社会心理语言学视域下
言者个体与群体身份的编码和解码

摘要：言语交流中，听者如何快速有效地感知言者的身份和个性是社会心理语言学的重要问题。关注言者间身份变异解码的传统研究发现听者区分言者间身份的正确率受听者音系知识及言者基频和声道长度的影响。新近研究发现，言者会因交际意图变化而调整发声策略（语言结构、语言风格和发声生理基础），听者能通过适应言者内部的变异进而识别言者身份。本文回顾了音系规则对身份编码的特殊制约，梳理了底层声学参数如何表征言者间及内部身份变异、进而影响言者身份感知；引入了内/外群体概念，探讨了言者在群体身份渗透意图下会采用不同发声策略这一现象如何支持交际调节理论；基于以上提出了言语互动场景下的言者身份编码及解码模型，并展望了三个研究方向。

关键词：言者身份；嗓音表情；交际意图；社会分组；社会心理语言学

A Social Psycholinguistics Perspective: Encoding and Decoding Mechanisms for Speakers' Individual and Group Identities

Wenjun Chen,[1] Yanbin Hu[1], Xiaoming Jiang,[1*]

[1] *Institute of Linguistics, Shanghai International Studies University, Shanghai 201620, China*

Abstract: How listeners quickly and effectively perceive speakers' identity and personality in verbal communication remains a widely researched topic for social psycholinguistics. Traditional research focusing on the perception of between-speaker identity variation reported that the correct rate for between-speaker differentiation is subject to listeners' phonological knowledge and speakers' Fundamental Frequency (F0) and Vocal Tract Length (VTL). Recent research has found that speakers modulate their vocalisation strategies (language structure, language style, and physiological basis of vocalisation) according to their changing communicative intentions, whereas listeners could adapt to within-speaker variations and recognise speakers' identities. This article reviews the unique constraints on speaker identity encoding imposed by phonological rules and unpacks how underlying acoustic parameters characterise within- and between- speaker identity variations that influence speaker identity perception. It further introduces the concept of in-/out-group and explores how the phenomenon where speakers would adopt varied vocalisation strategies when motivated by group identity permutation intentions support the Communication Accommodation Theory (CAT). Based on such, it proposes Speaker Identity Encoding and Decoding Model for Verbal Interaction Scenarios and calls for future

* The corresponding author is to be contacted at xiaoming.jiang@shisu.edu.cn.

research's attention in three directions.

Keywords: speaker identification; vocal expression; communication intention; social grouping; social psycholinguistics

1. Introduction

A horse looks at four hooves, and a person looks at four faces. In Dream of the Red Chamber, Lin Daiyu can quickly perceive Wang's domineering personality and her prominent position in the Jia household through Wang Xifeng's voice precisely by virtue of her 'phonetic phase'. In verbal communication, the human voice not only conveys linguistic information but also contains information about the identity and emotions of the speaker (Belin et al. 2004). The listener not only hears from the voice who the person is but also forms a general impression of who the person is. The human voice, like the human face, carries identity information and is also referred to as the 'auditory face' (Schirmer 2018). The identity of the speaker, which includes information such as gender, age, and body size (Campanella & Belin 2007), is encoded by a combination of speech signals based on fundamental frequency and vocal tract length (Frühholz & Schweinberger 2021; Lavan et al. 2019c), which the listener decodes mainly by the right anterior superior temporal sulcus (RSTS) and right anterior superior temporal sulcus (right STS) (Formisano et al. 2008). Speaker identity information shares speech signals such as the fundamental frequency with linguistic information in speech and even accent stress to indicate pragmatic purposes (Frühholz & Schweinberger 2021; Tang et al. 2017), i.e., speaker identity changes continuously with the speech task of vocalisation. However, a large number of speaker identity studies have not considered the dynamic nature of speaker identity encoding and decoding in speech interaction scenarios, i.e., they have paid much less attention to the social interaction dimension than to the cognitive-psychological dimension (Shuang Dingfang & Zhang Lifei 2021), so this paper specifically explores the interaction between speaker identity and the encoding and decoding of linguistic information in the communicative interaction dimension of speech.

In the psychosocial view of language, verbal communication is a conscious speech activity, and the study of its specific speech patterns and psychosocial speech mechanisms requires the integration of the disciplines of sociolinguistics, psycholinguistics, and engineering linguistics (Wang Dechun & Sun Rujian 1992a, 1992b). Sok Dingfang (1992) cites the phenomenon that teachers in Dublin can infer the social status of poor students' families from linguistic cues in their speech (e.g., substandard pronunciation) and thus lower their evaluations of students (i.e., judgments of speaker group identity influence social interactions); introduces the concepts of causal attribution and the group The concepts of causal attribution and group distinctiveness are introduced to explain the phenomenon of linguistic convergence under Giles et al.'s (1991) adaptation theory. The main object of research in psychosocial linguistics. In his outlook, Sok Dingfang (1992) calls on the linguistic community to investigate the relationship between linguistic change, language structure, language style, and group language and social psychology. The study of speaker identity is, therefore, an interdisciplinary issue in the context of psychosocial linguistics, which requires a synthesis of evidence from the intersection of psycholinguistics, sociolinguistics, communicative science, experimental psychology, experimental pragmatics, and cognitive neuroscience. Research into the encoding and decoding of speaker identity will contribute to the understanding of issues such as the relationship between language and psychosocial aspects, the use of artificial intelligence for speech cloning, and language learning and cross-linguistic processing. This paper, therefore, explores the mechanisms of encoding and decoding individual and group identities of speakers in dynamic speech interactions within an interdisciplinary perspective on language.

A key motivation for this paper's focus on the relationship between linguistic rules (phonological and syntactic structure) and language style in spoken communication and listeners' decoding of speaker identity and subsequent decisions about social interaction schemes is that classical linguistic theory does not consider

how the variability of acoustic information in multimodal interaction scenarios, such as spoken communication, will affect listeners' perceptions of individual and group identity of speakers. For example, Chomsky argues that language users have an innate ability that allows them to produce and understand an infinite number of sentences; this ability allows them, as listeners and when they hear the same sentence, to understand it in the same way even if these listeners again have different backgrounds and experiences (Chomsky 2014). Speech act theory tasks that the speaker will adopt discourse to achieve a particular speech act (e.g., making requests and giving orders) and that listeners will understand the speaker's speech act based on the context of the communicative situation, at which point who the speaker is does not affect the listener's understanding of the speaker's speech act (Austin 1975). Grice argues that there is more implied than the literal meaning of the speaker and that listeners will use the implied meaning in the conversation to understand the sentence; this theory suggests that context and non-verbal information play a key role in understanding the discourse without emphasising the internal identity of the individual speaker (Grice 1975).

