Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-024-04137-4 OPEN

Unpacking the gender-role interaction of prosodic entrainment in Chinese long-and-short turn-taking: evidence from perceptual and acoustic similarities

Yanting Sun [1] & Hongwei Ding[1,2,3]✉

Prosodic entrainment, the phenomenon of speakers adapting their speech patterns to each other, has been extensively studied in short turn-taking conversations. However, little is known about its dynamics in long-and-short turn-taking contexts, such as talk shows. This study investigates prosodic entrainment in Mandarin Chinese talk shows, focusing on the effects of gender and role on entrainment behaviour and the correlation between perceived similarity and acoustic-prosodic features. We analysed a corpus of 50 audio files from a popular Mandarin Chinese talk show, examining three acoustic-prosodic features: pitch, intensity, and speaking rate. Prosodic entrainment was measured using proximity and synchrony-related distances at both local and global levels. Additionally, a perceptual judgement task was conducted to assess perceived prosodic similarity. Results showed that negative entrainment was more prevalent than positive entrainment across all three feature sets, particularly in speaking rate synchrony. The female host demonstrated larger positive entrainment compared to both male and female guests. Gender-role interactions revealed complex entrainment patterns: mixed-gender pairs showed greater local synchrony, while same-gender pairs exhibited more global entrainment. Perceived similarity correlated moderately with pitch for male guests in both local proximity and synchrony, suggesting pitch plays a crucial role in listeners' perception of prosodic similarity. These findings challenge conventional assumptions about positive entrainment's benefits in communication and highlight the complex interplay between gender, role, and entrainment behaviour in talk show settings. The prevalence of negative entrainment suggests that contrasting speech patterns can contribute to engaging and dynamic conversations in specific contexts. This study enhances our understanding of prosodic entrainment in long-and-short turn-taking contexts and emphasises the need to consider gender, role, and communicative settings in future research on speech accommodation.

[1] Institute of Language Sciences, Shanghai International Studies University, Wenxiang Road 1550, Songjiang District, 201306 Shanghai, China. [2] Speech-Language-Hearing Center, School of Foreign Languages, Shanghai Jiao Tong University, Dongchuan Road 800, Minhang District, 200240 Shanghai, China. [3] National Research Centre for Language and Well-Being, Shanghai, China. ✉email: hwding@sjtu.edu.cn

Introduction

Speech entrainment refers to the phenomenon wherein interlocutors exhibit similar speech behaviour[1] (Wynn and Borrie, 2022). This process of communication involves cooperation and adaptation between the parties involved, leading to the creation of a cohesive and smooth dialogue. According to Mondada (2001) and De Looze et al. (2014), dialogue is a social interaction requiring continual joint adjustments. Prosodic entrainment, a specific aspect of speech entrainment, is defined as the alignment of prosodic features—such as pitch, intensity, and duration—between speakers (Levitan et al., 2012; Xia et al., 2014). This alignment is influenced by a range of cognitive, psychological, linguistic, social, and contextual factors. Given its seeming complexity, the question arises as to whether speech convergence is underpinned by automatic or socially motivated mechanisms. There have been several models positing entrainment as an automatic process. For instance, the Interactive Alignment Model suggests that this process is achieved through a "primitive and resource-free priming mechanism" (Pickering and Garrod, 2004, 172), enabling conversational partners to share linguistic representations and thereby reducing cognitive load. In contrast, Communication Accommodation Theory (CAT) describes potential socially motivated mechanisms, viewing convergence as a non-automatic behaviour aimed at decreasing social distance or gaining approval. This theory predicts that when a power imbalance exists between interlocutors, the less dominant speaker will converge more, while a talker may choose to diverge from an interlocutor to increase the social distance between them (Bourhis and Giles, 1977) or to show a desire to gain the interlocutor's approval (Giles et al., 1991; Giles and Ogay, 2007).

Despite the unclear mechanisms, strong evidence suggests that entrainment is productive and useful. Interlocutors adjust speech behaviour to create smooth dialogue, which can be measured through continuous acoustic-prosodic features like speaking rate (Cohen Priva et al., 2017; Wynn et al., 2022), intensity (Borrie et al., 2015; Rahimi et al., 2017), pitch (Babel and Bulatov, 2012; Bradshaw and McGettigan, 2021; Loveday, 1981), voice quality (e.g., jitter, shimmer, noise-to-harmonics ratio; Levitan et al., 2018). Studies have focused on duration, intensity, and pitch as fundamental components of prosody (Levitan et al., 2015; Sun and Ding, 2023; Weise et al., 2019; Xia and Ma, 2019), essential for smooth communication (Levitan and Hirschberg, 2011). Given the established body of research on these three features and their potential significance in conversational dynamics, the present study focuses on these features to examine prosodic entrainment.

Research has also investigated perceptual similarity tasks to measure entrainment, juxtaposing subjective perceptual approaches with objective acoustic measures. The human perceptual system can also be a holistic assessor by conducting perceptual similarity tasks. Perceptual similarity judgements have been the gold standard of prosodic entrainment research for nearly two decades. Perceptual similarity judgements, such as *AXB* tests (Babel et al., 2014; Lewandowski and Nygaard, 2018; Pardo, 2006; Pardo et al., 2018) or *Likert* scales (Abel and Babel, 2017), assess multiple dimensions important to listeners. Pardo et al. (2017) found perceptual similarity tasks detected significant entrainment effects, with each acoustic feature predicting similarity scores. In our study, we follow Abel and Babel (2017)'s research method because the *Likert* scale can provide the perceptual similarity score for each dyad, which is beneficial for observing the listeners' rating score for each dyad.

Although previous studies have relied on the entrainment categorisation system introduced by Levitan and Hirschberg (2011), recent frameworks by Wynn and Borrie (2022) have advanced this field by classifying entrainment into eight distinct types based on class, level, and dynamicity. The first classification factor *class* differentiates proximity from synchrony. Proximity refers to the overall similarity of speech features between interlocutors across an entire interaction. For example, two speakers might have similar average pitch ranges throughout a conversation. Synchrony, however, focuses on the dynamic, moment-to-moment coordination of speech features. It measures how speakers adjust their prosodic features in real-time in response to each other, including both the direction and magnitude of change. For instance, synchrony would capture how one speaker might increase their speech rate immediately after their partner does. The second classification factor *level* specifies local or global entrainment. Local entrainment occurs between units equal to or smaller than adjacent turns, while global entrainment occurs across any time scale greater than adjacent turns. The third classification factor *dynamicity* considers changes over time, categorising into static and dynamic types. Dynamic entrainment is further classified into dynamic proximity (i.e., a change in the similarity of speech features over time) and dynamic synchrony (i.e., a change in the similarity of movement of speech features over time), depending on the type of change observed. Positive or negative changes in dynamic entrainment can also occur. The direction of the change may vary—similarity can increase or decrease over time, respectively referred to as positive or negative dynamic entrainment. By operationalising these classification factors, Wynn and Borrie's framework facilitates a clearer and more precise categorisation of entrainment. In this study, therefore, we use Wynn and Borrie's framework for examining prosodic entrainment. The above classification can examine the valence of entrainment in spoken interactions, specifically positive or negative valence. Positive prosodic entrainment demonstrates accommodative behaviour, where interlocutors adjust their speech patterns to become more similar, while negative prosodic entrainment indicates a lack of entrainment tendency. The degree and valence of entrainment have varied correlations concerning different conversational contexts. The degree and valence of entrainment vary across conversational contexts, influenced by social factors such as gender and the interlocutor's role. In this context, the role of the interlocutor refers to the social or interactional position an individual occupies during the conversation. This role, such as leader, follower, expert, or novice, can affect how participants adapt their prosodic features. Subordinate interlocutors, such as students or employees, may exhibit higher levels of positive entrainment when interacting with dominant figures like teachers or managers. Additionally, gender dynamics can further modulate this effect. Research has suggested that women may exhibit more accommodative behaviour through entrainment, particularly in mixed-gender interactions where they may adopt more affiliative roles (Reichel et al., 2018). Conversely, men may demonstrate lower degrees of entrainment when occupying dominant roles, potentially as a reflection of different social expectations (Weise et al., 2019).

Prosodic entrainment is evident in contexts of short-and-short turn-taking, where both interlocutors typically produce brief utterances in rapid succession. In these contexts, turns are generally limited to a few seconds each, allowing for frequent exchanges between speakers. This pattern of interaction is characteristic of task-/topic-oriented conversations (Cohen Priva et al., 2017; Guydish and Fox Tree, 2022; Lee et al., 2018; Levitan and Hirschberg, 2011; Reichel et al., 2018; Šturm et al., 2021; Ulbrich, 2021; Weise et al., 2019), therapy (Lee et al., 2010; Nasir et al., 2018), tutoring (Ward and Litman, 2007), judicial hearings (Danescu-Niculescu-Mizil et al., 2012) and interviews (Gregory and Webster, 1996; Levitan et al., 2018; Street, 1984; Sun and Ding, 2023; Weise et al., 2019; Weizman, 2006). Several studies,

including Matarazzo and Wiens (1967) and Natale (1975), highlighted evidence that interviewees tended to make prosodic entrainment to the interviewer regarding the pause duration and mean vocal intensity entrainment, respectively. Global mean pitch prosodic entrainment was observed by Collins (1998) in English interviews. Levitan et al. (2018) also conducted a study on deceptive interviews and found that differences in entrainment behaviour between deceptive and truthful speech could be identified in acoustic-prosodic and lexical dimensions. Recently, Weise et al. (2019) investigated two forms of local acoustic-prosodic entrainment between English and Chinese native speakers in English conversations, analysing individual differences in large corpora. They found no variable entrainment based on factors such as gender, native language, or their combination. However, they hypothesised that gender and its complex role might be significant among other factors, including sociocultural norms and conversational contexts.

Many of the above studies have only explored the impact of each factor separately (Gregory and Webster, 1996; Guydish et al., 2021; Guydish and Fox Tree, 2022; Levitan et al., 2015; Levitan and Hirschberg, 2011; Pardo et al., 2018; Sun and Ding, 2023; Weise et al., 2019; Xia et al., 2014), failing to account for potential inconsistencies observed in the results. Therefore, it is vital to investigate the effects of both gender and role on prosodic entrainment to further understand this phenomenon. Notably, Pardo (2006) discovered that males in a dependent role displayed more prosodic entrainment than those in positions of power in task-based conversations. The findings of Kendall (2009) indicated that a person's speaking rate was more influenced by the interviewer's gender rather than their own. Additionally, cooperative interactions have been shown to be impacted by gender, with female describers and male followers entraining the most, while male describers and female followers entraining the least (Reichel et al., 2018).

While previous studies have attempted to explore prosodic entrainment in short-and-short turn-taking, little research has focused on long-and-short turn-taking contexts like talk shows (e.g., Sun and Ding, 2023), where one speaker consistently produces longer utterances and the other typically responds with shorter turns. This pattern is common in talk shows, where guests often provide extended responses to brief questions or prompts from the host. It is expected that prosodic entrainment will occur in these types of contexts, as successful communication generally requires mutual adaptation. However, the dynamics of entrainment in long-and-short turn-taking may differ from those in short-and-short interactions. The asymmetry in turn length could potentially affect the opportunities for and patterns of prosodic adaptation. For instance, the speaker with longer turns might have more time to modulate their prosody, while the speaker with shorter turns might need to adapt more quickly or selectively. To investigate these potential differences, we examined long-and-short turn-taking in a Mandarin Chinese talk show corpus in which the length of the guests' turns was almost five times that of the host.

