INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Speech Entrainment in Chinese Story-Style Talk Shows: The Interaction Between Gender and Role

Yanting Sun[1], *Hongwei Ding*[2]

[1]Institute of Corpus Studies and Applications, Shanghai International Studies University, China

[2]Speech-Language-Hearing Center, School of Foreign Languages, Shanghai Jiao Tong University, China

sunyanting@shisu.edu.cn, hwding@sjtu.edu.cn

Abstract

Speech entrainment is evident in short-and-short turn-taking, but entrainment in long-and-short turn-taking, like talk shows, is expected to be different but also evident. We examined three prosodic feature sets of pitch, intensity, and duration (speaking rate) to explore the impact of gender and role interaction between the host and guests on speech entrainment in a Mandarin Chinese talk show corpus. This research consistently showed that intensity remained the robust entraining feature, and the speaking rate was steadily divergent. Another vital result was that the rare occurrence of the final-rising pitch in question types led to dynamic local positive proximity. Besides, it was interesting to note that mixed-gender pairs with different roles showed more dynamic local positive proximity and synchrony on intensity and pitch than same-gender pairs. Taken together, these results suggest the complexity of gender and role interaction on speech entrainment in long-and-short turn-taking.

Index Terms: speech entrainment, prosody, gender, role

1. Introduction

Speech entrainment is similar speech behavior between interlocutors in conversations [1]. Entrainment occurs in many elements of speech, and in all articulatory, rhythmic, and phonatory dimensions. Most studies on prosody entrainment have been conducted mainly using the three acoustic-prosodic features of duration, intensity, and pitch [2-4]. Therefore, the present work focused on the above three acoustic-prosodic features in speech entrainment because speakers adapt prosody to that of their interlocutors and align in conversation to ensure smooth and successful communication [5].

Previous studies mainly adopted the entrainment categorization system from Levitan and Hirschberg's [5] entrainment classification framework. It has recently been extended by Wynn and Borrie [1] to include eight entrainment types divided by three classification factors (class, level, and dynamicity). For the class, "proximity" is the similarity in speech features between interlocutors. "Synchrony" is the similarity in the movement of speech features. As for the level, "local" is the similarity between units that are equal to or smaller than adjacent turns. "Global" is the similarity across any time scale greater than adjacent turns. For dynamicity, "static" is similarity without the statistical consideration of changes over time. "Dynamic" is a change in similarity over time. All in all, the highlights are reflected in two aspects. First, they consider "convergence" as a subtype of proximity; second, entrainment is further divided into static and dynamic types. In this study, therefore, Wynn and Borrie's framework was adopted for an initial attempt, as it could provide consistent terminology and clearer entrainment classification.

The above classification involves studying the positive (convergent) or negative (divergent) valence of entrainment, which is highly variable in social factors such as gender and role of the interlocutor. However, numerous studies have focused only on the effects of gender or role alone on speech entrainment [3-9], highlighting the need to investigate the impact of gender and role on speech entrainment to explain the inconsistencies mentioned above. Pardo [10] suggested that males in a dependent role entrained more than those in a position of power in a task conversation. Kendall [11] indicated that the interviewee's speaking rate was stronger influenced by the interviewer's gender. Reichel *et al*. [12] found gender-dependent strategies in cooperative interactions, i.e., female describers and male followers entrained most, while male describers and female followers least.

Although the above-mentioned attempts have been made to explore entrainment in short-and-short turn-taking, such as task-oriented conversations [3, 5, 12-15] and interviews [3, 9, 16-18], there remained a paucity of evidence of the effects on the degree and valence of speech entrainment in other contexts, such as long-and-short turn-taking situations. It is to be expected that entrainment will also be evident in such a context, because successful communication is reflected in mutual accommodation. Therefore, we investigated long-and-short turn-taking in a Mandarin Chinese talk show corpus in which guests' turns were nearly four times longer than the host's. This context is fundamental to extending previous research on the complex links between entrainment and spoken interactions.

Accordingly, this study focused on two specific questions: (1) How does the gender and role interaction participate in speech entrainment in long-and-short turn-taking in terms of three feature sets? (2) What are the degree and valence of speech entrainment in long-and-short turn-taking?