However, the above linguistic theories are centred on the view that the ontological structure of language (which does not take into account multimodal interaction) is such that the listener's understanding of the discourse does not change depending on who the speaker is, and the brain does not seem to process sentences differently depending on the identity of the speaker. However, psycholinguistic and neurolinguistic experiments under speech-based communication interaction have shown that listeners' processing of speakers' discourse is influenced by inter- and intra-individual differences in speaker identity. For example, indicators of neural activity in the brain are sensitive to conditions in which the use of the honorifics' you (您)' and 'you (你)' is violated when socially interacting between parties of different status in the social domain (Jiang et al. 2013). Similar findings have been found in the context of syntactic structure, for example, in studies of contrasting SOV and OSV

114 syntactic structures in German, where listeners expect speakers to speak simple SOV
115 sentences, but experimenters observe P600 when they hear speakers speaking difficult
116 OSV sentences (Kroczek & Gunter 2021). A similar P600 effect was observed for
117 whether individuals often expressed themselves sarcastically in reading experiments
118 under the pragmatic category (Regel et al. 2010). Thus, from the speaker's
119 perspective, linguistic rules consisting of phonological structure and linguistic style
120 (focusing on communication at the phonological level) influence the speaker's speech
121 output, with differences in output reflected in sophisticated acoustic analyses (e.g.,
122 intergroup differences in parameters such as fundamental frequency, vocal tract length
123 parameters, sound intensity, duration, jitter, and shimmer, which can be: confident,
124 neutral, sceptical "sense of knowing" driven speech rhythm differences (Jiang & Pell
125 2017)). Listeners, on the other hand, will be sensitive to acoustic variation in the
126 speech produced by the speaker. In particular, the fundamental frequency and vocal
127 tract length critically characterise identity differences within the individual speaker
128 and between the individual speaker and other individual speakers; thus, individuals are
129 sensitive to changes in speaker identity in speech. One piece of evidence for the above
130 inference comes from an EEG study exploring listeners' decoding of confidence levels
131 in speech produced by speakers of English with different accents, which found that
132 linguistic structure and speaker identity information, including phonological structure,
133 are processed at an early stage of ERP (Jiang et al. 2020). Therefore, this paper
134 focuses on the individual's understanding of the speaker's produced speech in the
135 context of phonological transmission of information, focusing on how listeners
136 interpret the identity of the speaker and thus influence the listener's vocalisation
137 strategies to engage in social interaction.

138 Much of the research on speaker identity encoding and decoding has focused on
139 the cognitive processing patterns involved in individual vocalisations under controlled
140 conditions (i.e., experimental stimuli recorded with neutral voice expressions).
141 Experiments on identity decoding based on vocal cues have focused on two paradigms

to explore listeners' recognition mechanisms for unfamiliar and familiar voices: speaker discrimination, in which listeners determine whether an unfamiliar speaker is the same person based on the two sentences they listen to; and speaker identification, a paradigm derived from the judicial practice of suspect identification, where the person listens to an array of speech and then calls on memory to point to the identity of the suspect (Frühholz & Belin 2018; Levi 2019). There are several related reviews in China: Wu Ke et al. (2020) introduced a dual-pathway model, a multi-stage model, and an integration model involved in human voice speech, emotion and identity processing from the perspective of neural mechanisms of perception; Zhou Aibao et al. (2021) distinguished the differences in damaged brain regions between patients with acquired and developmental vocal agnosia; Ming Lili and Hu Xueping (2021) compared normal sighted and blind people to process human brain mechanisms that differ in voice identity. In addition, Chen Zhongmin (2021) discussed the characteristics of speech perception and the anatomical and physiological mechanisms involved according to the anatomical and physiological configuration of the nervous system above the auditory organs, suggesting a physiological basis for the encoding of speaker identity.

Most of these reviews have explored the encoding and decoding of vocal identity based on individuals who are members of the same group and less on the differences in vocal identity and processing mechanisms between social groups. However, individuals also vocalise in more complex ways than just neutral vocal expressions. Individual vocalisations vary internally according to speaking style, environmental and social contexts (e.g., imitating others), stage of cognitive development, emotional, physiological and psychological states (Lavan et al., 2019b), i.e., the listener's perception of the "who" and "what kind of person" the speaker is from the voice is not constant. The results are not constant and can be influenced by these factors.

On the basis that language structure and language style, together with the physiological basis of vocalisation, influence vocal strategies, this paper reviews the

literature on individual and group identity encoding and decoding of speakers, proposes an integrated model of speaker identity encoding and decoding in speech interaction scenarios, and provides a research outlook based on this.

2. The physiological basis of individual speaker identity and linguistic-acoustic coding

Similar to the broad language faculty (FLB) proposed by Hauser, Chomsky, and Fitch Chomsky et al. (2022), the encoding and decoding of vocal identity is a universal ability of the species; and similar to the recursive nature of their narrow language faculty (FLN), the decoding and encoding of human speaker identity have become more complex as a result of language evolution. After determining language structure and language style at the speech planning stage, speakers execute a specific 'vocal strategy' by invoking the physiological basis of vocalisation, which results in the production of speech sounds carrying vocal identity. These include phonological structure (which can be brought about by accent or stylistic variant) and syntactic structure (e.g., speakers' preference for SOV or OSV structure in the experimental design (Kroczek & Gunter 2021)); and linguistic style, which includes gender-binary speaking styles, and specific, pragmatic choices (e.g., a tendency to speak sarcastically (Regel et al. (Regel et al. 2010)), etc.

2.1 Encoding and recognition of voice identity as a universal competence

The body size of living organisms is a key element of acoustic identity coding, for example, wolves howl to indicate territory to their mates when hunting or calving, and the perception of the size of conspecifics is common in the social organisation of many species (Harrington and Mech 1979). Reby and McComb (2003) analysed howl, body weight and reproductive success data from 24 male red deer and found that vocal tract length, based on resonance peak spacing, was positively correlated with body weight and that the maximum tract length corresponding to a normal state howl was positively correlated with reproductive success. Similar findings have been found in human societies. .ebesta (2019) et al. invited 84 heterosexual participants from Brazil

198 and 68 heterosexual participants from the Czech Republic to read short sentences and
199 sing aloud and report on socialised sexuality and found that shorter vocal tract length
200 in short speeches and longer vocal tract length in singing predicted female sexual
201 behaviour. This suggests that a species' judgement of identity is a universal ability.

202 The human ability to decode each other's identity from sounds preceded the
203 emergence of verbal communication: Polka et al. (2022) synthesised vowels/i/audio
204 from infants with widely spaced resonance peaks (F2-F1=3761) and adult females
205 with less spacing (F2-F1=2315) and found that infants with an average age of 220
206 days had a preference for vowels that mimicked the infant's vocal state. And this
207 ability was strongly related to language acquisition: infants who had been exposed
208 only to English for most of their mean age 136 days were presented by Fecher and
209 Johnson (2019) et al. with English, Polish and Spanish sentences recorded by four
210 bilingual (two English and Polish speakers; two English and Spanish speakers)
211 females; with gaze duration as the dependent variable, speaker (different/same) and
212 language (native/non-native) as the main fitting parameters, the results showed no
213 main effects for either speaker or language but an interaction between them,
214 suggesting that whether the stimulus was the infant's native language or not
215 moderated the infant's gaze duration when listening to audio from the same or
216 different speakers.

217 Studies based on other species and early language acquisition suggest that
218 humans have retained the ability to recognise each other from sound during evolution
219 and that this ability predates verbal communication, but that language is a uniquely
220 human phenomenon that complicates this ability compared to other species.

