2016.

[9] A. Li, Q. Fang, F. Hu, L. Zheng, H. Wang, and J. Dang, "Acoustic and articulatory analysis on Mandarin Chinese vowels in emotional speech," in *7th Int. Symp. China Spoken Language Processing*, 11 2010, pp. 38–43.

[10] P. Liu and M. Pell, "Recognizing vocal emotions in Mandarin Chinese: A validated database of chinese vocal emotional stimuli," *Behav. Res. Methods*, vol. 44, pp. 1042–1051, 04 2012.

[11] T.-L. Pao, Y.-T. Chen, J.-H. Yeh, and W.-Y. Liao, "Detecting emotions in Mandarin speech," *Int. J. Computer Linguist. and Chinese Language Processing*, vol. 10, 2004.

[12] J. Yuan, L. Shen, and F. Chen, "The acoustic realization of anger, fear, joy and sadness in Chinese," in *Proc. 7th Int. Conf. Spoken Language Processing*, 2002, pp. 2025–2028.

[13] Z. Zhang, M. Huang, and Z. Xiao, "A study of correlation between physiological process of articulation and emotions on Mandarin Chinese," *Speech Commun.*, vol. 147, pp. 82–92, 2023.

[14] T. Schatz, N. H. Feldman, S. Goldwater, X.-N. Cao, and E. Dupoux, "Early phonetic learning without phonetic categories: Insights from large-scale simulations on realistic input," *Proc. Natl. Acad. SCI USA*, vol. 118, no. 7, p. e2001844118, 2021.

[15] Y. Matusevych, T. Schatz, H. Kamper, N. H. Feldman, and S. Goldwater, "Infant phonetic learning as perceptual space learning: A crosslinguistic evaluation of computational models," *Cogn. Sci.*, vol. 47, no. 7, p. e13314, 2023.

[16] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Commun.*, vol. 116, pp. 56–76, 2020.

[17] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for voice research and affective computing," *IEEE Trans. Affect. Comput.*, vol. 7, no. 2, pp. 190–202, 2016.

[18] M. Ekberg, G. Stavrinos, J. Andin, S. Stenfelt, and Dahlström, "Acoustic features distinguishing emotions in Swedish speech," *J. Voice*, 2023.

[19] J. Sol, M. Aaen, C. Sadolin, and L. ten Bosch, "Towards automated vocal mode classification in healthy singing voice—an XGBoost decision tree-based machine learning classifier," *J. Voice*, 2023.

[20] E. Tang, J. Gong, J. Zhang, J. Zhang, R. Fang, J. Guan, and H. Ding, "Chinese Emotional Speech Audiometry Project (CESAP): Establishment and validation of a new material set with emotionally neutral disyllabic words," *J. Speech Lang. Hear. Res.*, 2024 in press.

[21] M. Liu, Y. Chen, and N. O. Schiller, "Context matters for tone and intonation processing in Mandarin," *Lang. Speech*, vol. 65, no. 1, pp. 52–72, 2022.

[22] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the Munich versatile and fast open-source audio feature extractor," *Proc. 18th ACM int. conf. Multimedia*, 2010.

[23] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the Munich open-source multimedia feature extractor," 10 2013, pp. 835–838.

[24] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[25] Z. Huiting, J. Yuan, and L. Chen, "Short-term load forecasting using EMD-LSTM neural networks with a XGBoost algorithm for feature importance evaluation," *Energies*, vol. 10, p. 1168, 08 2017.

[26] S. A. Memon, "Acoustic correlates of the voice qualifiers: A survey," *ArXiv*, vol. abs/2010.15869, 2020.

[27] C. Dogdu, T. Kessler, D. Schneider, M. Shadaydeh, and S. Schweinberger, "A comparison of machine learning algorithms and feature sets for automatic vocal emotion recognition in speech," *Sensors*, vol. 22, p. 7561, 10 2022.

[28] P. Liu and M. Pell, "Processing emotional prosody in Mandarin Chinese: A cross-language comparison," in *Proc. Speech Prosody 2014*, 2014, pp. 95–99.

[29] C. Pak and W. Katz, "Recognition of emotional prosody by Mandarin-speaking adults with cochlear implants," *J. Acoust. Soc. Am.*, vol. 146, pp. EL165–EL171, 08 2019.

[30] H. Elfenbein and N. Ambady, "On the universality and cultural specificity of emotion recognition: A meta-analysis," *Psychol. Bull.*, vol. 128, pp. 203–35, 03 2002.

[31] T. Wang and Y.-C. Lee, "Does restriction of pitch variation affect the perception of vocal emotions in Mandarin Chinese?" *J. Acoust. Soc. Am.*, vol. 137, p. EL117, 2015.

[32] C. Martinuzzi and J. Schertz, "Sorry, not sorry: The independent role of multiple phonetic cues in signaling the difference between two word meanings," *Lang. Speech*, vol. 65, no. 1, pp. 143–172, 2022.