2. Methodology

2.1. Speech corpus

The present study collected 232 audio files from a popular Mandarin Chinese talk show in the Himalaya audio sharing platform [19] -- *Lu Yu's Appointment: Tell Your Story* (2021-2022). We chose this talk show as our corpus for the following three reasons. First, this program has had a powerful impact on the brand over its 20 years of development, which is representative and typical. It has been called "the most valuable Chinese TV show in the past 15 years" by *Time Magazine*. Secondly, it is typical for this talk show that the turn-taking of

10.21437/Interspeech.2023-1154

the guests (6,326 seconds) is almost four times as long as that of the host (1,760 seconds). Thus, long-and-short turn-taking contexts are presented. Third, as one of the most famous hosts in mainland China, Lu Yu is regarded as more likable by most audiences. A friendly atmosphere is created between the host and the guests; thus, speech entrainment is expected in this long-and-short turn-taking.

For the present study, 30 audio files were selected from this corpus based on the following criteria. First, the conversation was a dyadic one between a host and a guest; second, the guest spoke Mandarin Chinese to avoid the influence of dialects; third, at least three but no more than five audio files were required for a guest, so the total audio duration for a single guest was about 26 minutes. Three hours and 32 minutes of conversations with eight guests (4 males and 4 females) were selected.

2.2. Alignment and annotation

The speech corpus was annotated at two levels in *Praat* [20]. For the Chinese character level, the Montreal Forced Aligner [21] with the Mandarin China MFA dictionary and acoustic model aligned the signal automatically by outputting an annotation in TextGrid format. The number of inaccurate alignments was manually adjusted by accounting for changes in waveforms, spectrograms, and perceptual cues if necessary. For the turn level, we annotated a turn as a maximal sequence of inter-pausal units (IPU) of a single speaker [22]. The Chinese IPU was at least 80 ms [2]. The turn identification method followed the judgment of Caspers [22] and Liu [23] that a turn exists when the context in which repairs, backchannels, overlaps, or interruptions are present meets the following three standards. First, the listeners interrupt the speaker's words; second, their roles are switched; third, the listeners' turn provides new information. We also annotated the role of the speaker (host *ER*, guest *EE*) in IPU.

The present corpus contained 722 turns (4,988 IPUs) with a mean of 68.28 Chinese characters and a mean duration of 11.2 seconds per turn. On average, there were 361 turns (930 IPUs) for the female host, and 361 turns (4,058 IPUs) for the guests (206 for females, and 155 for males). The distribution of the duration of all turns in three speaker types (female guest *f_EE*, male guest *m_EE*, and female host *f_ER*) is in Table 1.

Table 1: *Duration of turns for speaker types*

Speaker type	Duration in each turn (seconds)		
	Maximum	Minimum	Mean
f_EE	86.63	0.24	13.88
m_EE	139.43	0.34	22.08
f_ER	37.39	0.28	4.87

2.3. Acoustic-prosodic features extraction

The present study measured 12 acoustic-prosodic features in three feature sets: intensity (*Int*), pitch (*F0*), and duration (i.e., speaking rate *SR*) in each IPU.

Using *Prosody Pro* [24], the time-normalized intensity (dB) was taken from 30 points evenly spaced in each IPU, preserving the original duration. We calculated five features of the maximum (*max_int*), minimum (*min_int*), mean (*mean_int*), median (*med_int*), and standard deviation of the intensity (*std_int*). Besides, the number of Mandarin Chinese characters (equating to syllables σ) was measured in each IPU based on orthography. The speaking rate (s/σ) was then calculated by factoring in the duration (sec) of each syllable in each IPU.

A high accuracy of f0 estimation measured in Hertz was automatically achieved in a two-step process. As a first step, we used an open-source pitch tracker *Reaper* [25], to deal with creaky voices produced frequently by multiple speakers. As a second step, we performed pitch tracking via a two-pass procedure following Hirst [26]. In the first pass, we fixed a precise search range of 75–400 Hz for all IPUs to cover all reasonable *f0* samples and then extracted *f0*. We computed the first and third quartiles (i.e., *q1* and *q3*) across all *f0* samples for each IPU. The second pass gave new values for the pitch floor and pitch ceiling, where the pitch floor was provided by the formula *0.75 * q1*, and the pitch ceiling was given by the formula *1.5 * q3*. We calculated six features of maximum (*max_f0*), minimum (*min_f0*), mean (*mean_f0*), median (*med_f0*), standard deviation of pitch (*std_f0*), and f0 range.