221 **2.2 Specificity in the encoding and decoding of individual speaker identity: the constraints**
222 **of linguistic phonological rules**

223 Listeners are able to integrate information about the identity of the speaker (i.e.,
224 social goal) and the content information in the discourse for communicative purposes

during the interaction, thus forming a linguistic goal (Kuhl 2011), and it is the decoding of linguistic goals by listeners that makes human speaker recognition different from recognition of conspecifics by ordinary animals. Perrachione et al. (2011) found that impairments in the phonological rules of English (as characterised by scores on the Comprehensive Test of Phonological Processing, or CTOPP's Phonological Memory and Phonological Awareness subtests) The dyslexic group, as judged by clinical diagnosis, was found to be equally accurate in recognising the identity of speakers coded in native English and in completely unfamiliar Chinese, and both were much less accurate than healthy controls. In particular, from a phonological perspective, individuals who are completely unaware of a language are less likely to accurately hear and produce the sounds and sound patterns of that language, thus lacking knowledge of the phonological rules of the particular language (Goldsmith et al., 2014). Individuals with English monolingual dyslexia lack knowledge of the phonological rules of Chinese, and their dyslexia impairs the phonological rules of English, resulting in knowledge of the phonological rules of their native language at an unfamiliar language level, and this impairment makes their speaker identity recognition accuracy in the native language condition similar to that in the unfamiliar language condition. Human speaker identity decoding is highly dependent on the listener's knowledge of phonological rules, and it is the greater knowledge of the native language that leads to the 'language familiarity effect': even when listening to sentences played backwards without semantic fluency (Fleming et al. 2014. Goggin et al. 1991), monolingual listeners will identify the speaker more accurately in the native language condition (Perrachione and

Wong 2007). Notably, Orena et al. (2015) found that English monolingual adults in Montreal, Canada, were able to learn and identify the speaker identity of French speakers more quickly and accurately than English monolinguals in Connecticut, USA, suggesting that being exposed to usage scenarios with specific phonological rules can also contribute to the emergence of a language familiarity effect.

The above evidence suggests that the integrated processing of identity and phonological information in speaker discourse by human listeners may have processing mechanisms that differ from those used by ordinary animals to identify their counterparts.

2.3 Specificity in the encoding and decoding of individual speaker identity: the binding of syntactic structure and linguistic style

Kroczek and Gunter (2021) trained subjects to be exposed to a design in which a particular speaker spoke 70% OSV and 30% SOV sentences or vice versa, whereby subjects built up an image of the particular speaker as an OSV speaker or an OSV speaker; when tested subjects heard the ERP component of the SOV speaker speaking OSV sentences and found a mid-posterior P600 - characterising syntactic re analysis or repair. A similar experimental manipulation also revealed that listeners could anticipate the high/low syntactic attachment style of the speaker (Kamide 2012).

More research has also shown that listeners bind speaker identity to a particular linguistic style. For example, Regel et al. (2010) used a similar design to the above to allow readers to form bindings and expectations about the tendency of two personal names to use irony, and a similar P600 was found when readers read ironic discourse by non-ironic stylists.

Notably, Walker and Perry (2022) manipulated male- and female-specific rhymes (a woman using habitual rhymes vs imitating male rhymes) and language style (masculine/feminine diction scenarios), and ERPs when subjects heard the woman speaking in male rhymes induced greater N400s characterising semantic inconsistencies or unanticipated stimuli, and ERP results also reported greater rhyme and The ERP results also reported an interaction between rhyme and language style, i.e., greater negative waves were produced in conditions consistent with female speaker identity and female rhyme and female speech style than in other inconsistent conditions. This study suggests that the social category (male-female dichotomous

rhymes) and the linguistic ontology category (gender-specific vocabulary) jointly influence the outcome of speech output and that the listener's processing of both categories is linked to the listener's decoding of the speaker's identity.

It can be deduced from the above that the ability of individuals to learn the tendency to use syntactic structures and language styles of others in a strictly controlled laboratory is closely related to the individual's long-term practice of combining features of others' language use and others' identities in natural interaction scenarios.

2.4 Encoding and decoding of individual identities based on inter-speaker variation parameters

Language is a product of the advanced evolution of the human species, which has made the encoding of speaker identity more complex compared to other species. Winters et al. (2008) recruited English monolingual subjects and presented them with German meta-auxiliary-meta-words recorded by English-German bilinguals during the familiarisation phase, i.e. listeners were required to associate the identity behind the German word with the corresponding name; after eight rounds of familiarisation-refamiliarisation-recognition, the After eight rounds of familiarisation-re-familiarisation-recognition, the subjects were able to associate the ten identities behind the German audio with the corresponding names; finally, in the test and generalisation phase they were presented with another set of English words recorded by the bilinguals and asked to indicate the identity of the speaker behind the audio; it was found that after listening to the German words to familiarise themselves with the identity of the speaker, the listeners were able to discriminate the identity behind the English vocalisation of the bilingual speaker well above the chance level. This suggests that there are acoustic cues independent of phonological structure that consistently encode vocal identity, namely the fundamental frequency (Matsumoto et al. 1973; Xu et al. 2013) and the resonance peak spacing that characterises vocal tract length (Ghazanfar and Rendall 2008. Johnson 2020), with fundamental frequency and

vocal tract length interactively influencing timbre and encoding speaker identity (von Kriegstein et al. 2006).

With regard to the physiological basis of vocalisation, the airflow during vocalisation is transmitted from the respiratory system (lungs) through the trachea to the vocal system (vocal folds) and then through the larynx to the articulatory system (the tuning area consisting of the oral, pharyngeal and nasal cavities, i.e. the vocal tract), which ultimately produces speech (Nakagawa et al. 1995). Firstly, the frequency of vocal fold vibrations is characterised as the auditorily perceptible pitch or fundamental frequency. The fundamental frequency is the inter-individual glottal-pulse rate, an acoustic representation that differs due to differences in the construction of the vocal folds, and although it is not strongly related to individual size, there are clear gender differences: adult males have vocal folds that are approximately 60% longer and wider and thicker than those of females, resulting in generally lower glottal pulses in males than in females, so that male fundamental frequencies are generally one octave lower than those of females (Titze 1989; Künkel et al. (Titze 1989; Künzel 1989). Secondly, the distance between resonance peaks is statistically related to vocal tract length; the smaller the spacing between resonance peaks, the longer the vocal tract length, and vocal tract length is directly related to individual body size (Lee et al. 1999; Johnson 2020), with individual vocal tract lengths being approximately 8 cm at birth and ranging from 13 to 20 cm in adults (Lammert and Narayanan 2015). Also, fundamental frequency and vocal tract length interact to influence speaker identity coding. Voices from vocal tracts of equal length will thus sound larger in size when the vocal tract pulse rate is lower and smaller if the vocal tract pulse rate is lower (Smith and Patterson 2005).

In addition to spectral information such as fundamental frequency and vocal tract length based on resonance peaks, there are additional parameters that characterise speaker identity differences.

335 The following parameters were analysed: root means square energy (RMS
336 energy), which reflects temporal information; spectral centroid and spectral roll-off,
337 which reflect spectral information; shape mel frequency cepstrum coefficients
338 (MFCCs), which reflect the spectral envelope; and spectral entropy, which reflects the
339 amount of information in the signal. Spectral entropy, probability density function
340 (PDF entropy), permutation entropy, and singular value decomposition (SVD
341 entropy). No significant differences were found between races for the above
342 parameters; however, significant differences were found between men and women
343 (regardless of race) for several indicators.

344 In general, inter-individual differences in fundamental frequency and vocal tract
345 length parameters due to physiological structure primarily characterise the inter-
346 speaker identity, and a wider range of temporal, spectral, cepstral, and entropy-related
347 parameters may also reflect gender group identity.

348 **2.5 Encoding and decoding of individual identities based on intra-speaker variation**
349 **parameters**

350 The field of vocal identity has mostly used vocalisations in speakers' neutral
351 voices as stimuli to explore listeners' recognition mechanisms of vocal identity
352 (Perrachione et al. 2011; Fleming et al. 2014). However, the language used and the
353 requirements of the paralinguistic information expressed in the context of the scene
354 (e.g., the speaker's mood when confident and doubtful) allow for a high degree of
355 intra-speaker variability. For example, Voigt et al. (2016) analysed recordings of 25
356 German-French and 20 German-Italian bilinguals during interviews on light topics and
357 found that German-French bilingual women used higher mean basal frequencies when
358 speaking French, and German-Italian bilingual women spoke Italian at lower basal
359 frequencies. In an analysis of declarative showing different levels of speakers' 'sense
360 of knowing', it was found that the basal frequency of confidence was higher when
361 looking at sentence-initial and mid-sentence components compared to non-confidence,
362 yet the basal frequency of non-confidence was higher when looking at sentence-final

363 components (Jiang and Pell 2017), which supports the idea that speaker identity is
364 only coded through acoustic parameters that characterise inter-speaker variation. This
365 challenges the idea that speaker identity is encoded solely through acoustic parameters
366 that characterise inter-speaker variation, i.e. listeners rely on more complex cues when
367 decoding speaker identity.