Previous research has been limited by small corpus samples and a lack of perceptual experiments (Levitan and Hirschberg, 2011; Sun and Ding, 2023; Weise et al., 2019). To address these gaps, our current study expands the corpus sample size and includes perceptual experiments to provide more evidence. Accordingly, the present study addressed three specific questions: (1) What is the degree and valence of prosodic entrainment in long-and-short turn-taking? (2) Does the perceptual similarity judgement task correlate with acoustic-prosodic features, and to what extent does it validate the results from the acoustic-prosodic analysis? (3) To what extent do gender and role influence prosodic entrainment in long-and-short turn-taking, and how do they affect the three vital acoustic-prosodic features?

Methods

Acoustic similarity analysis

Corpus collection. The study utilised a corpus derived from the Mandarin Chinese talk show *Luyu's Appointment: Tell Your Story* (2021–2022), aired on the open-source Himalaya platform (https://www.ximalaya.com/album/25484870). The authors compiled 232 audio files, downloading them as WAV files with corresponding transcripts saved as TXT files. The show, which features people with extraordinary life stories, was selected for its significant market impact over 20 years, earning recognition from *Time Magazine* as "the most valuable Chinese TV show of the past 15 years". The format typically involves guests speaking for longer durations than the host, with an average turn duration of 19.09 s for guests compared to 4.72 s for the host, making it ideal for examining long-and-short turn-taking interactions. The female host, *Luyu CHEN*, often referred to as the "Chinese Oprah", is known for her approachable demeanour and ability to foster genuine, in-depth conversations, creating an environment conducive to observing natural speech patterns and potential entrainment.

This study was comprised of 50 individual audio files selected based on three criteria. First, the conversation was dyadic, involving only the host and a guest; Second, to ensure linguistic homogeneity, only conversations conducted entirely in Standard Mandarin Chinese were chosen, excluding any instances of code-switching to other Chinese varieties or the use of strongly accented regional Mandarin. This selection was to minimise the influence of dialectal variation on the analysis. Lastly, each guest contributed between three to five audio files, resulting in approximately 22 min of audio per guest. The final dataset consisted of 14 guests (7 males and 7 females), with a total of 5 h and 41 min of recorded conversations and 82,371 words. Further details on data pre-processing are in sections 'Alignment and annotation' and 'Materials'. All corpus data and scripts are available in the OSF repository at https://osf.io/ybx86/.

Alignment and annotation. We annotated the speech corpus at two levels using *Praat* (Boersma and Weenink, 2024). At the Chinese character level, we employed the *Montreal Forced Aligner* (McAuliffe et al., 2017) with the Mandarin China MFA dictionary and acoustic model to automatically align the signal, producing a *TextGrid* annotation. The present research excluded laughing, coughing, sneezing and so on, which contained no linguistic contents and were not annotated in *Praat*. The filled pauses, repairing, restarting, backchannel and so on, which contained linguistic contents, were included, and they were considered valid speaking and were annotated in *Praat*. The number of inaccurate alignments of Chinese characters was manually adjusted by accounting for changes in waveforms, spectrograms, and perceptual cues if necessary.

At the turn level, our study defined a turn as a maximal sequence of inter-pausal units (IPU) produced by a single speaker, as outlined by Caspers (2003). An IPU is a unit used for analysing prosodic entrainment, defined as the stretch of speech produced by a single speaker that is bounded by pauses (Levitan and Hirschberg, 2011). In our study, we set the threshold for pauses between Chinese IPUs to be at least 80 ms, as suggested by Xia and Ma (2019), and annotated pause boundaries automatically with a *Praat* script[2]. The automatic annotation was then manually refined, ensuring IPU boundaries did not split individual characters and verifying the 80 ms pause threshold was correctly applied. Speaker turns were determined based on Caspers (2003) and Liu (2004)'s criteria. A turn was identified when repairs, backchannels, overlaps, or interruptions met three standards: (1) the listener interrupts the speaker's words (e.g., Host: "*I think the economy is-*"; Guest: "*Actually, recent data*

Table 1 Four types of turn pairs based on the roles of the host (*ER*) and guest (*EE*), with corresponding counts for each type.

Type	Turn pair	Description	Number of turn pairs	
			ER	***EE***
Adjacent	*atp*	a turn that directly precedes another turn.	511	551
Non-adjacent	*natp*	a turn being randomly selected from the preceding part to pair with another turn.	456	456
Same dyad	*stp*	a turn paired with another turn randomly drawn from the preceding part of the same dyad.	337	365
Different dyads	*tp*	two turns of a host and guest being in different dyads.	556	556

The distribution of turn pairs across different types is not necessarily balanced, as the pairing of turns is based on the principle of maximising the formation of turn pairs rather than ensuring an equal number representation of the two roles.

shows…"); (2) their roles switch (e.g., Host: "*What's your view on this issue?*" Guest: "*Well, in my experience…*"); and (3) the listener's turn provides new information (e.g., Host: "*Tell us about your new project.*" Guest: "*Certainly, our latest initiative focuses on…*"). Furthermore, we manually annotated the role of the speaker (i.e., host/interviewer *ER* or guest/interviewee *EE*) in each IPU again, and all annotations were double-checked by the authors to ensure accuracy.

The present corpus contained 1112 turns (8151 IPUs), with an average of 71.89 Chinese characters (excluding punctuation) and a mean duration of 11.91 s per turn. On average, there were 556 turns (1383 IPUs) for the female host and 556 turns (6768 IPUs) for the guests (2897 for females and 3871 for males). These IPUs were extracted using a *Praat* script[3], which split the long sound files into shorter segments.

Acoustic-prosodic feature extraction. The present study measured twelve acoustic-prosodic features across three feature sets: intensity, pitch, and duration (specifically speaking rate). These features were extracted from each IPU.

To collect intensity (i.e., *Int*) data, we used the *Prosody Pro* script (Xu, 2013). Time-normalised intensity, measured on a scale of 0–100 dB, was recorded at 30 points, equidistantly spaced across each IPU while preserving the original duration of the recording. From these recordings, we calculated five measures: maximum (*max_int*), minimum (*min_int*), mean (*mean_int*), median (*med_int*), and standard deviation (*sd_int*) of the intensity. Additionally, the number of Mandarin Chinese characters, corresponding to syllables (σ) in each IPU, was based on orthographic transcriptions. We calculated speaking rate (*SR*, in ms/σ) by dividing the duration of each syllable by the number of syllables in the IPU.

We employed a two-step process for pitch (f0) estimation, measured in Hertz, to achieve higher accuracy, especially in creaky voices. First, we used a wrapper for the open-source pitch tracker, *Reaper* (Dallaston, 2023), to handle the irregular vocal fold vibrations typical of creaky voices, which can lead to errors in traditional f0 estimation methods (Keating et al., 2015). *Reaper*'s EpochTracker class estimated glottal closure instants (GCI), voicing state, and local f0 by inverting the time between successive GCIs, effectively tracking pitch even in irregular signals (Talkin, 2015). The features of the *Reaper* f0 tracking algorithm allowed for more accurate f0 measures when there is creaky phonation. Second, we used a two-pass pitch tracking procedure (Hirst, 2011). In the first pass, we set fixed pitch ranges of 100–400 Hz for females and 75–300 Hz for males to capture all reasonable f0 samples. To address potential f0 aliasing and the limitations of using a full 100% f0 range, a refined approach was adopted by trimming the data to 90%, 91%, and ultimately 98% ranges. This strategy mitigated the effects of extreme f0 values that could result from mistracking or creaky voice, ensuring a more accurate representation of the f0 range. Pitch analysis settings were validated by repeating the analyses with varying parameters, confirming that f0 aliasing did not occur and establishing the robustness of the chosen settings. After exploring multiple trim ranges, 98% proved optimal, yielding the smallest average standard deviation in the f0 mean values and providing a reliable summary (results of the range exploration are available at https://osf.io/ybx86/). We then computed the first and third quartiles (i.e., q1 and q3) across all f0 samples for each IPU. This approach was adopted as errors were more likely to occur at the extreme distribution values, and the first and third quartiles of distribution were more robust estimates of the dispersion. The refinement of the pitch floor and ceiling effectively excluded many outliers that were likely due to measurement errors rather than genuine pitch extremes. This automated approach resulted in the apparent "disappearance" of long tails in the f0 histograms, as the extreme values were no longer included in the final f0 calculations. In the second pass, new values for the pitch floor and pitch ceiling were obtained. 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 pitch features (i.e., *f0*), namely the maximum (*max_f0*), minimum (*min_f0*), mean (*mean_f0*), median (*med_f0*), standard deviation (*sd_f0*) of pitch, and f0 range (*f0_range*), based on these new values.

Proximity/synchrony-related distance calculation. To measure entrainment, we focused on the relative feature distance from a reference value, where positive entrainment shows lower feature distances and negative entrainment shows higher feature distances (Reichel et al., 2018). We measured entrainment in turn pairs, with each turn containing multiple IPUs. For each turn, we computed a weighted average of all IPUs (Xia and Ma, 2019). Specifically, we combined four types of turn pairs (Table 1) for both guests and the host, using directed pairing of turns to the left dialogue context only. We then assessed the similarity of the second speaker to the first speaker in each turn pair. Local entrainment was measured by comparing adjacent and non-adjacent turn pairs within the same dyad, while global entrainment was measured by comparing turn pairs within the same and different dyads. For example, we compared an adjacent turn pair (a host's turn followed by a guest's turn) with a non-adjacent turn pair (a guest's turn located earlier in the conversation). Positive local entrainment was observed if adjacent turn pairs were more similar than non-adjacent ones.

To calculate proximity/synchrony-related distances, we followed the method proposed by (Reichel et al., 2018). We assigned point-wise distance values (positive or negative values) to each raw feature within single-turn pairs to represent *proximity-related distance*; smaller values indicated greater proximity. We then subtracted the mean values of each speaker from the feature values and calculated the absolute residuals as *synchrony-related*

distance. Low synchrony-related distances indicated high synchrony, reflecting parallel movements of speakers' features around their respective means. We used the Shapiro-Wilk test in *R* (R Core Team, 2024) to check if feature distance values followed a normal distribution. As the data were non-normally distributed, we used the Mann–Whitney *U*-test to compare proximity and synchrony-related distances in adjacent and non-adjacent turn pairs for local entrainment and between the same and different dyad turn pairs for global entrainment. We evaluated *p*-values for local and global proximity and synchrony across three feature sets for each speaker type. A p-value > 0.05 indicated negative entrainment, while $p < 0.05$ suggested significant differences, necessitating further analysis to determine if entrainment was positive or negative. This judgement involved examining the relative position of the vertical line in the density plot (see section 'Level and class of entrainment across speaker types') of the feature set.

To understand the valence of entrainment through acoustic-prosodic features, we examined the conditional probabilities of positive and negative entrainment across speaker types. We calculated *conditional probabilities* by dividing the number of positive or negative entrainment occurrences by the total entrainment occurrences in proximity or synchrony, locally or globally. Specifically, the *condition* here referred to the type of entrainment (positive or negative) and its scope (local or global, and proximity or synchrony). This analysis included all turn pairs with observed entrainment, excluding those without entrainment, resulting in an overall ratio of four types of global/local positive/negative proximity and synchrony entrainment.

Dynamicity measures. To assess the dynamic changes over time, we performed the Augmented Dickey-Fuller (ADF) test on the mean absolute distance values. Unlike previous studies that utilised *t*-tests (De Looze et al., 2014; Ko et al., 2015; Levitan and Hirschberg, 2011) or linear mixed-effect models (Michalsky and Schoormann, 2017) to compare mean feature values between the first and last thirds of conversations, we employed the ADF test from the *fUnitRoots* package in *R* to determine the stationarity of a time series (Hamilton, 2020; Phillips and Perron, 1988). This method is validated as a robust approach for constructing new change (Livieris et al., 2021; Silva et al., 2021).