2.4. Feature distances calculation

Weighted averages of IPUs [26] were calculated by the mean feature values of all IPUs in one turn before analysis. Thus, the present study analyzed speech entrainment at the turn level, and a weighted average of all the IPUs in a turn was computed as the representative value of this turn. All entrainment measures were performed on turn pairing [12].

First, we combined four types of turn pairs (Table 2) for guests and the host, respectively, because we applied a directed pairing of turns to the left dialog context only. We examined how similar the second speaker got to the first for each turn pair.

Second, we calculated proximity and synchrony-related distances following Reichel *et al*. [12]. The absolute point-wise distance values of each raw feature within single-turn pairs represented a proximity-related distance. The smaller the distance, the greater the proximity. In addition, we first subtracted the mean values of each speaker from the feature values and then calculated the absolute residual values as synchrony-related distance. If the speakers realized a feature either above or below their respective means, the synchrony-related distance was thus low, and synchrony was higher.

Third, we first checked whether the feature distance values conformed to the normal distribution using the Shapiro-Wilk test in *R* [27]. The result showed that the data were not normally distributed due to different numbers of turn pairing in the two groups, so we chose the Mann-Whitney U test. Thus, local entrainment was measured according to whether the proximity and synchrony-related distance differed significantly in the adjacent and non-adjacent turn pairs were significantly different. Likewise, global entrainment between the same and different conversation turn pairs could be observed.

Finally, the Augmented Dickey-Fuller (ADF) test was performed for the change in mean distance values to assess if it was dynamic over time. Previous studies using t-tests [5, 28, 29] or linear-mixed models [30] compared only the difference between the mean feature values of the interlocutors in the first/ last third of the conversation. However, they did not reflect the changing process. Thus, we applied the ADF test to determine whether a time series is stationary [31, 32]. This approach has been validated as a robust method to construct a new change detector [33, 34]. We used a three-step method to determine if each feature set's changing process was static/dynamic. First, we plotted line charts of mean distance values over 20 sections for each feature set and speaker type. Second, we used the ADF test, in which $p > 0.05$ indicated a dynamic process while $p <= 0.05$ indicated static. Third, we determined dynamic directions by overall line plot upward/downward trend.

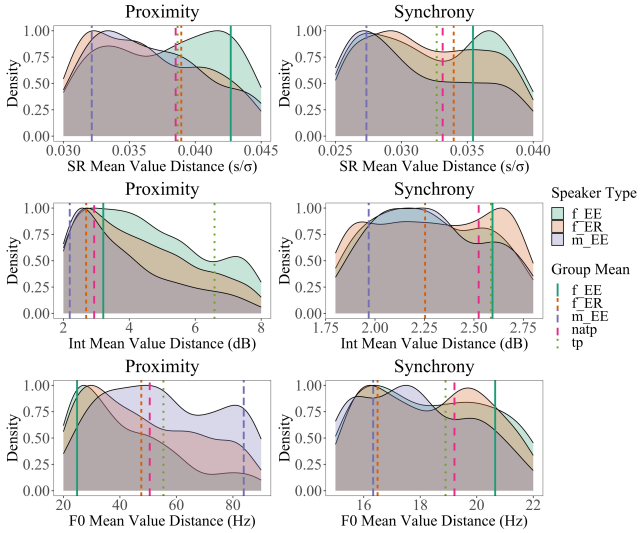

Figure 1: *Proximity and synchrony in feature sets.*

3. Results

3.1. Level and class of entrainment across speaker types

3.1.1. The degree of proximity and synchrony

We compared proximity and synchrony entrainment for each feature set across speaker types at local and global levels. The plot of the smoothed density estimates is shown in Figure 1. The respective mean feature distance values between turn pairs are plotted on the *x*-axis, and the estimated density values are on the *y*-axis, scaled to a maximum of 1. Thus, differences in the degree of entrainment among the three speaker types are shown in colored areas. In addition, mean distance values are divided into three turn pair types: *atp*, *natp*, and *tp*. The *atp* type is further split by the speaker's gender and role to visualize the impact of speaker types on entrainment; the *natp* and *tp* lines serve as the reference line for local and global entrainment, respectively. Each type's mean value is also plotted and represented with vertical lines.