368 Research has found that listeners have the ability to adapt to internal variations in
369 speaker identity. Different speakers have different prototype-based vocal identities,
370 which are distributed in a multidimensional sound space (Latinus and Belin 2011).
371 Based on this, Lavan et al. (2019c) manipulated identities on a two-dimensional space
372 with fundamental frequency and vocal tract length by adjusting semitones, creating
373 four source-independent vocal identities by shifting the fundamental frequency
374 up/down by 1.6 semitones and the vocal tract length left/right by 2.36 semitones in a
375 space with vocal tract length as the horizontal axis and fundamental frequency as the
376 vertical axis (the voice in the lower left corner was excluded due to unnatural
377 excluded). Around the midpoints of the three new sound prototypes, 16 new inner
378 perimeters were created closer to the prototypes and 18 outer perimeters further away,
379 each within a range of 2.25 semitones to the left and right of the channel length and
380 3.6 semitones to the fundamental frequency; the inner/outer perimeters reflect the
381 internal variation of the three sounds. The listeners only heard the peripheral point
382 during the training phase, but they reported hearing the peripheral point during the test
383 phase when judging the sound as “old/new”. This suggests that listeners have a
384 prototypical awareness of the speaker’s voice identity and can adapt to internal
385 variations in speaker identity.

386 However, listeners’ ability to adapt to internal variations in speaker identity is
387 limited. Lavan et al. (2019a) normalised 1.2 to 4-second-long emotional vocal spikes
388 from the two male leads of the American drama series Desperado to 0.400 Pa and,
389 after low-pass filtering at 10 kHz, asked listeners to drag and drop classify the speaker
390 identities behind the synthesised manipulated audio; it was found that even Xu and

Armony (2021) prepared four sentences from each of the 12 speakers (2 emotional rhythms: fear/neutral* 2 semantic contents) and presented them with one sentence each in the neutral/fear rhythm of 6 of the speakers during the listener familiarisation phase, and then in the During the test phase, 48 sentences from all 12 recorders were played and listeners were asked to determine whether the identity of the voice was present or not. It was found that when listening to sentences with the same content, subjects were generally more than 80% accurate if the rhymes were the same but only around the chance level if the rhymes did not match. Thus, listeners' adaptation to intra-speaker variation was limited to a specific threshold.

The above suggests that speaker identity varies internally according to the language used or the specific situation and that the listener is able to adapt to such variation within a certain threshold, normalising the identity of a speaker whose voice identity has changed to the same person.

3. Speaker group identity penetration intentions modulate vocal strategies

Social psychology explains the division of social groups in terms of archetypal theory, whereby a fuzzy collection of individual-related attributes such as attitudes, behaviours, and customs form a prototypical conception of human groups in the mental representations of communicating individuals, and the attributes represented by this prototype maximise group solidity and lead to stereotyping; among other things, language and speech style is one of the identity symbols of the group an individual is a member of (Hogg 2016). (Hogg 2016), whereby people classify objects of social interaction as in-group/out-group members (Jiang et al. 2020), and social group divisions are permeable, i.e. members can change their identity representations (Hogg 2016). The following section reviews how speaker group identities are encoded and how individuals can perform group permeability through moderated vocalisation before proposing an integrative framework for decoding speaker identity representations based on a group interaction perspective.

3.1 Decoding the identity of the speaker group

The language used by the speaker is one of the criteria used by the listener to classify the in/out-group. Kenyans argue that the use of Swahili and Giriama differently defines the self, rights, entitlements, and religion of language speakers (Kinzler 2021). Empirical research on infants provides evidence of in-group preferences among native speakers. For example, 12-month-old infants are more likely to take food handed to them by native in-group speakers (Shutts et al. 2009); 10-month-old native English-speaking infants prefer toys shown to them by English in-group speakers, and English/French monolingual children around 2.5 years of age are more likely to hand objects to in-group native speakers for playful interaction (Kinzler et al. 2012). Begus et al. (2016) presented infants at around 11 months of age with videos of their native in-group (English) and out-group (Spanish) female speakers pointing to objects unfamiliar to the infant to teach nouns and observed infants' activity in the 3-5 Hz theta band (neural oscillations in the theta band are commonly used to characterise information processing and learning, and in adults, the theta band is 4-8 Hz) and found that infants had more strongly active theta oscillations in the in-group speaker condition. This suggests that listeners are sensitive to the linguistic structure of a particular language and that this leads to a judgement of the speaker's group identity and, thus, to differentiated social interaction.

Rubin (1992) showed North American university students audio lectures in a standard Southern American accent accompanied by pictures of Asian or white faces and found that when a standard Southern American accent was associated with an Asian face, students perceived the speaker to have a heavier non-standard accent, poorer teaching qualifications, and more difficult to understand lectures. Jiang et al. (2020) recruited 44 native English-speaking listeners from Quebec, Canada, who had considerable French language skills and knowledge of the Australian English accent, and the subjects were given credibility ratings after listening to audio recorded by a

447 Canadian English accent, an Australian English accent, and a person with a Quebec
448 French accent, with either a confident or sceptical voice expression, and found that in
449 the confidence condition, the Canadian Tamagawa et al. (2011) provided evidence
450 beyond real speakers: subjects using a New Zealand accent listened to audio based on
451 British, American, and New Zealand accents trained on the same blood pressure
452 monitor. After introducing synthetic speech from a robot with the same blood pressure
453 monitor, it was concluded that the American accent was more machine-like than the
454 synthetic voice with the New Zealand accent and performed worse than the robot with
455 the New Zealand synthetic accent. The robot's mouth in this experiment was part of
456 the speech structure's representation of a different phonological structure. The three
457 experiments above suggest that speakers' choice of phonological structure critically
458 encodes their group identity as perceived by listeners; and that listeners will vary their
459 interaction decisions based on their perception of different speaker identities.

460 Speech styles based on gender dichotomies also mark group identity. Men's
461 speech styles are typically characterised by the use of slang (vulgarity), more blunt
462 speech, lower voices, aggressiveness, and appearing more authoritative, whereas
463 women's are characterised by greater variation in rate and fundamental frequency,
464 gentler speech, openness, self-disclosure, and appearing more emotional (Giles et al.
465 1983; Hogg 1985). The "big two" model proposed by Martin et al. (2021) suggests
466 that individuals' judgments and evaluations of others' traits follow two dimensions
467 that overlap significantly with gender roles: a. agency/masculinity, which is assertive,
468 competitive, dominant, independent, self-interested, and goal-seeking; and b.
469 community/femininity, which is nurturing, warm, expressive, concerned about others,
470 and social orientation. Of these, masculinity is strongly associated with perceived
471 'competence' by others; for example, masculine facial features make individuals
472 appear more competent (Oh et al. 2019), women with lower, i.e. more masculine,
473 voices are perceived as more dominant, while feminine voices are associated with
474 naivety and sexual immaturity (Borkowska and Pawlowski 2011). Thus, listeners

distinguish their group identity based on differences in male and female speech styles and associate this division with specific stereotypical images.