We applied a three-step method to assess whether the feature set changes by speaker type were static or dynamic. First, we plotted line charts showing the changes in mean distance values across ten sections[4] for each feature set and speaker type. Second, we used the ADF test to statistically examine the trends in these line plots, categorising them as static or dynamic based on the obtained *p*-values. A *p*-value greater than 0.05 indicated a dynamic process, while a *p*-value less than 0.05 indicated a static process. Third, we determined whether the dynamic process was positive or negative by analysing the overall trend in the line plots. An upward trend indicated negative dynamic entrainment, whereas a downward trend indicated positive dynamic entrainment.

To further understand the dynamicity of each feature set for all speaker types, we also calculated the conditional probabilities of each feature type within all 84 dyads. This analysis involved tallying the occurrences of each feature type and calculating the ratio of each type's occurrence to the total, providing insight into the prevalence of dynamic changes across different feature sets and speaker types.

Perceptual judgement experiment

Participants. Forty-one Mandarin Chinese undergraduate students (mean age = 24.19 ± 3.12 years) were recruited from Shanghai International Studies University. All participants self-identified as native speakers of Mandarin Chinese and hailed from North China, specifically from Beijing, Tianjin, Hebei, and Shandong provinces. Participants reported speaking Standard Mandarin as their primary language of communication both at home and in academic settings. These participants reported no speech, language, or hearing impairments and were remunerated for their participation. All participants successfully completed the tasks, with no drop-outs or exclusions from the final data analyses.

Materials. In each audio file, we annotated adjacent turn pairs in two *TextGrid* tiers (*EE* for the guest/interviewee and *ER* for the host/interviewer) in *Praat* (Boersma and Weenink, 2024). Overlapping speech and turns containing only non-linguistic information (e.g., laughter) were also excluded, resulting in 956 clear turn pairs. These turn pairs were extracted using the same *Praat* script[3], which split the long sound files into shorter segments. Following research practices focusing on the first and third portions of conversations as indicators of prosodic entrainment (Abel and Babel, 2017; Kim et al., 2011; Pardo, 2006), 476 turn pairs were identified across all dyads. To ensure a balanced distribution of turn pairs within each dyad, we randomly selected four turn pairs from each dyad (two from the first portion and two from the third portion for both the host and guest roles). This process yielded a final selection of 152 turn pairs (2 turn pairs × 2 roles × 38 dyads). The distribution included 36 turn pairs for female guests, 40 for male guests, and 76 for the female host. Analysis revealed differences in turn length: guests produced a total of 63,219 words, while the host produced 16,707 words, highlighting the "long-and-short" turn-taking dynamic characteristic of talk show interactions.

Procedures. The experiment was conducted in a sound-attenuated booth to minimise noise and distractions, using *PsychoPy* software (Peirce et al., 2022). Following the turn pair selection procedure from section 'Alignment and annotation', a total of 152 trials were presented in a random order within blocks. The presentation of the blocks and trials was randomised for each participant, with the trial presentation being randomised within each block. To maintain ecological validity, the full speech signal was presented, as prosody is typically perceived alongside other speech features in natural conversations. Participants were instructed to evaluate the similarity of the prosody in each trial on a seven-point *Likert* scale, where 1 represented "not similar at all" and 7 represented "extremely similar". They were specifically directed to focus on prosodic aspects, such as speaking rate and rhythm, while attempting to disregard lexical content as much as possible. The task required evaluating whether the second speaker's prosody matched the first speaker's, considering global aspects like speaking rate and loudness, to capture a holistic judgement of prosodic similarity (Abel and Babel, 2017; Pardo et al., 2013).

Prior to the formal task, participants completed a practice block with five items from the initial construction task piloting. This practice familiarised them with the task and rating criteria, emphasising prosodic features. Participants were given five audio examples with reference scores to aid scoring accuracy. If their score deviated significantly, further explanations clarified the reference criteria. For instance, a score of 1 indicated markedly different prosody (e.g., one speaker rushing, the other speaking measuredly), while a score of 4 suggested moderate similarity (e.g., similar rhythm but differing loudness). A score of 7 denoted near-identical prosody, with speakers closely mirroring each other's intonation and pacing. Once participants reached a general agreement on the criteria, they proceeded to the formal

task. Responses were recorded using a slider bar on a mouse, allowing continuous scores from 1 to 7. They heard the stimuli through Logitech ZONE VIBE 125 Studio headphones; the response prompt timed out after 3000 ms, at which time the next trial began. Participants were permitted to listen to each trial once and were provided brief self-timed breaks between blocks. The entire procedure lasted approximately 40 min.

Statistical analysis. To analyse the perceptual judgement results, we first calculated mean similarity ratings for each dyad in each conversation third for each speaker type (i.e., female guest *f_EE*; male guest *m_EE*; and female host *f_ER*, where *EE* stands for guest/interviewee and *ER* for host/interviewer). We then computed a similarity difference value measure by averaging and subtracting the similarity rating scores for the first and third conversation thirds from each dyad, following the method described in previous studies (Babel, 2010, 2012; Pardo et al., 2012). Specifically, we subtracted the mean similarity rating score of the first portion from that of the third portion. The resulting value indicated whether linguistic similarity increased or decreased over time. A negative value implied increasing linguistic similarity, while a positive value suggested linguistic divergence. We computed the similarity difference measure for each speaker type and then aggregated these values by listeners to obtain the average measure per listener.

We also conducted one-sample *t*-tests for each speaker type in two conversation thirds and applied a Bonferroni correction to account for multiple comparisons, resulting in an adjusted alpha level of 0.008. We then employed a linear mixed-effects model to examine the effects of speaker type on the similarity difference value. In this model, similarity difference value served as the response variable, with speaker type as the fixed effect predictor. Participant was included as a random intercept to account for individual variability. The analysis was conducted using the *lmer()* function from the *lme4* package (Bates et al., 2015) in *R*-studio (R Core Team, 2024). Following the establishment of the main effect of speaker type, we performed post-hoc comparisons with the *emmeans* package to identify significant differences among the three speaker type levels to observe the interaction of gender and role.

Correlation analysis. To analyse the correlations between the perceived similarity degree and the twelve acoustic-prosodic features used to determine local proximity and synchrony entrainment for each speaker type, we also employed linear mixed-effect models using the *lmer()* function in the *lme4* package in *R* (Bates et al., 2015). The dependent variable in these models was the calculated distances in each acoustic-prosodic feature. The fixed effects included perceived similarity score and speaker type. We examined significant interactions ($p < 0.05$) of the fixed effects using the *anova()* function from the *car* package in *R* (Fox and Weisberg, 2011). Corresponding subsets were then tested based on these interactions. To account for the many tests being carried out, we corrected our *p*-values for the false discovery rate (Benjamini and Yekutieli, 2001). Additionally, we employed Pearson correlation coefficients to quantify the strength and direction of the relationships between perceived similarity and acoustic-prosodic features.

Results

Acoustic-prosodic analysis results

Level and class of entrainment across speaker types. For proximity (Fig. 1a), we found that both female and male guests exhibited strong local and global proximity in speaking rate, as indicated by the *f_EE* and *m_EE* (*atp*)-lines being positioned to the left of the *natp*-reference and *tp*-reference lines. Conversely, the female host did not exhibit proximity in speaking rate. In terms of intensity, female guests and the female host showed minimal proximity, with their *atp*-lines slightly left of the *natp*-reference line but right of the *stp*-reference line, whereas male guests showed greater disentrainment on intensity. For pitch, the female host demonstrated strong local and global proximity, while male guests tended towards divergent proximity. Additionally, the vertical line of the mean distance value for female guests was positioned around three reference lines, indicating that there was no obvious proximity entrainment found. In coloured density areas, for the speaking rate, the degree of global and local proximity was as follows: (1) *f_EE* ($M = -23.23$, $SD = 52.96$) > *m_EE* ($M = -17.56$, $SD = 42.96$) > *f_ER* ($M = 17.64$, $SD = 50.54$). For the intensity, the degree was different: (2) *f_ER* ($M = -0.48$, $SD = 4.09$) > *f_EE* ($M = -0.44$, $SD = 3.76$) > *m_EE* ($M = 1.50$, $SD = 4.03$). Finally, the density plot in the pitch set for three speaker types was distinct, with the order of degree of proximity as follows: (3) *f_ER* ($M = -34.15$, $SD = 56.01$) > *f_EE* ($M = 3.15$, $SD = 30.74$) > *m_EE* ($M = 69.96$, $SD = 57.92$).

In terms of synchrony (Fig. 1b), male guests showed a slight degree of global and local synchrony in speaking rate, while female hosts and guests did not, as their *atp*-lines were on the right of the reference lines. In intensity, all speaker types exhibited synchrony, with the *m_EE* (*atp*)-line being closest to the reference lines. For pitch, male guests displayed strong synchrony, while female guests and the female host did not. With regards to the coloured areas of the speaking rate, the order of degree of global and local synchrony for the three speaker types was: (1) *m_EE* ($M = 29.35$, $SD = 29.81$) > *f_ER* ($M = 32.73$, $SD = 35.09$) > *f_EE* ($M = 33.78$, $SD = 38.26$)[5]. Additionally, the order of degree of synchrony in the coloured area of the intensity was reversed: (2) *f_EE* ($M = 2.18$, $SD = 2.11$) > *f_ER* ($M = 2.29$, $SD = 2.11$) > *m_EE* ($M = 2.34$, $SD = 1.81$). The pitch set demonstrated the order of degree of synchrony: (3) *f_EE* ($M = 19.70$, $SD = 16.31$) > *f_ER* ($M = 19.60$, $SD = 18.17$) > *m_EE* ($M = 17.19$, $SD = 18.28$).

Our analysis using the Mann–Whitney *U*-test revealed that the female host showed statistically significant local positive proximity in pitch ($p < 0.001$). Extending this analysis, we observed notable discrepancies in proximity and synchrony among different speaker types at both local and global levels (Table 2). The pitch set showed negative values for female guests but mixed results of positive and negative values for the female host and male guests. In intensity, all speaker types showed negative proximity, but female guests displayed some synchrony. For speaking rate, guests demonstrated positive proximity but negative synchrony, suggesting a strategy to convey their narratives.

An examination of entrainment valence across three speaker types revealed nuanced patterns at both local and global levels (Fig. 2). Positive synchrony was predominant among female hosts and female guests globally, contrasting with male guests who showed higher synchrony locally. Positive proximity, while less frequent than synchrony overall, was notably present among female guests at the local level. Additionally, the same role of different gender pairs exhibited distinct entrainment behaviour. Male guests exhibited greater local synchrony, whereas female guests demonstrated higher global synchrony. Similarly, within same-gender pairs of different roles, the female host exhibited the most positive proximity and synchrony across both local and global levels.