From visual observation, local positive entrainment is indicated by *atp*-lines left of the *natp*-reference line. Global positive entrainment is marked by both *atp*-lines and the *natp*-reference line left of the *tp*-reference line. Likewise, the opposite order indicates a negative entrainment for both the local and global levels. The distance and speaker types always refer to the asymmetric turn pair in which the speaker utters the later turn in one turn pair.

A notable observation was that entrainment behavior differed among the three feature sets. Only male guests showed significant global and local positive proximity and synchrony entrainment for the speaking rate. For the intensity set, male guests and the female host showed the strongest global and local entrainment for proximity and synchrony. For the pitch set, only female guests typically entrained the most for proximity, but not for global and local synchrony. Moreover, of the six figures, only proximity entrainment differed significantly in pitch, whereas female guests showed strong global and local positive proximity entrainment. In addition, all other figures showed that the degree of entrainment of the female host was somewhere between female and male guests.

Table 2: *Description and the number of turn pairs*

Type	Turn pair	Description	ER	EE
adjacent	atp	a turn that directly precedes another turn	337	355
non-adjacent	natp	a turn being randomly selected from the preceding part to pair with another turn	301	301
same conversation	stp	a turn paired with another turn randomly drawn from the preceding part of the same conversation	226	234
different conversation	tp	two turns of a host and guest being in different conversations	361	361

3.1.2. The valence of proximity and synchrony

The conditional probabilities for positive and negative entrainment in each feature across different speaker types are shown in Figure 2. Based on the Whitney U test, if $p < 0.05$ for a speaker type in a feature, we count it as positive; otherwise, we count it as negative. In this way, we obtain the number of global/local positive and negative entrainment for each feature in different speaker types. The conditional probability is the ratio of positive/negative entrainment to the total turn pairs. Therefore, the overall ratio of the four entrainment tendencies is 1, with four types of global/local proximity and synchrony accounting for a certain proportion.

First, global positive synchrony was more likely to occur in the intensity set. In contrast, global negative synchrony and positive proximity were frequently found in intensity and pitch features except for speaking rate. Second, local positive synchrony was more likely to occur in intensity and pitch sets; local positive proximity and negative synchrony were prominent in intensity and pitch features but in speaking rate. The first three features with the highest valence of local positive proximity entrainment were the median, mean, and standard deviation of the pitch. Third, three speaker types presented different valence of entrainment behavior. Male guests entrained the most in global and local positive proximity and synchrony; female guests entrained more in global and local positive proximity; and the female host was more likely to entrain in local positive synchrony. Therefore, the valence of proximity and synchrony entrainment was: i) (-) sync > (+) prox > (-) prox > (+) sync; ii) m_EE > f_ER > f_EE.

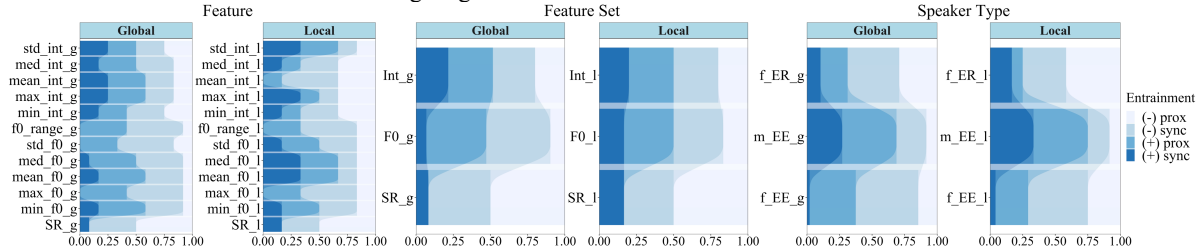

Figure 2: *Conditional global (*_g) and local (*_l) entrainment probabilities for proximity (prox) and synchrony (sync).*

3.2. Dynamicity of entrainment across speaker types

3.2.1. The degree of dynamicity

We calculated mean value feature distances in 20 sections over time for each feature set and speaker type. The ADF test results were recorded in Table 3, with three types of dynamicity, positive *(+) dyn*, negative dynamic *(-) dyn,* and static.