The specific type of language used in verbal communication, the accent displayed, and the degree of masculinity or femininity encode the speaker's group identity, on the basis of which the listener identifies the interacting party as an in-group or out-group member and adapts the interaction scheme differently. It is worth noting that a large body of existing literature has focused on how listeners' perception of speakers' accents affects social interactions, but most of this has been based on verbal communication between real people, with little research focusing on the recent emergence of artificially intelligent human voice cloning and speech synthesis technologies.

3.2 In/out-group penetration mechanisms for speaker identity coding

The group identity of speakers can be modified by their intention-based adjustment of vocal strategies. On the one hand, the theory of communicative accommodation (CAT) suggests that speakers converge with each other in terms of accent, speed, volume, pauses and content use in order to reduce social distance, promote mutual understanding and increase communicative efficiency (Coupland et al. 1988; Bernhold and Giles 2020). Such rhyme-level convergence has also been found in natural conversational contexts in Chinese (Xia Zhihua & Ma Qiowu 2019). An example of convergence in a typical social context is the adoption of each other's dominant speech patterns by parties of higher and lower status levels (Shuang Dingfang 1992). Another case is that adults will use child-oriented speech to communicate with infants, characterised by richer base frequency variation (Stern et al. 1982), higher base frequency and lengthened final syllables (Albin and Echols 1996), more repetition (Hills 2013), and shorter utterances (Soderstrom et al. 2008). Sorokowski et al. (2019) recorded audio of 27 male and 24 female scientists working at a university talking about everyday topics (asking for directions) and the authoritative topic "how to become a scientist and is it worth it" and found that both

503 males and female speakers had lower fundamental frequencies when giving
504 professional advice and that women (33Hz) made more frequent comments than
505 Harrington et al. (2000) investigated the vowels in the audio of Queen Elizabeth II's
506 speeches from the 1950s to the 1980s and found a tendency to move towards a
507 younger demographic and a commoner approach to her vocalisation. These examples
508 suggest that speakers strategically change their choice of speech style in order to
509 appear friendly or more professional and that such changes are reflected not only in
510 the level of effort put into modulating the vocal base but also in the way the stylistic
511 variant is calculated and executed.

512 On the other hand, Pisanski et al. (2021) suggest that vocalic complexity in
513 human-voiced speech may have its origins in a common phenomenon in the animal
514 kingdom: species lower vocal tract resonance (i.e., lower resonance peaks) to achieve
515 vocal body exaggeration, a phenomenon that exists in humans with complex language
516 systems and is also common in other groups that do not have human-like It is also
517 common in other animal groups that do not have human-like language systems. In
518 human speech scenarios, vocal tract lengths are longer when speakers are aggressive
519 compared to neutral vocalisations (Pisanski et al. 2022), vocal tract lengths are shorter
520 when speakers are happy than when they are angry, sad or neutral (Kim et al. 2020),
521 and speakers produce lower fundamental frequencies when they are confident
522 compared to unconfident recordings (Jiang and Pell 2017), such acoustic parameters
523 suggest a physiologically grounded effort to modulate the lengthening/shortening of
524 the vocal tract as speakers encode their identity; and these changes in vocal strategies
525 will directly affect the listener's perception of the speaker's identity. That is, humans
526 have retained the ability to alter vocal body shape by lengthening/shortening the vocal
527 tract, raising/lowering the fundamental frequency, and subconsciously performing
528 vocal body shape changes in specific speech scenarios (e.g., speakers make
529 themselves sound larger when aggressive) after they have evolved a language system.
530 From the above, it is clear that shorter-duration paralinguistic messaging may be

related to longer-duration human speech structure and the lineage of speech styles, similar to the relationship between broad language faculties (FLB) and narrow language faculties (FLN).

This section shows that speakers follow specific linguistic rules to modify their speech production to make themselves sound like part of a particular group for specific communicative purposes, a mechanism that may have evolutionary significance due to the exaggeration of body size prevalent in the animal kingdom. However, the finer points of such vocal modulation and whether there is cross-cultural consistency in language rules (given linguistic diversity) remain to be answered. At the same time, research into how language rules are acquired or used by aberrant groups to encode and decode speaker identity would help to understand the relationship between identity, language and paralinguistic information in the voice.

3.3 A theoretical framework for the encoding and decoding of speaker identity in speech social interaction scenarios

The above review shows that speakers adjust their vocal strategies in response to communicative intentions in order to influence the impressions that listeners receive of who and what the speaker is. In interactive scenarios, the listener takes the turn and adjusts the speech output based on the integrated information about the speaker's identity and language. Therefore, this paper integrates (1) the human voice processing model proposed by Belin et al. (2004), which states that three neural pathways in the human brain are activated separately to refine speech, emotion and identity information after recognising a voice as a human voice; (2) the human speech communication cycle proposed by Cummings (2015) from the perspective of the listener-speaker discourse wheel cycle; (3) the human speech communication cycle proposed by Jiang et al. (2020) proposed a cognitive processing model of voice expressions from the perspective of voice identity and emotion processing time course, in which listeners process vocal information structure (vocal structure) in speaker speech, including voice identity information and voice speech information

559 (which includes linguistic structure) in the early stages of ERP; finally, a framework
560 for encoding and decoding speaker identity in social interaction scenarios was
561 proposed (Figure 1).

562 The two basic elements of the framework are (1) emotional information, which
563 includes basic emotions such as surprise, happiness, anger, sadness, fear, disgust, and
564 neutrality, as well as assertive voice signals that express a “sense of knowing” (Jiang
565 2020); and (2) identity information, which is characterised by variables such as
566 gender, age, education, attractiveness, ability, and ethnicity (Frühholz and Belin
567 2018). (Frühholz and Belin 2018).

568 At the beginning of the discourse round, the speaker completes linguistic
569 encoding by selecting the listener-specific phonological and syntactic structures driven
570 by communicative intentions that permeate the identity of the other group (e.g.,
571 whether to adopt a typically masculine speech style); during the speech motor
572 encoding phase, the speaker completes a plan for how to invoke the vocal base of the
573 tongue, lips, and vocal folds based on linguistic structural information (e.g., changing
574 the degree of tongue roll for [r] sounds (Labov 2006)) and rules for expressing
575 paralinguistic information (e.g., whether to lengthen the vocal tract length, lowering
576 the fundamental frequency to appear more confident (Jiang and Pell 2017)) completes
577 the planning of how to invoke the vocal base of the tongue, lips and vocal folds; in the
578 speech motor execution phase, neural signals are sent from the speaker’s brain to
579 control the anatomical base of vocalisation to complete speech production and
580 transmit sound waves.

581 The listener’s auditory system converts mechanical wave vibrations into neural
582 signals that are transmitted to the auditory centre, completing the reception of auditory
583 information. During the speech perception phase, the listener performs a structural
584 analysis of the voice at around 100 ms, simultaneously processing voice identity,
585 emotional and content information (understanding the syntactic information structure

that represents the function of the sentence); at around 200 ms, the listener performs voice importance detection (comparing the tone of the speaker, similarity to the listener's own accent, etc.) to determine the amount of attention to be allocated; at 250 ms After 250 ms, the listener enters a language comprehension phase, where he/she reconfirms/disambiguates ambiguous semantics, makes pragmatic inferences based on identity information, and integrates identity information into the context. The speaker then takes over and strategically vocalises based on intention, and the cycle repeats itself.