Dynamicity of entrainment across speaker types. We analysed the dynamicity of entrainment across different speaker types for each feature set (Fig. 3) and identified the valence of static, positive dynamic, and negative dynamic entrainment using conditional probabilities (Fig. 4). The line chart revealed that (1) in the

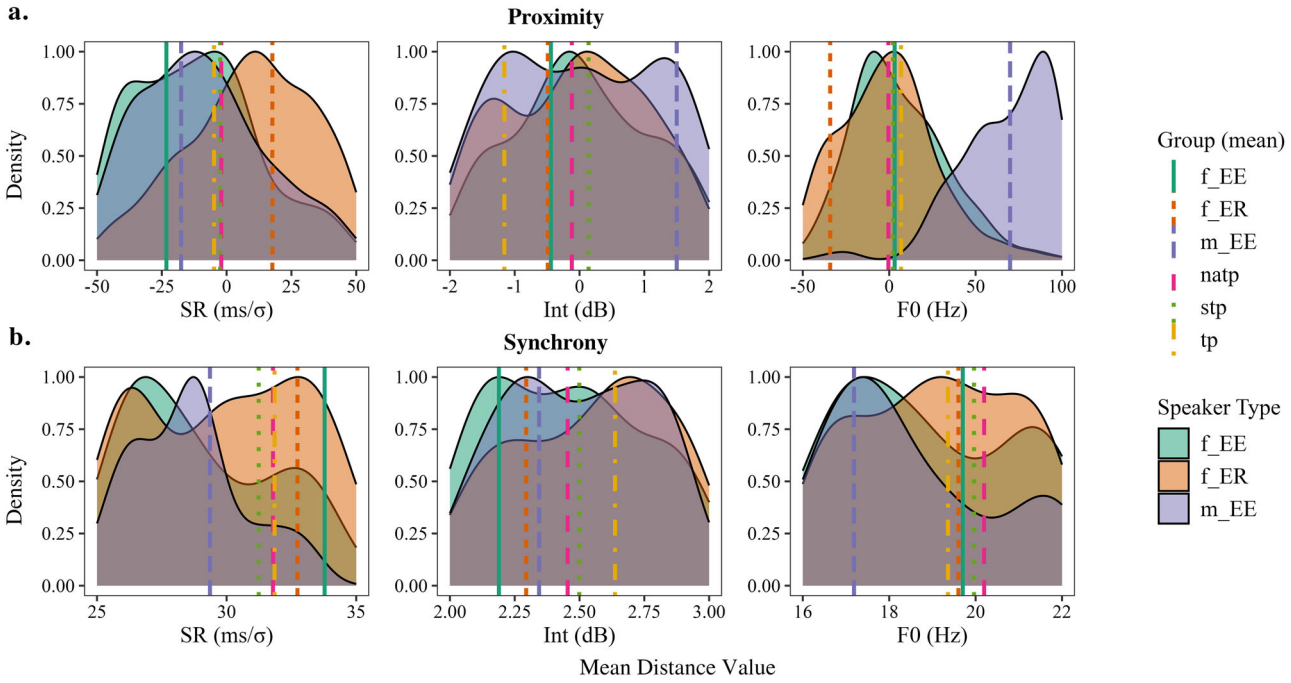

Fig. 1 Density distributions of global and local entrainment across speaker types. The distribution of global and local proximity (i.e., **a**)/synchrony (i.e., **b**) are across three feature sets (speaking rate [SR], intensity [Int], and pitch [F0]) for each speaker type. Smoothed density estimates are shown, with the mean feature distance values between turn pairs on the *x*-axis and the estimated density values scaled to a maximum of 1 on the *y*-axis. Coloured areas represent differences in entrainment among the three speaker types. Vertical lines indicate the mean values of four types of turn pairs (*atp*, *natp*, *stp*, *tp*), with *atp* further divided by the speaker's gender and role. The *natp*-line served as the reference line for local entrainment, while the *stp*- and *tp*-lines served as reference lines for global entrainment. Local positive entrainment was indicated by *atp*-lines (*f_EE*; *f_ER*; *m_EE*), which were located far left of the *natp*-reference line. Global entrainment, on the other hand, was marked by both *atp*-lines and the *stp*-reference line that were placed far left of the *tp*-reference line. Conversely, the opposite order indicated negative entrainment for both the local and global levels. f_EE female guest, m_EE male guest, f_ER female host, atp adjacent turn pairs, natp non-adjacent turn pairs, stp same-dyad turn pairs, tp different-dyads turn pairs.

Table 2 The results of the Mann–Whitney *U*-test for classifying the types of local/global proximity and synchrony across three speaker types.

Feature set	Speaker type	Local proximity		Local synchrony		Global proximity		Global synchrony	
		p-value	type	*p*-value	type	*p*-value	type	*p*-value	type
F0	*f_EE*	0.341 NS	(−) prox	0.958 NS	(−) sync	0.031*	(−) prox	0.791 NS	(−) sync
	f_ER	<0.001***	(+) prox	0.297 NS	(−) sync	<0.001***	(+) prox	0.412**	(−) sync
	m_EE	<0.001***	(−) prox	<0.001***	(+) sync	<0.001***	(−) prox	<0.001***	(+) sync
Int	*f_EE*	0.223 NS	(−) prox	0.017*	(+) sync	0.281 NS	(−) prox	<0.001***	(+) sync
	f_ER	0.056 NS	(−) prox	<0.001***	(−) sync	0.241 NS	(−) prox	<0.001***	(+) sync
	m_EE	<0.001***	(−) prox	0.636 NS	(−) sync	<0.001***	(−) prox	0.221 NS	(−) sync
SR	*f_EE*	<0.001***	(+) prox	0.513 NS	(−) sync	<0.001***	(+) prox	0.756 NS	(−) sync
	f_ER	<0.001***	(−) prox	0.454 NS	(−) sync	<0.001***	(−) prox	0.679 NS	(−) sync
	m_EE	<0.001***	(+) prox	0.501 NS	(−) sync	<0.001***	(+) prox	0.301 NS	(−) sync

The significant threshold is represented by the number of symbols ('***' $p < 0.001$, '**' $p < 0.01$, '*' $p < 0.05$, 'NS' $p >= 0.05$). *f_EE* female guest, *m_EE* male guest, *f_ER* female host, *F0* pitch, *Int* intensity, *SR* speaking rate, *(+/−) prox* positive/negative proximity, *(+/−) sync* positive/negative proximity.

intensity set, male guests and the female host showed a positive dynamic change over time, resulting in lower distance values between interlocutors, despite fluctuations in the middle of the dyad indicating dynamic adaptation rather than gradual entrainment. Female guests showed a fluctuating but downward trend in the line graph; (2) In the pitch set, female guests' distance values rose slightly at the beginning but decreased over time, indicating positive dynamic entrainment. Both male guests and the female host showed a decreasing trend, with the female host exhibiting a slight downward tendency mid-conversation and steep changes at the beginning and end; (3) For the speaking rate, the female host displayed negative dynamic entrainment with high mean distance values and some falling points, while both female and male guests showed a stepped-down descent during the middle of the conversation, with rising tendencies in the last turns. The ADF test revealed that all three feature sets exhibited dynamic nature ($p > 0.05$), indicating positive dynamic entrainment in pitch and intensity within the Chinese talk show (Table 3). However, the speaking rate showed a lower degree of positive dynamic entrainment compared to the other features.

Positive dynamic entrainment was the most prevalent, followed by negative dynamic entrainment, with static entrainment being the least apparent. The pitch set exhibited the most positive dynamic valence and the most static valence, despite the

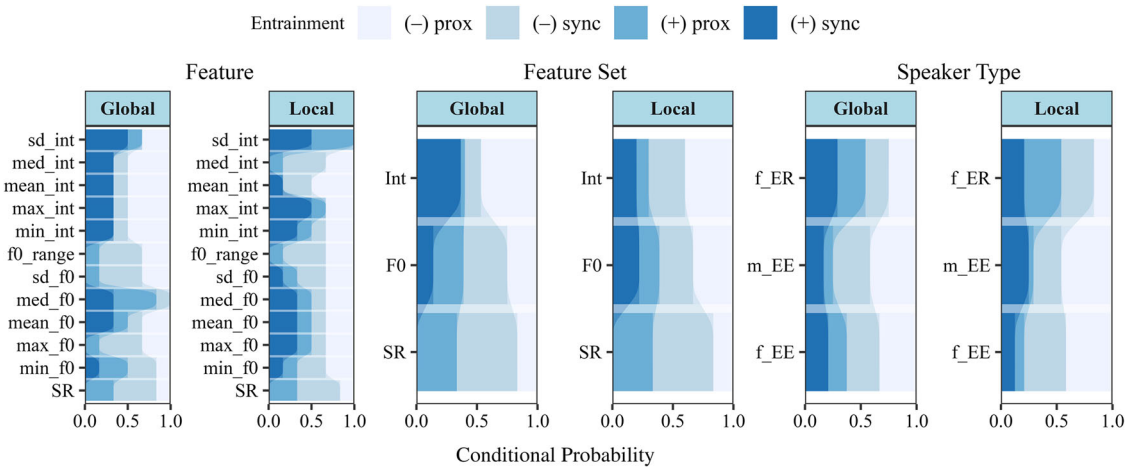

Fig. 2 Conditional global and local entrainment probabilities for proximity and synchrony in each feature, feature set and speaker type. Conditional probabilities are calculated by dividing the number of positive or negative entrainment occurrences by the total entrainment occurrences in proximity or synchrony, locally or globally. (+/-) pro positive/negative proximity, (+/-) sync positive/negative proximity, f_EE female guest, m_EE male guest, f_ER female host, *F0* pitch, *Int* intensity, *SR* speaking rate.

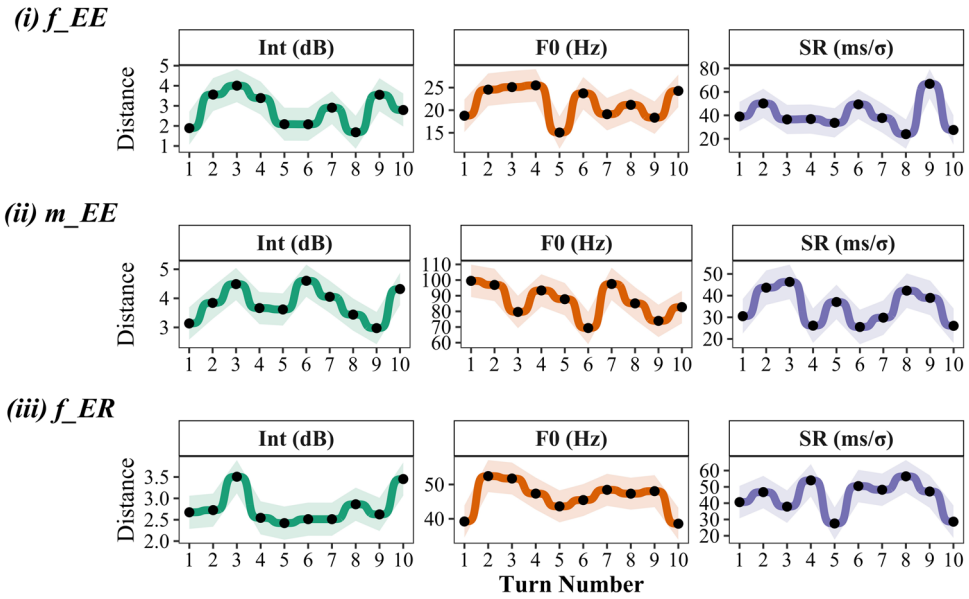

Fig. 3 Mean distances over ten turn pairs within each conversation in three feature sets across three speaker types. f_EE female guest, m_EE male guest, f_ER female host, F0 pitch, Int intensity, SR speaking rate.

minimum pitch. Intensity and speaking rate had similar dynamicity valence. Male guests displayed more positive dynamic entrainment towards their interlocutors, while female guests showed a lesser degree. Thus, the valence of dynamicity was: positive dynamic > negative dynamic > static.

Perceptual judgement results. Our results revealed different patterns in perceived similarity across different speaker combinations (Fig. 5). Male guests showed the most pronounced increase in similarity over time, with their scores ranging from 4 to 5, showing a significant difference between the first and third conversation thirds ($p = 0.04$). In contrast, the female host and female guests exhibited lower similarity rating scores (3–4 and approximately 4, respectively) with less significant changes over time (*f_ER*: $p = 0.13$; *f_EE*: $p = 0.36$). These results suggested a potential ceiling effect for same-gender (i.e., female-female) conversations. The higher initial similarity perceived in female-female conversations compared to female-male conversations might have limited further increases in similarity over time, explaining the less pronounced increases in similarity ratings for the female host and guests. While there was some variability across dyads, with some showing more similarity in the first portion of the conversation third than others (e.g., a decrease from 5 to 3), our study primarily focused on the general trend of similarity ratings. The mean similarity rating scores in the first (i.e., M_{1st}) and third (i.e., M_{3rd}) portions of the conversation thirds for all three speaker types displayed an increasing tendency (*m_EE*: M_{1st} (20) = 3.47, M_{3rd} (20) = 3.86; *f_ER*: M_{1st} (38) = 4.07, M_{3rd} (38) = 4.29; *f_EE*: M_{1st} (18) = 4.22, M_{3rd} (18) = 4.39)[6].