Table 3: The *ADF test of three feature sets*

Feature set	Speaker type	Statistic: DF	*p* value	Dynamicity
SR	f_EE	-5.47	0.08	(-) dyn
	m_EE	-3.04	0.21	(-) dyn
	f_ER	-6.30	0.06	(-) dyn
Int	f_EE	-4.25	0.13	(+) dyn
	m_EE	-4.75	0.11	(+) dyn
	f_ER	-2.96	0.21	(+) dyn
F0	f_EE	-3.78	0.16	(-) dyn
	m_EE	-3.48	0.17	(+) dyn
	f_ER	-4.60	0.11	(+) dyn

Three feature sets were dynamic. On the one hand, different speaker types shared something in common. For instance, negative dynamic entrainment in speaking rate and positive dynamic entrainment in the mean intensity was for all speaker types; On the other hand, different speaker types also differed in some features. Only the female guests showed a negative dynamic entrainment of the pitch. In contrast, male guests and the female host were dynamically entrained more to the interlocutors in the Mandarin Chinese talk show.

3.2.2. The valence of dynamicity

Based on the dynamicity of all features for each speaker type, we counted the numbers of three types in each feature, feature set, and speaker type. We further calculated the conditional probability of three types occupying the ratio of 1 (Figure 3).

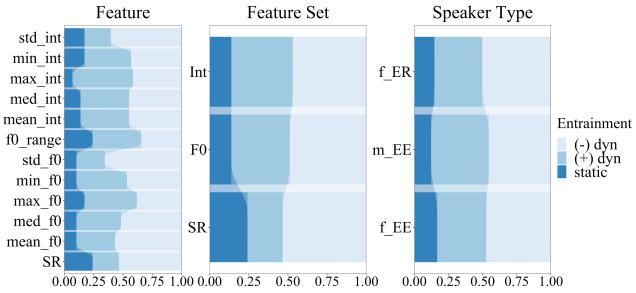

Figure 3: *Conditional probabilities of dynamicity.*

The host and guests entrained most dynamically on pitch range and maximum pitch while dynamically disentrained on the standard deviation of intensity and pitch. The speaking rate exhibited the largest static and negative dynamic proportions, indicating that the host and guests disentrained on the speaking rate. In contrast, intensity and pitch sets presented a similar valence of dynamicity, with dynamic ones for a significant proportion. Male guests showed more positive dynamic entrainment to interlocutors within each speaker type. Female guests and the host displayed slightly more static entrainment than male guests. Therefore, the valence of dynamicity was: i) (+) dyn ≈ (-) dyn > static; ii) f_EE ≈ f_ER > m_EE.

4. Discussion

For feature sets, the intensity remains the most robust and consistent entraining feature in long-and-short turn-taking. The above finding mirrors the results of Beňuš *et al*. [35] and Levitan *et al*. [4], who also found that intensity is the strongest entrainment indicator in the short-and-short turn-taking. Moreover, the host and guests steadily diverge the most on the speaking rate at the local and global levels, which is consistent with Levitan *et al*. [18], and also extends previous evidence for entrainment in speaking rate [2, 5, 11, 13]. There is evidence of dynamic local positive proximity for pitch. This result is broadly consistent with De Looze *et al*. [28]. The host and guests approximate each other's pitch because most of the questions the host asks in this talk show are typically chit-chat, where a final rise in pitch is unusual. The greater the functional importance of pitch in lexical differences in Chinese, the more significant entrainment can be at the local level.

For the valence of entrainment, local positive entrainment is greater than that of global positive entrainment, which shows a total negative local and global entrainment. These results do not support previous research by Reichel *et al*. [12] on short-and-short turn-taking. One possible explanation is that entrainment in long turns decreases over time as guests tell stories, showing less positive entrainment at the global level.

Another important finding is that the female host entrains more to male than to female guests, as evidenced primarily by local positive proximity and synchrony. This suggests that the female host has a lower dominance of conversation for mixed-gender pairs but a higher dominance for same-gender pairs. This result regarding gender-role interaction is quite complex compared to previous studies that have examined the effects of gender or role alone on speech entrainment [3, 6, 7, 12, 36]. The reason for this is unclear but may be related to different gender-role strategies to establish common ground in a talk show. Male guests may accept the interviewee's responsibility to establish common ground. In contrast, female guests believe that it is the host's responsibility to lead the conversation. However, these results are also due to individual differences and corpus size, so further studies with larger corpus are suggested.