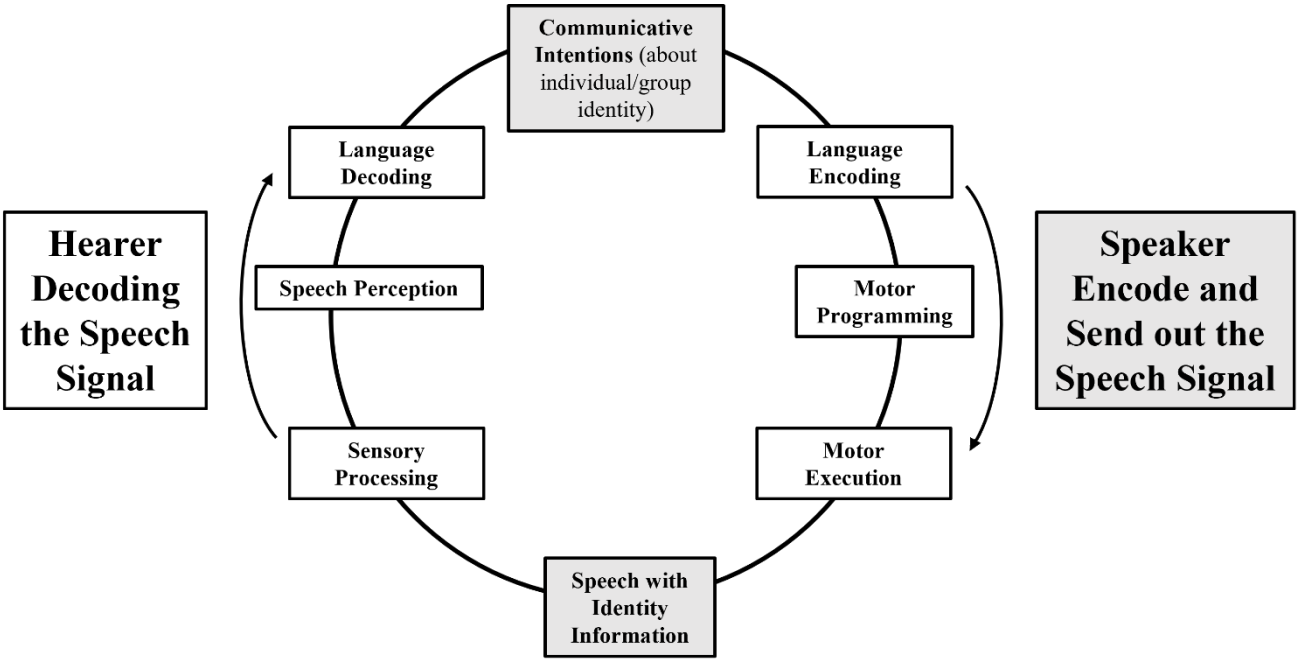

Figure 1. Speaker Identity Encoding and Decoding Model for Verbal Interaction Scenarios

4. Research outlook

Based on scholarly results on the encoding and decoding of individual and group identities of speakers, future research could explore (1) what indicators characterise speakers' efforts to adapt vocal movements for group penetration purposes; (2) the brain mechanisms of listeners' processing of artificially intelligent cloned human voices and the moderating effect of affective-rhythmic cues on cross-group vocal

decoding; and (3) the internal variation in encoding and decoding speaker identities in speakers with language rule deficits.

First, how individuals regulate their vocal strategies based on linguistic rules is yet to be explored. Social norms and behavioural habits are acquired by individuals through verbal and non-verbal social interactions, which have cross-cultural commonalities, and thus interaction participants can understand the cultural symbols of the speaker across cultural contexts (Jiang 2020). So, can speech outputs guide by specific discourse functions, such as the vocalisation of social attitudes like assertiveness, dominance and conformity, be compared across cultures through acoustic parameters of speaker audio and imaging techniques for vocal modality? That is, can assertive vocalisations be observed to lengthen the vocal tract while suspicions reduce the tract length, and is there intercultural consistency in such patterns? Possible research tools are acoustic parametric analysis, physiological motion measurement techniques of the vocal tract, and magnetic resonance imaging techniques.

Second, the mechanisms of human cognitive processing of cloned human voices need to be urgently explored. First, technology has made it possible to clone a model of a real person's voice based on seconds-long audio and use it to falsify an individual's speaker identity (Jia et al. 2018), so how will listeners define the in/out-group identity of the cloned voice and make decisions about social interactions when perceiving a real person's sound source and its cloned counterpart? For example, Pernet et al. (2015) found that three patches in temporal lobe sound areas were selectively sensitive to human voices, and Zhang et al. (2021) found by cortical EEG ECoG that specific electrode sites in the left anterior temporal lobe of epileptic patients responded only to native human voices, and, more notably, Di Cesare et al. (2022) presented listeners with words that conveyed social intention After the word "hello" was presented to listeners, the real voice specifically activated the dorsal-central insula region of the listener compared to the synthetic voice of a neutral voice. So are there specific neural correlates that characterise individuals who process the

real person and their cloned counterparts differently? Second, speech synthesis techniques can yield audio rich in expressive vocalisations such as crying, laughing, and yawning (Kharitonov et al. 2022) or even allow two synthetic voice models to engage in spontaneous but real-time small talk with natural overlaps and pauses (Kreuk et al. 2021); if the above techniques are combined with cloned human voices, i.e., the cloned voices become more "anthropomorphic", will listeners' group categorisation of cloned voices still be altered? Future research could combine electrophysiological and magnetic resonance imaging with exploring listeners' differentiated perception of speaker identity in different conditions in terms of time course and spatial dimensions. At the same time, the classical adaptation paradigm, based on the theory that listener-specific neuronal responses diminish with increasing exposure to the same type of stimulus and become more intense if the stimulus features change, could be used to further explore how speech rhythm modulates human listeners' perception of reality and cloned voices (Belin and Zatorre 2003; Grill-Spector et al. 2006). Again, how will vocal recognition technology respond to the speaker identity crisis posed by cloned human voices? Just as upgrading the vocal length normalisation algorithm may improve the accuracy of vocal identity products (Tan 2021), it is unclear how the broader spectrum, cepstrum and other parameters mentioned earlier play a role in altering recognition accuracy.

Third, the mechanisms of speaker output and listener perception in people with language rule deficits also need to be investigated. First, in order for transgender people to achieve vocal penetration into gender groups opposed to their biological sex, they need to acquire correspondence rules through (supplemented by visual) oral resonance speech therapy or undergo cricothyrotomy (Neumann and Welzel 2004; Hardy et al. 2016; Dahl and Mahler 2020), interventions that affect speech representation and application of rules, and how will this inform the implementation of correctional programmes for groups with crisis gender identity? Secondly, people on the autism spectrum with persistent impairments in social communication/interaction

have abnormalities in their 'social brain', and these patients often show impairments in the use of phonological rules, which are associated with their inferior frontal gyrus (IFG), superior temporal gyrus (STG), and the use of phonological rules. This is often associated with over-activation of the inferior frontal gyrus (IFG), superior temporal gyrus (STG) and amygdala (Peng et al. 2020). However, the superior temporal gyrus (STG) is involved in unfamiliar voice identity processing, and the inferior frontal gyrus (IFG) is involved in familiar voice identity processing through a functional connection with the anterior superior temporal sulcus (anterior STS) (Wu Ke et al. 2020), and the bilateral middle and posterior superior temporal sulcus (posterior STS/ superior STS) are involved in familiar voice identity processing (Wu Ke et al. 2020). posterior STS/ superior STS) characterise individual decoding of emotional rhythmic information in the voice (Leipold et al. 2022); thus, do groups on the autism spectrum differ in integrating real and cloned speaker identities compared to typical subjects? How does the introduction of rhythmic information as a within-speaker vocal variable moderate behavioural outcomes and the corresponding neural correlate representations? Differences in behavioural consequences resulting from rule deficits will further test the causal mechanisms of language rules in the coding and decoding of speaker identity.