We then examined the differences between the first and third portions of the conversation thirds by analysing similarity rating scores for each listener (Fig. 6). Male guests were perceived as becoming the most similar over the course of the dyad, with the lowest negative mean similarity difference values ($M = -0.38$, $SD = 0.36$). The female host ($M = -0.14$, $SD = 0.26$) and female

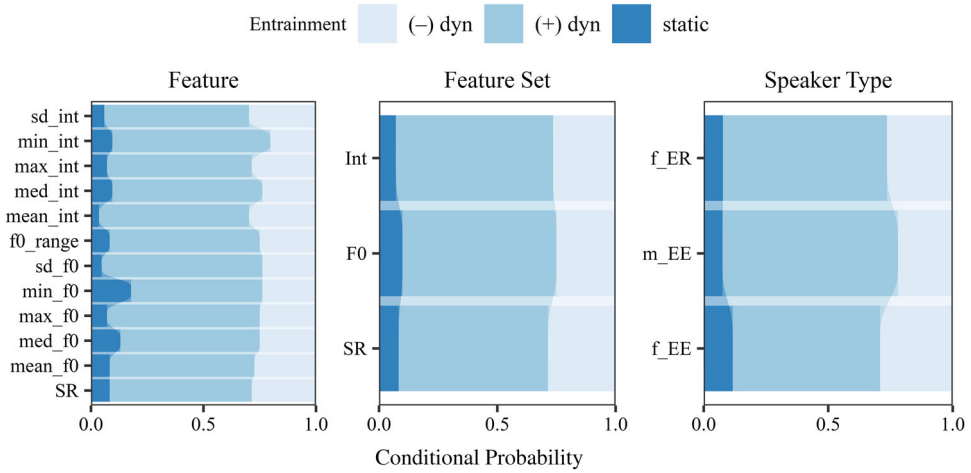

Fig. 4 Conditional probabilities of dynamicity entrainment for each feature, feature set, and speaker type. A line chart represents the changing process of mean distance values across ten sections, with coloured ribbons illustrating the fluctuation range of each point. The range of each point is depicted between $[y + y_{min}, y + y_{max}]$, where y_{min} and y_{max} represent the minimum and maximum values of all mean feature distance values (y), respectively. Static static entrainment, (+/-) dyn positive/negative dynamic entrainment. f_EE female guest, m_EE male guest, f_ER female host, F0 pitch, Int intensity, SR speaking rate.

Table 3 The dynamicity types obtained from the ADF test results for three feature sets, namely intensity, pitch, and speaking rate, across three speaker types.

Feature set	Speaker type	Statistic: *DF*	*p*-value	Dynamicity
Int	*f_EE*	-2.711	0.205	(+) dyn
	m_EE	-4.333	0.099	(+) dyn
	f_ER	-2.567	0.215	(+) dyn
F0	*f_EE*	-2.614	0.210	(+) dyn
	m_EE	-6.510	0.054	(+) dyn
	f_ER	-1.458	0.348	(+) dyn
SR	*f_EE*	-5.238	0.066	(+) dyn
	m_EE	-3.479	0.146	(+) dyn
	f_ER	-1.618	0.327	(-) dyn

ADF test Augmented Dickey-Fuller test, *DF* degree of freedom, *f_EE* female guest, *m_EE* male guest, *f_ER* female host, *F0* pitch, *Int* intensity, *SR* speaking rate.

guests ($M = -0.14$, $SD = 0.44$) also showed increased similarity, but with slightly higher mean difference values. Notably, the positive values found for some female guests indicate the least similarity between the first and third portions of the conversation.

Statistical analysis using a Bonferroni-corrected alpha level of 0.008 for three *t*-tests confirmed that all distributions were significantly different from zero (*m_EE*: $t(40) = -6.78$, $p < 0.001$, Cohen's $d = 1.50$; *f_ER*: $t(40) = -3.45$, $p = 0.001$, Cohen's $d = 0.76$; *f_EE*: $t(40) = -2.06$, $p = 0.046$, Cohen's $d = 0.46$). The results from a linear mixed-effects model revealed a significant main effect of speaker type ($F(2, 120) = 6.06$, $p = 0.003$, ${}_p\eta^2 = 0.09$). Post-hoc analyses indicated significant differences between male guests and the female host ($\beta = 0.24$, $SE = 0.08$, $t = 3.04$, $p = .009$), as well as between male guests and female guests ($\beta = 0.23$, $SE = 0.08$, $t = 2.99$, $p = 0.01$). However, no significant difference was observed between the female host and female guests ($\beta = -0.003$, $SE = 0.08$, $t = -0.05$, $p = 0.99$).

Correlations between perceived similarity and acoustic-prosodic measures. We computed correlations between perceived similarity and 12 acoustic-prosodic features among different speaker types, focusing on those with significant correlations[7]. Negative values in mean similarity difference indicate increased similarity over time, so a negative correlation suggests positive entrainment behaviour. In Fig. 7 for local proximity entrainment, among male guests, the correlation between perceived similarity and pitch was significant ($r = 0.46$, $p = 0.002$, 95% CI [0.18, 0.67]). For female guests, none of the correlations reached statistical significance. Linear mixed-effect models showed a strong interaction between perceived similarity and speaker type ($F(2, 117) = 4.12$, $p = 0.01$) for local proximity in pitch, predominantly in male guests. This result indicated that as the acoustic distances in pitch decreased for male guests, perceived similarity also decreased. No significant correlation was found for speaking rate ($F(1, 117) = 3.11$, $p = 0.08$) or intensity ($F(1, 117) = 0.03$, $p = 0.87$), suggesting pitch was a more salient factor in perceptual similarity judgements when interlocutors accommodated each other over time. Speaker type significantly affected speaking rate ($F(2, 117) = 8.03$, $p < 0.001$) and pitch ($F(2, 117) = 66.69$, $p < 0.001$).

Concerning local synchrony entrainment (Fig. 8), the correlations between perceived similarity and most of the features were generally weak and negative, with perceived similarity and maximum pitch showing a significant negative correlation ($r = -0.22$, $p = 0.01$, 95% CI [-0.38, -0.04]), as well as perceived similarity and mean pitch ($r = -0.22$, $p = 0.03$, 95% CI [-0.35, -0.01]), and perceived similarity and pitch range ($r = -0.20$, $p = 0.02$, 95% CI [-0.36, -0.02]). These negative correlations suggested that higher perceived similarity scores are associated with lower pitch values and reduced pitch variability. The relationship between perceived similarity and intensity was not significant ($r = -0.03$, $p = 0.69$, 95% CI [-0.21, 0.14]), but a weak negative correlation was observed between perceived similarity and the intensity of speech in male guests ($r = 0.32$, $p = 0.04$, 95% CI [0.01, 0.57]). Linear mixed-effect models indicated weak negative correlations between perceived similarity and all three pitch features of maximum pitch ($F(1, 117) = 6.20$, $p = 0.01$), mean pitch ($F(1, 117) = 4.50$, $p = 0.04$), and pitch range ($F(1, 117) = 5.09$, $p = 0.03$), with the effect being predominantly observed in male guests. These findings suggested that parallel pitch changes between the host and male guests enhance listeners' recognition of prosodic similarity. There was no significant correlation between perceived similarity and speaking rate in local

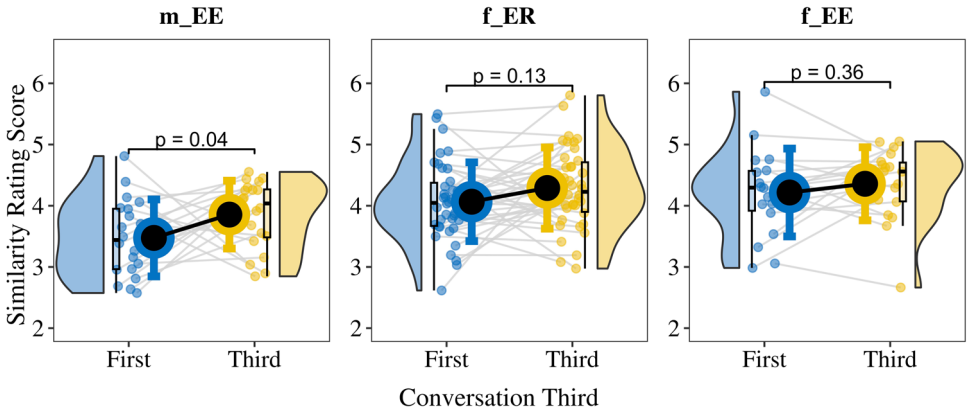

Fig. 5 Mean similarity rating scores in each conversation third for female guest (*f_EE*), male guest (*m_EE*), and female host (*f_ER*). Error bars represent the standard error of the mean. The large black circle indicates the mean similarity score for each conversation third (i.e., first vs. third), while the small circles show the mean similarity score of each dyad. Matching lines highlight the difference in mean similarity ratings between the first and third portions of the conversation thirds.

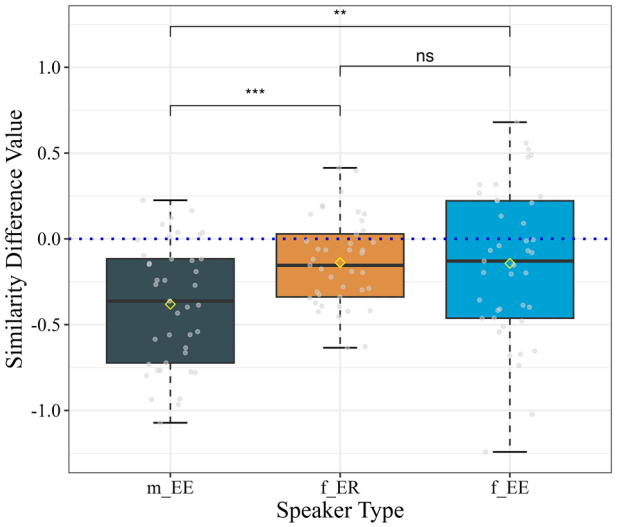

Fig. 6 Similarity rating difference values between the first and third portions of each conversation by speaker type. A negative value implies increasing linguistic similarity, while a positive value suggests linguistic divergence. f_EE female guest, m_EE male guest, f_ER female host.

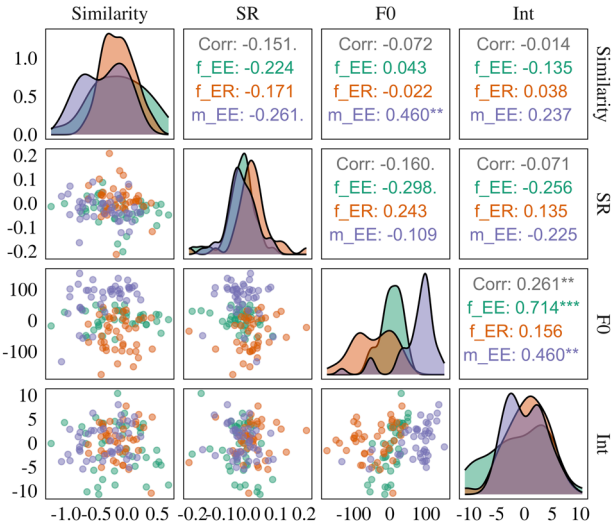

Fig. 7 Correlations between perceived similarity and acoustic-prosodic measures in local proximity across speaker types. Corr Pearson Correlation Coefficient (*r*), f_EE female guest, m_EE male guest, f_ER female host, F0 pitch, Int intensity, SR speaking rate.

synchrony ($F(1, 117) = 1.07$, $p = 0.30$), indicating that variations in speaking rate were less noticeable to listeners. Moreover, no significant effects of speaker types were found in any acoustic-prosodic features (*SR*: $F(2, 117) = 1.33$, $p = 0.26$; *max_f0*: $F(2, 117) = 0.37$, $p = 0.68$; *mean_f0*: $F(2, 117) = 0.10$, $p = 0.90$; *f0_range*: $F(2, 117) = 0.39$, $p = 0.67$; *Int*: $F(2, 117) = 0.71$, $p = 0.49$).