5. Conclusions

We focused on distinguishing three speaker types and used proximity and synchrony-related distance calculation and the ADF test to investigate the degree and valence of speech entrainment in long-and-short turn-taking. This study showed that overt entrainment is inextricably linked to gender and role interaction in three acoustic-prosodic features. Future research is needed to expand the corpus size and conduct an experiment providing perceptual evidence. Moreover, having only one host as a case study limits the ability to make broad claims to talk show hosts in general. Using multiple hosts in future work would allow for conclusions that can address hosts and speech entrainment more broadly.

6. Acknowledgements

This work was supported by the major program of the National Social Science Foundation of China [18ZDA293]; grants of the 2022 Graduate Research and Innovation Project of Institute of Corpus Studies and Applications, Shanghai International Studies University. The corresponding author is Hongwei Ding (hwding@sjtu.edu.cn).

7. References

[1] C. J. Wynn and S. A. Borrie, "Classifying conversational entrainment of speech behavior: An expanded framework and review," *Journal of Phonetics,* vol. 94, pp. 1-11, Sep. 2022.

[2] Z. Xia and Q. Ma, *Prosodic Entrainment in Mandarin Chinese Conversations: An Experimental Study*. Shanghai, China: Tongji University Press, 2019.

[3] A. Weise, S. I. Levitan, J. Hirschberg, and R. Levitan, "Individual differences in acoustic-prosodic entrainment in spoken dialogue," *Speech Communication,* vol. 115, pp. 78-87, Dec. 2019.

[4] R. Levitan, Š. Beňuš, A. Gravano, and J. Hirschberg, "Acoustic-prosodic entrainment in Slovak, Spanish, English and Chinese: A cross-linguistic comparison," in *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Prague, Czech Republic, Sep. 2015, pp. 325-334.

[5] R. Levitan and J. Hirschberg, "Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions," in *Proceedings of INTERSPEECH 2011 - 12th Annual Conference of the International Speech Communication Association*, Florence, Italy, Aug. 2011.

[6] J. S. Pardo *et al.*, "A comparison of phonetic convergence in conversational interaction and speech shadowing," *Journal of Phonetics,* vol. 69, pp. 1-11, Jul. 2018.

[7] Z. Xia, R. Levitan, and J. Hirschberg, "Prosodic entrainment in Mandarin Chinese and English: A cross-linguistic comparison," in *the 7th International Conference on Speech Prosody*, Dublin, Ireland, May 2014, pp. 65-69.

[8] Š. Beňuš, A. Gravano, and J. Hirschberg, "Pragmatic aspects of temporal accommodation in turn-taking," *Journal of Pragmatics,* vol. 43, no. 12, pp. 3001-3027, Sep. 2011.

[9] S. W. J. Gregory and S. Webster, "A nonverbal signal in voices of interview partners effectively predicts communication accommodation and social status perceptions," *Journal of Personality and Social Psychology,* vol. 70, no. 6, pp. 1231-1240, Jun. 1996.

[10] J. S. Pardo, "On phonetic convergence during conversational interaction," *The Journal of the Acoustical Society of America,* vol. 119, no. 4, pp. 2382-2393, Jan. 2006.

[11] T. Kendall, "Speech Rate, Pause, and Linguistic Variation: An Examination through the Sociolinguistic Archive and Analysis Project," Dissertation, English, Duke University, 2009.

[12] U. D. Reichel, S. Beňuš, and K. Mády, "Entrainment profiles: Comparison by gender, role, and feature set," *Speech Communication,* vol. 100, no. 1, pp. 46-57, Jun. 2018.

[13] U. Cohen Priva, L. Edelist, and E. Gleason, "Converging to the baseline: Corpus evidence for convergence in speech rate to interlocutor's baseline," *The Journal of the Acoustical Society of America,* vol. 141, no. 5, pp. 2989-2996, May 2017.

[14] Y. Lee, S. Gordon Danner, B. Parrell, S. Lee, L. M. Goldstein, and D. Byrd, "Articulatory, acoustic, and prosodic accommodation in a cooperative maze navigation task," *PLoS ONE,* vol. 13, no. 8, pp. 1-26, Aug. 2018.

[15] P. Šturm, R. Skarnitzl, and T. Nechanský, "Prosodic accommodation in face-to-face and telephone dialogues," in *Proceedings of INTERSPEECH 2021 - 22nd Annual Conference of the International Speech Communication Association*, Brno, Czechia, Sep. 2021, pp. 1444-1448.