Reference

[1] Albin, D.D., Echols, C.H. Stressed and word-final syllables in infant-directed speech [J]. *Infant Behavior and Development*, 1996(4): 401-418.
[2] Austin, J.L. *How to do things with words*: Oxford university press, 1975.
[3] Begus, K., Gliga, T., Southgate, V. Infants' preferences for native speakers are associated with an expectation of information [J]. *Proceedings of the National Academy of Sciences*, 2016(44): 12397-12402.
[4] Belin, P., Fecteau, S., Bedard, C. Thinking the voice: Neural correlates of voice perception [J]. *Trends in Cognitive Sciences*, 2004(3): 129-135.
[5] Belin, P., Zatorre, R.J. Adaptation to speaker's voice in right anterior temporal lobe [J]. *Neuroreport*, 2003(16): 2105-2109.
[6] Bernhold, Q.S., Giles, H. Vocal accommodation and mimicry [J]. *Journal of Nonverbal Behavior*, 2020(1): 41-62.
[7] Borkowska, B., Pawlowski, B. Female voice frequency in the context of dominance and attractiveness perception [J]. *Animal Behaviour*, 2011(1): 55-59.
[8] Braber, N., Cummings, L., Morrish, L. *Exploring language and linguistics* [M]. Cambridge:

Cambridge University Press, 2015.
[9] Campanella, S., Belin, P. Integrating face and voice in person perception [J]. *Trends in cognitive sciences*, 2007(12): 535-543.
[10] Chen, X., Li, Z., Setlur, S., et al. Exploring racial and gender disparities in voice biometrics [J]. *Scientific Reports*, 2022(1): 1-12.
[11] Chomsky, N. *Aspects of the theory of syntax*: MIT press, 2014.
[12] Coupland, N., Coupland, J., Giles, H., et al. Accommodating the elderly: Invoking and extending a theory1 [J]. *Language in Society*, 1988(1): 1-41.
[13] Dahl, K.L., Mahler, L.A. Acoustic features of transfeminine voices and perceptions of voice femininity [J]. *Journal of Voice*, 2020(6): 961-e919.
[14] Di Cesare, G., Cuccio, V., Marchi, M., et al. Communicative and affective components in processing auditory vitality forms: An fmri study [J]. *Cerebral Cortex*, 2022(5): 909-918.
[15] Fecher, N., Johnson, E.K. By 4.5 months, linguistic experience already affects infants' talker processing abilities [J]. *Child Development*, 2019(5): 1535-1543.
[16] Fleming, D., Giordano, B.L., Caldara, R., et al. A language-familiarity effect for speaker discrimination without comprehension [J]. *Proceedings of the National Academy of Sciences*, 2014(38): 13795-13798.
[17] Formisano, E., De Martino, F., Bonte, M., et al." Who" is saying" what"? Brain-based decoding of human voice and speech [J]. *Science*, 2008(5903): 970-973.
[18] Frühholz, S., Belin, P. *The oxford handbook of voice perception* [M]. Oxford: Oxford University Press, 2018.
[19] Frühholz, S., Schweinberger, S.R. Nonverbal auditory communication–evidence for integrated neural systems for voice signal production and perception [J]. *Progress in Neurobiology*, 2021: 101948.
[20] Ghazanfar, A.A., Rendall, D. Evolution of human vocal production [J]. *Current Biology*, 2008(11): R457-R460.
[21] Giles, H., Coupland, N., Coupland, I. 1. Accommodation theory: Communication, context, and [J]. *Contexts of accommodation: Developments in applied sociolinguistics*, 1991.
[22] Giles, H., Scholes, J., Young, L. Stereotypes of male and female speech: A british study [J]. *Central States Speech Journal* 1983(4).
[23] Goggin, J.P., Thompson, C.P., Strube, G., et al. The role of language familiarity in voice identification [J]. *Memory Cognition*, 1991(5): 448-458.
[24] Goldsmith, J.A., Riggle, J., Alan, C.L. *The handbook of phonological theory*: John Wiley & Sons, 2014.
[25] Grice, H.P. Logic and conversation *Speech acts*. Brill: 41-58, 1975.
[26] Grill-Spector, K., Henson, R., Martin, A. Repetition and the brain: Neural models of stimulus-specific effects [J]. *Trends in Cognitive Sciences*, 2006(1): 14-23.
[27] Hardy, T.L.D., Boliek, C.A., Wells, K., et al. Pretreatment acoustic predictors of gender, femininity, and naturalness ratings in individuals with male-to-female gender identity [J]. *American Journal of Speech-Language Pathology*, 2016(2): 125-137.
[28] Harrington, F.H., Mech, L.D. Wolf howling and its role in territory maintenance [J]. *Behaviour*, 1979(3-4): 207-249.
[29] Harrington, J., Palethorpe, S., Watson, C.I. Does the queen speak the queen's english? [J]. *Nature*, 2000(6815): 927-928.
[30] Hauser, M.D., Chomsky, N., Fitch, W.T. The faculty of language: What is it, who has it, and how did it evolve? [J]. *science*, 2002(5598): 1569-1579.
[31] Hills, T. The company that words keep: Comparing the statistical structure of child-versus adult-directed language [J]. *Journal of Child Language*, 2013(3): 586-604.
[32] Hogg, M.A. Masculine and feminine speech in dyads and groups: A study of speech style and gender salience [J]. *Journal of Language and Social Psychology*, 1985(2): 99-112.
[33] Hogg, M.A. Social identity theory [M].Shelley McKeown, R. H., Neil Ferguson. *Understanding peace and conflict through social identity theory: Contemporary global perspectives*. Switzerland: Springer: 3-17, 2016.
[34] Jia, Y., Zhang, Y., Weiss, R., et al. Transfer learning from speaker verification to multispeaker text-