Discussion

Our study on prosodic entrainment in Mandarin Chinese talk shows has yielded several important findings. First, negative entrainment was more prevalent than positive entrainment across the three feature sets (pitch, intensity, and speaking rate), particularly evident in the negative synchrony in speaking rate. This unexpected finding challenges conventional assumptions about the benefits of positive entrainment in communication. Second, we found that the female host demonstrated larger positive entrainment compared to both male and female guests, indicating a complex interplay between gender, role, and entrainment behaviour, and different gender-role strategies may be employed to establish common ground in talk show settings. Lastly, we observed a moderate negative correlation between perceived similarity and pitch for male guests in both local proximity and synchrony, suggesting that pitch plays a crucial role in listeners' perception of prosodic similarity.

Prosodic entrainment patterns in long-and-short talk shows. Our analysis revealed that negative entrainment was more prevalent than positive entrainment across the three feature sets, particularly evident in the negative synchrony in speaking rate. While this finding might seem counterintuitive, it can be interpreted as beneficial for the success of a talk show when viewed from a broader Communication Accommodation Theory perspective. According to CAT, negative entrainment can occur in instances of patronising communication or when speakers deliberately diverge from their interlocutors' speech patterns (Giles et al., 1991; Paquette-Smith et al., 2022; Soliz and Giles, 2014). In the context of our study, the larger negative entrainment observed in the three feature sets can be interpreted as a deliberate pattern employed by guests to tell their stories, which

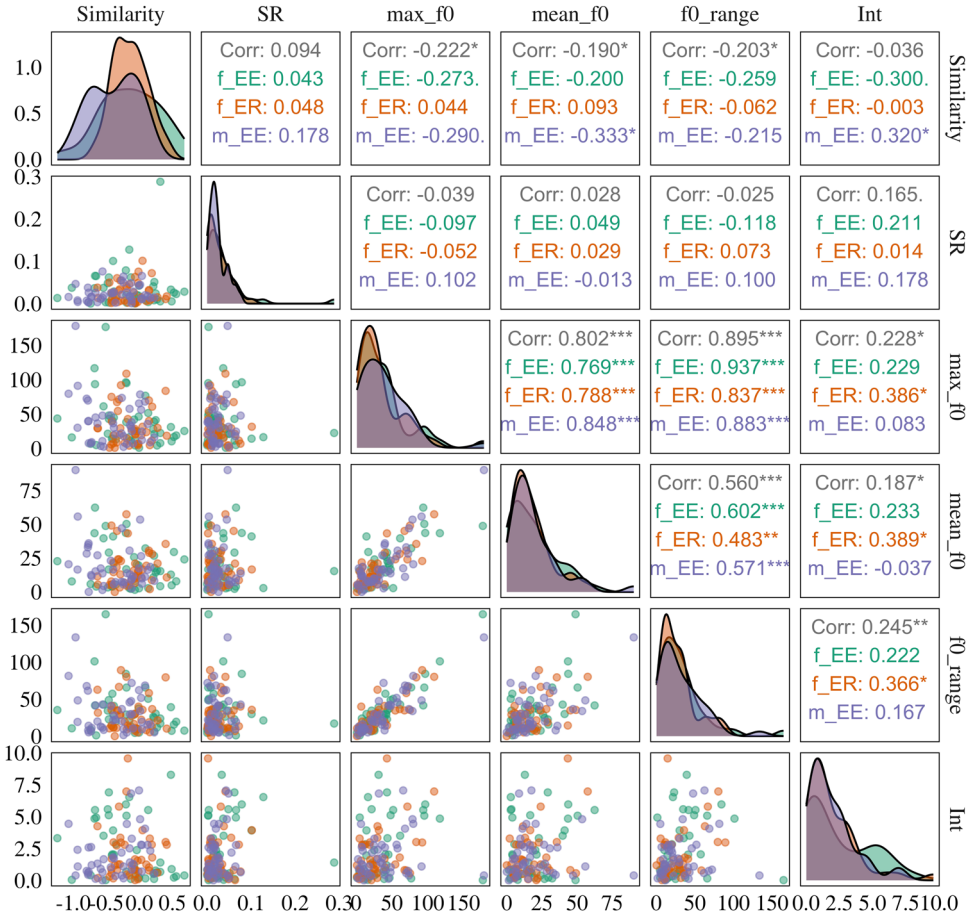

Fig. 8 Correlations between perceived similarity and acoustic-prosodic measures in local synchrony across speaker types. Corr Pearson correlation coefficient (*r*), f_EE female guest, m_EE male guest, f_ER female host, F0 pitch, Int intensity, SR speaking rate.

diverges from the host's preferred method of brief chatting. This interpretation is supported by examples from our speech corpus where guests (e.g., *Beining SA*; *Bo HUANG*) imitated other people's voice features, such as dialectal and tonal elements, to enhance the atmosphere of their stories. As a result, the host and guests exhibited contrasting speaking styles, emphasising the nature of negative entrainment in this particular setting.

The observed negative entrainment might serve several functions in the talk show context: (1) The host and guests may use contrasting prosodic features to maintain their distinct roles, enhancing the dynamic nature of the conversation (Giles et al., 1991); (2) Contrasts in prosodic features can create a more lively and engaging dialogue, potentially increasing audience interest (Levitan et al., 2012); (3) Negative entrainment in certain prosodic features might help signal turn transitions, contributing to smoother conversation flow (Levitan and Hirschberg, 2011). These findings challenge the conventional understanding that positive entrainment is always beneficial for communication and suggest that in certain contexts, such as talk shows, negative entrainment can play a crucial role in creating engaging and dynamic conversations.

Gender, role, and prosodic entrainment behaviour. Our study also provides novel insights into the effects of gender and role on entrainment behaviour during talk show conversations. We found that the female host demonstrated larger positive entrainment compared to both male and female guests. This large valence suggests that the female host possessed a higher degree of control in terms of turn-taking, leading to a more natural and appropriate flow of conversation. This finding aligns with research that revealed an imbalance in power between guests and hosts in most interviews, with hosts generally viewed as having higher power and greater control over the conversation (Drew and Heritage, 1992). The phenomenon of larger positive prosodic entrainment in the female host is also consistent with CAT, which posits that the less powerful speaker tends to entrain more when a power imbalance exists in a discussion. The female host's distinct approach, characterised by soothing and friendly questions, establishment of trust, and facilitation of relaxed conversations, may have contributed to this positive entrainment. This approach likely helped relieve guests' nervousness and bridge the gap between them, regardless of the guests' status.

Interestingly, our results revealed different entrainment patterns for male and female guests. At the local level, mixed-gender pairs with different roles (i.e., male guests with the female host) showed greater positive synchrony entrainment during turn-taking. At the global level, same-gender pairs with different roles (i.e., female guests with the female host) demonstrated greater positive entrainment behaviour across the entire conversation. These findings suggest that different gender-role strategies may be employed to establish common ground in a talk show setting. The observed differences in entrainment patterns between male and female guests could be attributed to various factors, including sociocultural norms, individual communication styles, or the specific dynamics of the talk show format (Holmes, 2013; Holmes, 1988; Tannen, 1990). For instance, male guests might focus more on immediate turn-by-

turn interactions, while female guests might prioritise overall conversational harmony. However, it's important to note that these patterns are likely influenced by a complex interplay of individual, situational, and cultural factors, rather than being solely determined by gender (Eckert and McConnell-Ginet, 2013). Our results highlight the complex effects of gender-role interactions on prosodic entrainment, which previous studies (Bilous and Krauss, 1988; Levitan et al., 2012; Pardo et al., 2018; Reichel et al., 2018; Weise et al., 2019; Xia et al., 2014) have not fully explored. These findings suggest that entrainment cannot be predicted straightforwardly by the function of dominance alone, as proposed by CAT (Giles et al., 1991; Giles et al., 1987; Giles and Ogay, 2007). Furthermore, our results indicate that the perception-behaviour link (Chartrand and Bargh, 1999) may not serve as an exhaustive explanation for the biological gender influence on entrainment, as different roles among guests also play a vital role in entrainment behaviour. We thus view this outcome as the interaction between gender and role, suggesting a more nuanced understanding of entrainment in conversation dynamics in the long-and-short turn-taking talk shows.

Perceived similarity and acoustic-prosodic features. A moderate negative correlation was observed between perceived similarity and pitch for male guests in both local proximity and synchrony. However, it's noteworthy that the overall similarity rating scores were relatively low. This finding may be attributed to biological gender differences, which could make it easier for listeners to distinguish between speakers' turn-taking and judge changes in prosodic similarity more readily when dealing with male guests than female guests. Whilst our study focused on the impact of speakers' gender on perceived similarity, it is worth noting that listeners' own gender could potentially influence their perceptual judgements as well. Research has shown that listeners' experiences and characteristics, including gender, can affect speech perception in general (Johnson et al., 1999; Strand, 1999). In the context of our study, male and female participants might differ in their sensitivity to pitch and intensity changes, potentially affecting their judgements of speaker similarity. While this aspect is outside the immediate focus of our study, it underscores the complexity of how gender dynamics may affect similarity ratings, warranting further exploration in future research. Nevertheless, the inherent vocal differences between male and female speakers might allow listeners to detect prosodic changes in mixed-gender pairs more easily. Conversely, when judging two female speakers, listeners may perceive their voices as more similar than they actually are, making changes in similarity less evident than in mixed-gender pairs. These findings suggested that while entrainment occurred across all speaker combinations, the potential for increased similarity over time might be influenced by gender pairing. Same-gender pairs, particularly female-female interactions, might start at a higher baseline of perceived similarity, potentially limiting the observable increase in entrainment due to ceiling effects. In contrast, mixed-gender pairs (female host-male guest) showed a more pronounced increase in similarity, possibly due to a lower initial baseline and greater room for convergence.

Interestingly, our findings indicate that pitch features are more prominent than intensity in listeners' ability to judge similarity. This finding contrasts with previous studies that found intensity and pitch to be equally evident in terms of proximity and synchrony when analysing acoustic-prosodic measures (Beňuš et al., 2014; Levitan et al., 2015). Our study suggests that subjective perceptual judgement overall reflects that changes in lower and higher registers (pitch) are more prominent than volume changes (intensity) during speech accommodation in Mandarin Chinese interviews. The prominence of pitch in perceptual similarity judgements is closely tied to Mandarin's tonal nature. Pitch, governed by the vocal folds, is flexible and consciously controllable, carrying lexical meaning through tones in Mandarin (Xu, 1994). This character makes speakers highly sensitive to pitch variations. In Mandarin interviews, both interviewers and interviewees may skillfully adjust pitch to convey meaning, build rapport, and adapt to their conversational partner (Gussenhoven, 2004), reflecting the significant role of pitch in communicative accommodation.