[16] R. L. Street, "Speech convergence and speech evaluation in fact-finding interviews," *Human Communication Research,* vol. 11, no. 2, pp. 139-169, Dec. 1984.

[17] E. Weizman, "Roles and identities in news interviews: The Israeli context," *Journal of Pragmatics,* vol. 38, no. 2, pp. 154-179, Feb. 2006.

[18] S. I. Levitan, J. Xiang, and J. Hirschberg, "Acoustic-prosodic and lexical entrainment in deceptive dialogue," in *the 9th International Conference on Speech Prosody*, Poznań, Poland, Jun. 2018.

[19] Ximalaya.com. "Lu Yu's appointment: Tell your story." Available: https://www.ximalaya.com/album/25484870, Sep. 9, 2022. [Accessed: Sep. 10, 2022].

[20] *Praat: Doing Phonetics by Computer*. (2022). [Online]. Available: http://www.praat.org.

[21] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable text-speech alignment using Kaldi," in *Proceedings of INTERSPEECH 2017 - 18th Annual Conference of the International Speech Communication Association*, Stockholm, Aug. 2017, pp. 498-502.

[22] J. Caspers, "Local speech melody as a limiting factor in the turn-taking system in Dutch," *Journal of Phonetics,* vol. 31, no. 2, pp. 251-276, 2003.

[23] H. Liu, *Conversation Analysis: An Introduction*. Beijing: Peking University Press, 2004.

[24] Y. Xu, "ProsodyPro — A tool for large-scale systematic prosody analysis," in *Proceedings of Tools and Resources for the Analysis of Speech Prosody (TRASP 2013)*, Aix-en-Provence, France, Aug. 2013, pp. 7-10.

[25] *MacReaper*. (2022). Brisbane, Australia. [Online]. Available: https://kjdallaston.com/projects.

[26] D. Hirst, "The analysis by synthesis of speech melody: From data to models," *Journal of Speech Sciences,* vol. 1, no. 1, pp. 55-83, Jul. 2011.

[27] *The R Project for Statistical Computing*. (1997). [Online]. Available: https://www.r-project.org.

[28] C. De Looze, S. Scherer, B. Vaughan, and N. Campbell, "Investigating automatic measurements of prosodic accommodation and its dynamics in social interaction," *Speech Communication,* vol. 58, pp. 11-34, Mar. 2014.

[29] E.-S. Ko, A. Seidl, A. Cristia, M. Reimchen, and M. Soderstrom, "Entrainment of prosody in the interaction of mothers with their young children," *Journal of Child Language,* vol. 43, no. 2, pp. 284-309, Jun. 2015.

[30] J. Michalsky and H. Schoormann, "Pitch convergence as an effect of perceived attractiveness and likability," in *Proceedings of INTERSPEECH 2017 - 18th Annual Conference of the International Speech Communication Association*, Stockholm, 2017, pp. 2253-2256.

[31] J. D. Hamilton, *Time Series Analysis*. Princeton, New Jersey: Princeton University Press, 2020.

[32] P. C. Phillips and P. Perron, "Testing for a unit root in time series regression," *Biometrika,* vol. 75, no. 2, pp. 335-346, 1988.

[33] R. P. Silva, B. B. Zarpelão, A. Cano, and S. B. Junior, "Time series segmentation based on stationarity analysis to improve new samples prediction," *Sensors,* vol. 21, no. 21, p. 7333, 2021.

[34] I. E. Livieris, S. Stavroyiannis, L. Iliadis, and P. Pintelas, "Smoothing and stationarity enforcement framework for deep learning time-series forecasting," *Neural Computing and Applications,* vol. 33, no. 20, pp. 14021-14035, 2021.

[35] S. Beňuš, R. Levitan, J. Hirschberg, A. Gravano, and S. Darjaa, "Entrainment in Slovak collaborative dialogues," in *the 5th IEEE Conference on Cognitive Infocommunications (CogInfoCom)*, Nov. 2014, pp. 309-313.

[36] R. Levitan, A. Gravano, L. Willson, S. Beňuš, J. Hirschberg, and A. Nenkova, "Acoustic-prosodic entrainment and social behavior," in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Montréal, Canada, Jun. 2012, pp. 11-19.