In addition, the crucial role of pitch in listeners' perception of prosodic similarity in Mandarin Chinese interviews may be attributed to its unique characteristics compared to other acoustic cues: (1) In Mandarin Chinese, pitch plays a crucial role in listeners' perception of prosodic similarity due to its unique dual function as both a carrier of lexical meaning and a cue for intonation and emphasis (Xu and Wang, 2001). This dual function of pitch makes pitch a primary focus of attention in speech, especially in interviews, where it may enhance prosodic entrainment; (2) Adjusting pitch requires less cognitive effort compared to modifying speaking rate or intensity, which often involves changes in breathing, articulation, and pacing. Slowing down or speeding up speech can disrupt the rhythm and fluency of conversation, whereas adjusting pitch can be accomplished without significantly altering these elements (Patel and Schell, 2008). In the context of Mandarin Chinese interviews, where maintaining fluency and coherence is crucial, pitch adjustment may be the most efficient means of achieving prosodic entrainment; (3) The tonal nature of Mandarin, combined with common exposure to musical training, heightens speakers' sensitivity to pitch variation, increasing its prominence in perceptual similarity judgements (Patel, 2011). This enhanced awareness of pitch could contribute to its prominence in perceptual similarity judgements during interviews, where listeners may be particularly attuned to tonal and intonational patterns; (4) Pitch modulation in Mandarin also signals politeness, respect, and social status (Yip, 1980). In interview contexts, both interviewers and interviewees may adjust their pitch to navigate social dynamics and build rapport, reinforcing its role in prosodic entrainment. These factors highlight the central role of pitch in Mandarin Chinese interviews, where its lexical and pragmatic functions make it crucial for prosodic similarity judgements. The ease of pitch adjustment compared to other acoustic cues warrants further research on its influence in Mandarin, as well as cross-linguistic studies to explore prosodic entrainment patterns in both tonal and non-tonal languages.

Limitations and future studies. While our study provides some insights into prosodic entrainment in Mandarin Chinese talk shows, several limitations should be acknowledged. First, our study focused on three specific acoustic-prosodic features (pitch, intensity, and speaking rate). However, entrainment may manifest in other linguistic dimensions not captured in our analysis, such as segmental features or tones. Future research could expand on our findings by investigating a broader range of features, potentially uncovering additional patterns of entrainment in long-and-short turn-taking contexts. Second, the relatively small sample size of our corpus may have affected the results. Further studies with a larger corpus are recommended to explore these findings in greater detail and increase the generalisability of the results. Third, having only one host as a case study limits the ability to make broad claims about talk show hosts in general. Future work should include multiple hosts to allow for conclusions that can address hosts and speech entrainment more broadly. Additionally, the consideration of age as a variable was

not included in our study, which may affect the turn-taking styles and preferences exhibited by different age groups for interviewees. Future research could investigate how age influences prosodic entrainment patterns, providing a more comprehensive understanding of conversational dynamics. Lastly, our study focused on Mandarin Chinese talk shows, which may have unique cultural and linguistic features. Cross-cultural studies comparing entrainment patterns in talk shows across different languages and cultures could provide valuable insights into the universality or cultural specificity of our findings. While we considered gender and role, other contextual factors such as topic familiarity, emotional state, or interpersonal relationships between speakers were not accounted for. Future research could explore how these factors influence entrainment patterns in talk show settings.

Conclusions

This study provides novel insights into prosodic entrainment during long-and-short turn-taking in Mandarin Chinese talk shows, revealing complex interactions between gender, role, and acoustic-prosodic features (pitch, intensity, and speaking rate). We observed a prevalence of negative entrainment across pitch, intensity, and speaking rate, suggesting that such contrasting speech patterns can contribute to engaging and dynamic conversations in talk show contexts. Our results also highlight complex gender-role interactions in entrainment behaviour, with the female host demonstrating larger positive entrainment, while male and female guests showed different entrainment patterns at local and global levels. Additionally, we found that pitch plays a crucial role in listeners' perception of prosodic similarity, more so than intensity. These findings contribute to a more nuanced understanding of prosodic entrainment in long-and-short turn-taking contexts, emphasising the need to consider gender, role, and specific communicative settings in future research. While our study focused on Mandarin Chinese talk shows, future cross-cultural investigations and examination of a broader range of acoustic-prosodic features could further enhance our understanding of entrainment dynamics in various conversational contexts.

Data availability

All corpus data and scripts analysed during the current study are available in the OSF repository at https://osf.io/ybx86/. The data folder contains three main folders. The first main folder is corpus data, with WAV files and corresponding TXT files. The second main folder is scripts and data, containing acoustic-related scripts such as *Praat* scripts and *Reaper* shell, as well as original data from acoustic and perceptual analysis. The third folder contains all scripts for statistical analyses and visualisations of all acoustic, perceptual and correlation analyses.

Received: 21 July 2024; Accepted: 13 November 2024;
Published online: 27 November 2024

Notes

1 In computer science, entrainment is commonly referred to as alignment, accommodation, turn-takings, audience design, mimicry, or similar terms used in psychology, sociology, and other related fields.

2 This *Praat* script is designed to create a *TextGrid* object for a *LongSound* object and sets boundaries at pauses based on an intensity analysis. This script is written by Mietta Lennes and is distributed under the GNU General Public License.

3 This *Praat* script designed to split long sound files into several small ones was written by *Pengfei SHAO*, which can be found at https://blog.csdn.net/shaopengfei/article/details/109097404.

4 The ten sections were defined to evenly divide the conversation into ten equal intervals, allowing us to observe and compare the dynamic changes in entrainment over the entire duration. This segmentation provides a detailed and granular view of how entrainment evolves over time.

5 It is important to note that although the highest density for female guests (i.e., green colour area) was distributed around 27.5 mean distance values, the mean values of the female guest group was 34. Therefore, the order of degrees was mainly arranged according to the mean values of the group.

6 The mean similarity rating scores for each speaker type were calculated for two conversation thirds, with the number of conversations containing male guests indicated in parentheses. For instance, the first portion of the conversation third yielded a mean similarity rating score of 3.47, while the third portion yielded a mean score of 3.86 across all 20 conversations featuring male guests.

7 It should be noted that we have only listed the feature sets and specific features for which significant correlations occurred, given the relatively small sample size of our study ($n = 41$). Therefore, these correlations should be taken in stride.

References

Abel J, Babel M (2017) Cognitive load reduces perceived linguistic convergence between dyads. Lang Speech 60(3):479–502. https://doi.org/10.1177/0023830916665652

Babel M (2010) Dialect divergence and convergence in New Zealand English. Lang Soc 39(4):437–456. https://doi.org/10.1017/s0047404510000400

Babel M (2012) Evidence for phonetic and social selectivity in spontaneous phonetic imitation. J Phon 40(1):177–189. https://doi.org/10.1016/j.wocn.2011.09.001

Babel M, Bulatov D (2012) The role of fundamental frequency in phonetic accommodation. Lang Speech 55(2):231–248. https://doi.org/10.1177/0023830911417695

Babel M et al. (2014) Novelty and social preference in phonetic accommodation. Lab Phonol 5(1):123–150. https://doi.org/10.1515/lp-2014-0006

Bates D et al. (2015) Fitting linear mixed-effects models using lme4. J Stat Soft 67(1):1–48. https://doi.org/10.18637/jss.v067.i01

Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. Ann Stat 29(4):1165–1188. https://doi.org/10.1214/aos/1013699998

Beňuš S et al. (2014) Entrainment in Slovak collaborative dialogues. In: Proceedings of the 5th Conference on Cognitive Infocommunications (CogInfoCom), IEEE, Vietri sul Mare, Italy, 5–7 Nov 2014

Bilous FR, Krauss RM (1988) Dominance and accommodation in the conversational behaviours of same- and mixed-gender dyads. Lang Commun 8(3-4):183–194. https://doi.org/10.1016/0271-5309(88)90016-X

Boersma P, Weenink D (2024) Praat: Doing phonetics by computer. http://www.praat.org/. Accessed 2 Mar. 2023

Borrie SA, Lubold N, Pon-Barry H (2015) Disordered speech disrupts conversational entrainment: A study of acoustic-prosodic entrainment and communicative success in populations with communication challenges. Front Psychol 6:1187. https://doi.org/10.3389/fpsyg.2015.01187

Bourhis RY, Giles H (1977) The language of intergroup distinctiveness. In: Giles H (ed) Language, ethnicity and intergroup relations. Academic Press, London, UK, pp. 119–135

Bradshaw AR, McGettigan C (2021) Convergence in voice fundamental frequency during synchronous speech. PLoS ONE 16(10):e0258747. https://doi.org/10.1371/journal.pone.0258747

Caspers J (2003) Local speech melody as a limiting factor in the turn-taking system in Dutch. J Phon 31(2):251–276. https://doi.org/10.1016/S0095-4470(03)00007-X

Chartrand TL, Bargh JA (1999) The chameleon effect: The perception-behavior link and social interaction. J Pers Soc Psychol 76(6):893–910. https://doi.org/10.1037/0022-3514.76.6.893

Cohen Priva U, Edelist L, Gleason E (2017) Converging to the baseline: Corpus evidence for convergence in speech rate to interlocutor's baseline. J Acoust Soc Am 141(5):2989–2996. https://doi.org/10.1121/1.4982199

Collins B (1998) Convergence of fundamental frequencies in conversation: If it happens, does it matter? In: Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP 1998), International Speech Communication Association, Sydney, Australia, Nov 30–Dec 4 1998

Dallaston K (2023) MacReaper. https://kjdallaston.com/projects. Accessed 5 Apr. 2023

Danescu-Niculescu-Mizil C et al. (2012) Echoes of power: Language effects and power differences in social interaction. In: Proceedings of the 21st International Conference on World Wide Web, Association for Computing Machinery, Lyon, France, 16–20 Apr 2012

De Looze C et al. (2014) Investigating automatic measurements of prosodic accommodation and its dynamics in social interaction. Speech Commun 58:11–34. https://doi.org/10.1016/j.specom.2013.10.002
Drew P, Heritage J (1992) Analyzing talk at work: An introduction. In: Drew P, Heritage J (eds) Talk at work: Interaction in institutional settings. Cambridge University Press, Cambridge, pp 3–65
Eckert P, McConnell-Ginet S (2013) Language and gender. Cambridge University Press, Cambridge
Fox J, Weisberg S (2011) An R companion to applied regression. Sage, London
Giles H, Coupland N, Coupland J (1991) Accommodation theory: Communication, context, and consequence. In: Giles H, Coupland J, Coupland N (ed) Contexts of Accommodation. Cambridge University Press, Cambridge, UK, pp. 1–68
Giles H et al. (1987) Speech accommodation theory: The first decade and beyond. Ann Int Commun Assoc 10:13–48. https://doi.org/10.1080/23808985.1987.11678638
Giles H, Ogay T (2007) Communication accomodation theory. In: Whaley B, Samter W (ed) Explaining communication: contemporary theories and exemplars. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 293–310
Gregory SWJ, Webster S (1996) A nonverbal signal in voices of interview partners effectively predicts communication accommodation and social status perceptions. J Pers Soc Psychol 70(6):1231–1240. https://doi.org/10.1037/0022-3514.70.6.1231
Gussenhoven C (2004) The phonology of tone and intonation. Cambridge University Press, Cambridge
Guydish AJ, D'Arcey JT, Fox Tree JE (2021) Reciprocity in conversation. Lang Speech 64(4):859–872. https://doi.org/10.1177/0023830920972742
Guydish AJ, Fox Tree JE (2022) Reciprocity in instant messaging conversations. Lang Speech 65(2):404–417. https://doi.org/10.1177/00238309211025070
Hamilton JD (2020) Time series analysis. Princeton University Press, Princeton, New Jersey
Hirst D (2011) The analysis by synthesis of speech melody: From data to models. J Speech Sci 1(1):55–83. https://doi.org/10.20396/joss.v1i1.15011
Holmes J (2013) Women, men and politeness. Routledge, London and New York
Holmes VM (1988) Hesitations and sentence planning. Lang Cognitive Proc 3(4):323–361. https://doi.org/10.1080/01690968808402093
Johnson K, Strand EA, D'Imperio M (1999) Auditory–visual integration of talker gender in vowel perception. J Phon 27(4):359–384. https://doi.org/10.1006/jpho.1999.0100
Keating PA, Garellek M, Kreiman J (2015) Acoustic properties of different kinds of creaky voice. Paper presented at the 18th International Congress of Phonetic Sciences, Glasgow, Scotland, 10–14 Aug. 2015
Kendall T (2009) Speech rate, pause, and linguistic variation: An examination through the sociolinguistic archive and analysis project. Dissertation, Duke University
Kim M, Horton WS, Bradlow AR (2011) Phonetic convergence in spontaneous conversations as a function of interlocutor language distance. Lab Phonol 2(1):125–156. https://doi.org/10.1515/labphon.2011.004
Ko E-S et al. (2015) Entrainment of prosody in the interaction of mothers with their young children. J Child Lang 43(2):284–309. https://doi.org/10.1017/S0305000915000203
Lee C-C et al. (2010) Quantification of prosodic entrainment in affective spontaneous spoken interactions of married couples. In: Proceedings of INTERSPEECH 2010, International Speech Communication Association, Makuhari, Chiba, Japan, 26–30 Sept 2010
Lee Y et al. (2018) Articulatory, acoustic, and prosodic accommodation in a cooperative maze navigation task. PLoS ONE 13(8):1–26. https://doi.org/10.1371/journal.pone.0201444
Levitan R et al. (2015) 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, Association for Computational Linguistics, Prague, Czech Republic, 2–4 Sept 2015
Levitan R et al. (2012) 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, Association for Computational Linguistics, Montréal, Canada, 3–8 Jun 2012
Levitan R, Hirschberg J (2011) Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In: Proceedings of INTERSPEECH 2011, International Speech Communication Association, Florence, Italy, 27–31 Aug 2011
Levitan SI, Xiang J, Hirschberg J (2018) Acoustic-prosodic and lexical entrainment in deceptive dialogue. In: Proceedings of Speech Prosody 2018, International Speech Communications Association, Poznań, Poland, 13–16 Jun 2018
Lewandowski EM, Nygaard LC (2018) Vocal alignment to native and non-native speakers of English. J Acoust Soc Am 144(2):620–633. https://doi.org/10.1121/1.5038567
Liu H (2004) Conversation analysis: An introduction. Peking University Press, Beijing
Livieris IE et al. (2021) Smoothing and stationarity enforcement framework for deep learning time-series forecasting. Neural Comput Appl 33(20):14021–14035. https://doi.org/10.1007/s00521-021-06043-1
Loveday L (1981) Pitch, politeness and sexual role: An exploratory investigation into the pitch correlates of English and Japanese politeness formulae. Lang Speech 24(1):71–89. https://doi.org/10.1177/002383098102400105
Matarazzo JD, Wiens AN (1967) Interviewer influence on durations of interviewee silence. J Exp Res Pers 2:56–69
McAuliffe M et al. (2017) Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. In: Proceedings of INTERSPEECH 2017, International Speech Communications Association, Stockholm, Sweden, 20–24 Aug 2017
Michalsky J, Schoormann H (2017) Pitch convergence as an effect of perceived attractiveness and likability. In: Proceedings of INTERSPEECH 2017, International Speech Communications Association, Stockholm, Sweden, 20–24 Aug 2017
Mondada L (2001) Pour une linguistique interactionnelle. Marg Ling 1:1–21
Nasir M et al. (2018) Towards an unsupervised entrainment distance in conversational speech using deep neural networks. In: Proceedings of INTERSPEECH 2018, International Speech Communications Association, Hyderabad, India, 2–6 Sept 2018
Natale M (1975) Convergence of mean vocal intensity in dyadic communication as a function of social desirability. J Pers Soc Psychol 32:790–804. https://doi.org/10.1037/0022-3514.32.5.790
Paquette-Smith M, Schertz J, Johnson EK (2022) Comparing phonetic convergence in children and adults. Lang Speech 65(1):240–260. https://doi.org/10.1177/00238309211013864
Pardo JS (2006) On phonetic convergence during conversational interaction. J Acoust Soc Am 119(4):2382–2393. https://doi.org/10.1121/1.2178720
Pardo JS et al. (2012) Phonetic convergence in college roommates. J Phon 40(1):190–197. https://doi.org/10.1016/j.wocn.2011.10.001
Pardo JS et al. (2013) Phonetic convergence in shadowed speech: The relation between acoustic and perceptual measures. J Mem Lang 69(3):183–195. https://doi.org/10.1016/j.jml.2013.06.002
Pardo JS et al. (2017) Phonetic convergence across multiple measures and model talkers. Atten Percept Psychophys 79(2):637–659. https://doi.org/10.3758/s13414-016-1226-0
Pardo JS et al. (2018) A comparison of phonetic convergence in conversational interaction and speech shadowing. J Phon 69:1–11. https://doi.org/10.1016/j.wocn.2018.04.001
Patel AD (2011) Why would musical training benefit the neural encoding of speech? The OPERA hypothesis. Front Psychol 2:142. https://doi.org/10.3389/fpsyg.2011.00142
Patel R, Schell KW (2008) The influence of linguistic content on the Lombard effect. J Speech Lang Hear Res 51(1):209–220. https://doi.org/10.1044/1092-4388(2008/016)
Peirce J, Hirst R, MacAskill M (2022) Building experiments in PsychoPy. Sage, London
Phillips PC, Perron P (1988) Testing for a unit root in time series regression. Biometrika 75(2):335–346. https://doi.org/10.1093/biomet/75.2.335
Pickering MJ, Garrod S (2004) Toward a mechanistic psychology of dialogue. Behav Brain Sci 27(2):169–190. https://doi.org/10.1017/S0140525X04000056
R Core Team (2024) R: A language and environment for statistical computing. https://www.R-project.org/. Accessed 10 Oct. 2024
Rahimi Z et al. (2017) Entrainment in multi-party spoken dialogues at multiple linguistic levels. In: Proceedings of INTERSPEECH 2017, International Speech Communications Association, Stockholm, Sweden, 20–14 Aug 2017
Reichel UD, Beňuš ST, Mády K (2018) Entrainment profiles: Comparison by gender, role, and feature set. Speech Commun 100(1):46–57. https://doi.org/10.1016/j.specom.2018.04.009
Silva RP et al. (2021) Time series segmentation based on stationarity analysis to improve new samples prediction. Sensors 21(21):7333. https://doi.org/10.3390/s21217333
Soliz J, Giles H (2014) Relational and identity processes in communication: A contextual and meta-analytical review of communication accommodation theory. Ann Int Commun Assoc 38:107–144. https://doi.org/10.1080/23808985.2014.11679160
Strand EA (1999) Uncovering the role of gender stereotypes in speech perception. J Lang Soc Psychol 18(1):86–100. https://doi.org/10.1177/0261927X99018001006
Street RL (1984) Speech convergence and speech evaluation in fact-finding interviews. Hum Commun Res 11(2):139–169. https://doi.org/10.1111/J.1468-2958.1984.TB00043.X
Šturm P, Skarnitzl R, Nechanský T (2021) Prosodic accommodation in face-to-face and telephone dialogues. In: Proceedings of INTERSPEECH 2021, International Speech Communications Association, Brno, Czechia, Aug 30–Sept 3, 2021

Sun Y, Ding H (2023) Speech entrainment in Chinese story-style talk shows: The interaction between gender and role. In: Proceedings of INTERSPEECH 2023, International Speech Communication Association, Dublin, Ireland, 20–24 Aug 2023
Talkin D (2015) REAPER: Robust epoch and pitch estimator. https://github.com/google/REAPER. Accessed 5 Apr. 2023
Tannen D (1990) You just don't understand: Women and men in conversation. Morrow, New York
Ulbrich C (2021) Phonetic accommodation on the segmental and the suprasegmental level of speech in native–non-native collaborative tasks. Lang Speech PaPE 2019(Special Issue):1–27. https://doi.org/10.1177/00238309211050094
Ward A, Litman DJ (2007) Automatically measuring lexical and acoustic/prosodic convergence in tutorial dialog corpora. In: Proceedings of the SLaTE Workshop on Speech and Language Technology in Education, Farmington, PA, USA, 1–3, Oct. 2007
Weise A et al. (2019) Individual differences in acoustic-prosodic entrainment in spoken dialogue. Speech Commun 115:78–87. https://doi.org/10.1016/j.specom.2019.10.007
Weizman E (2006) Roles and identities in news interviews: The Israeli context. J Pragmat 38(2):154–179. https://doi.org/10.1016/j.pragma.2005.06.018
Wynn CJ, Barrett TS, Borrie SA (2022) Rhythm perception, speaking rate entrainment, and conversational quality: A mediated model. J Speech Lang Hear Res 65(6):2187–2203. https://doi.org/10.1044/2022_JSLHR-21-00293
Wynn CJ, Borrie SA (2022) Classifying conversational entrainment of speech behavior: An expanded framework and review. J Phon 94:101173. https://doi.org/10.1016/j.wocn.2022.101173
Xia Z, Levitan R, Hirschberg J (2014) Prosodic entrainment in Mandarin Chinese and English: A cross-linguistic comparison. In: Proceedings of Speech Prosody 2014, International Speech Communications Association, Dublin, Ireland, 20–23 May 2014
Xia Z, Ma Q (2019) Prosodic entrainment in Mandarin Chinese conversations: An experimental study. Tongji University Press, Shanghai, China
Xu Y (1994) Production and perception of coarticulated tones. J Acoust Soc Am 95(4):2240–2253. https://doi.org/10.1121/1.408684
Xu Y (2013) 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, 1 Aug 2013
Xu Y, Wang QE (2001) Pitch targets and their realization: Evidence from Mandarin Chinese. Speech Commun 33(4):319–337. https://doi.org/10.1016/S0167-6393(00)00063-7
Yip MJ (1980) The tonal phonology of Chinese. Dissertation, Massachusetts Institute of Technology

Acknowledgements

This work was supported by the major programme of the National Social Science Foundation of China (Grant No. 18ZDA293); the Supervisor Academic Guidance Programme of Shanghai International Studies University (Grant No. 2023DSYL001); and the Postgraduate Research & Innovation Programme of Institute of Language Sciences, Shanghai International Studies University.

Author contributions

Yanting Sun: Conceptualisation, Methodology, Software, Data curation, Visualisation, Formal analysis, Writing—original draft, Funding acquisition, Investigation. Hongwei Ding: Validation, Writing—review & editing, Resources, Supervision, Project administration, Funding acquisition.

Competing interests

The authors declare no competing interests.

Ethical approval

The study was conducted in accordance with the Declaration of Helsinki, and approved by the Ethics Committee of the Institute of Language Sciences of Shanghai International Studies University (No. 20230308005 on 8 March 2023).

Informed consent

Informed consent was obtained from all participants involved in this study prior to the commencement of the perceptual judgement experiment. Participants, all of whom were undergraduate students from Shanghai International Studies University, were provided with detailed information regarding the study's objectives, procedures, and their role in the research. Consent was obtained in written form between 14 and 15 March 2023 by the principal investigator. The scope of the consent included agreement to participate in the perceptual judgement experiment, the collection and use of their data for research purposes, and permission to publish the findings. Participants were informed that their data would remain confidential, and they were given the right to withdraw from the study at any time without any consequence. All participants voluntarily provided consent before engaging in the study.

Additional information

Correspondence and requests for materials should be addressed to Hongwei Ding.