to-speech synthesis [J]. *Advances in Neural Information Processing Systems*, 2018.
[35] Jiang, X., Gossack-Keenan, K., Pell, M.D. To believe or not to believe? How voice and accent information in speech alter listener impressions of trust [J]. *Quarterly Journal of Experimental Psychology*, 2020(1): 55-79.
[36] Jiang, X., Li, Y., Zhou, X. Is it over-respectful or disrespectful? Differential patterns of brain activity in perceiving pragmatic violation of social status information during utterance comprehension [J]. *Neuropsychologia*, 2013(11): 2210-2223.
[37] Jiang, X., Pell, M.D. The sound of confidence and doubt [J]. *Speech Communication*, 2017: 106-126.
[38] Johnson, K. The δf method of vocal tract length normalisation for vowels [J]. *Laboratory Phonology*, 2020(1).
[39] Kamide, Y. Learning individual talkers' structural preferences [J]. *Cognition*, 2012(1): 66-71.
[40] Kharitonov, E., Copet, J., Lakhotia, K., et al. Textless-lib: A library for textless spoken language processing [J]. *arXiv preprint arXiv:2202.07359*, 2022.
[41] Kim, J., Toutios, A., Lee, S., et al. Vocal tract shaping of emotional speech [J]. *Computer Speech Language*, 2020: 101100.
[42] Kinzler, K.D. Language as a social cue [J]. *Annual Review of Psychology*, 2021: 241-264.
[43] Kinzler, K.D., Dupoux, E., Spelke, E.S. 'Native'objects and collaborators: Infants' object choices and acts of giving reflect favor for native over foreign speakers [J]. *Journal of Cognition Development*, 2012(1): 67-81.
[44] Kreuk, F., Polyak, A., Copet, J., et al. Textless speech emotion conversion using decomposed and discrete representations [J]. *arXiv preprint arXiv:2111.07402*, 2021.
[45] Kroczek, L.O.H., Gunter, T.C. The time course of speaker-specific language processing [J]. *Cortex*, 2021: 311-321.
[46] Kuhl, P.K. Who's talking? [J]. *Science*, 2011(6042): 529-530.
[47]Künzel, H.J. How well does average fundamental frequency correlate with speaker height and weight? [J]. *Phonetica*, 1989(1-3): 117-125.
[48] Labov, W. *The social stratification of english in new york city* [M]. Cambridge: Cambridge University Press, 2006.
[49] Lammert, A.C., Narayanan, S.S. On short-time estimation of vocal tract length from formant frequencies [J]. *PloS one*, 2015(7): e0132193.
[50] Latinus, M., Belin, P. Anti-voice adaptation suggests prototype-based coding of voice identity [J]. *Frontiers in Psychology*, 2011: 175.
[51] Lavan, N., Burston, L.F., Ladwa, P., et al. Breaking voice identity perception: Expressive voices are more confusable for listeners [J]. *Quarterly Journal of Experimental Psychology*, 2019a(9): 2240-2248.
[52] Lavan, N., Burton, A.M., Scott, S.K., et al. Flexible voices: Identity perception from variable vocal signals [J]. *Psychonomic Bulletin Review*, 2019b(1): 90-102.
[53] Lavan, N., Knight, S., McGettigan, C. Listeners form average-based representations of individual voice identities [J]. *Nature communications*, 2019c(1): 1-9.
[54] Lee, S., Potamianos, A., Narayanan, S. Acoustics of children's speech: Developmental changes of temporal and spectral parameters [J]. *The Journal of the Acoustical Society of America*, 1999(3): 1455-1468.
[55] Leipold, S., Abrams, D.A., Karraker, S., et al. Neural decoding of emotional prosody in voice-sensitive auditory cortex predicts social communication abilities in children [J]. *Cerebral Cortex*, 2022.
[56] Levi, S. Methodological considerations for interpreting the language familiarity effect in talker processing [J]. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2019(2): e1483.
[57] Martin, A.E., Slepian, M.L. The primacy of gender: Gendered cognition underlies the big two dimensions of social cognition [J]. *Perspectives on Psychological Science*, 2021(6): 1143-1158.
[58] Matsumoto, H., Hiki, S., Sone, T., et al. Multidimensional representation of personal quality of vowels and its acoustical correlates [J]. *IEEE Transactions on Audio Electroacoustics*, 1973(5): 428-436.
[59] Nakagawa, S., Shikano, K., Tohkura, Y.i. *Speech, hearing and neural network models* [M].

Amsterdam: IOS Press, 1995.
[60] Neumann, K., Welzel, C. The importance of the voice in male-to-female transsexualism [J]. *Journal of Voice*, 2004(1): 153-167.
[61] Oh, D., Buck, E.A., Todorov, A. Revealing hidden gender biases in competence impressions of faces [J]. *Psychological Science*, 2019(1): 65-79.
[62] Orena, A.J., Theodore, R.M., Polka, L. Language exposure facilitates talker learning prior to language comprehension, even in adults [J]. *Cognition*, 2015: 36-40.
[63] Peng, Z., Chen, J., Jin, L., et al. Social brain dysfunctionality in individuals with autism spectrum disorder and their first-degree relatives: An activation likelihood estimation meta-analysis [J]. *Psychiatry Research: Neuroimaging*, 2020: 111063.
[64] Pernet, C.R., McAleer, P., Latinus, M., et al. The human voice areas: Spatial organisation and inter-individual variability in temporal and extra-temporal cortices [J]. *Neuroimage*, 2015: 164-174.
[65] Perrachione, T.K., Del Tufo, S.N., Gabrieli, J.D. Human voice recognition depends on language ability [J]. *Science*, 2011(6042): 595-595.
[66] Perrachione, T.K., Wong, P.C. Learning to recognise speakers of a non-native language: Implications for the functional organisation of human auditory cortex [J]. *Neuropsychologia*, 2007(8): 1899-1910.
[67] Pisanski, K., Anikin, A., Reby, D. Static and dynamic formant scaling conveys body size and aggression [J]. *Royal Society Open Science*, 2021(1): 211496.
[68] Pisanski, K., Anikin, A., Reby, D. Vocal size exaggeration may have contributed to the origins of vocalic complexity [J]. *Philosophical Transactions of the Royal Society B*, 2022(1841): 20200401.
[69] Polka, L., Masapollo, M., Ménard, L. Setting the stage for speech production: Infants prefer listening to speech sounds with infant vocal resonances [J]. *Journal of Speech, Language*, 2022(1): 109-120.
[70] Reby, D., McComb, K. Anatomical constraints generate honesty: Acoustic cues to age and weight in the roars of red deer stags [J]. *Animal Behaviour*, 2003(3): 519-530.
[71] Regel, S., Coulson, S., Gunter, T.C. The communicative style of a speaker can affect language comprehension? Erp evidence from the comprehension of irony [J]. *Brain research*, 2010: 121-135.
[72] Rubin, D.L. Nonlanguage factors affecting undergraduates' judgments of non-native english-speaking teaching assistants [J]. *Research in Higher Education*, 1992(4): 511-531.
[73] Schirmer, A. Is the voice an auditory face? An ale meta-analysis comparing vocal and facial emotion processing [J]. *Social Cognitive Affective Neuroscience*, 2018(1): 1-13.
[74] Šebesta, P., Mendes, F.D.C., Pereira, K.J. Vocal parameters of speech and singing covary and are related to vocal attractiveness, body measures, and sociosexuality: A cross-cultural study [J]. *Frontiers in Psychology*, 2019: 2029.
[75] Shutts, K., Kinzler, K.D., McKee, C.B., et al. Social information guides infants' selection of foods [J]. *Journal of Cognition Development*, 2009(1-2): 1-17.
[76] Smith, D.R.R., Patterson, R.D. The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age [J]. *The Journal of the Acoustical Society of America*, 2005(5): 3177-3186.
[77] Soderstrom, M., Blossom, M., Foygel, R., et al. Acoustical cues and grammatical units in speech to two preverbal infants [J]. *Journal of Child Language*, 2008(4): 869-902.
[78] Sorokowski, P., Puts, D., Johnson, J., et al. Voice of authority: Professionals lower their vocal frequencies when giving expert advice [J]. *Journal of Nonverbal Behavior*, 2019(2): 257-269.
[79] Stern, D.N., Spieker, S., MacKain, K. Intonation contours as signals in maternal speech to prelinguistic infants [J]. *Developmental Psychology*, 1982(5): 727.
[80] Tamagawa, R., Watson, C.I., Kuo, I.H., et al. The effects of synthesised voice accents on user perceptions of robots [J]. *International Journal of Social Robotics*, 2011(3): 253-262.
[81] Tan, Z.-H. Vocal tract length perturbation for text-dependent speaker verification with autoregressive prediction coding [J]. *IEEE Signal Processing Letters*, 2021: 364-368.
[82] Tang, C., Hamilton, L.S., Chang, E.F. Intonational speech prosody encoding in the human auditory cortex [J]. *Science*, 2017(6353): 797-801.
[83] Titze, I.R. Physiologic and acoustic differences between male and female voices [J]. *The Journal of the Acoustical Society of America*, 1989(4): 1699-1707.
[84] Voigt, R., Jurafsky, D., Sumner, M. *Between-and within-speaker effects of bilingualism on f0*

variation [C]. Interspeech. San Francisco, The United States, 2016:1122-1126.
[85] von Kriegstein, K., Warren, J.D., Ives, D.T., et al. Processing the acoustic effect of size in speech sounds [J]. *Neuroimage*, 2006(1): 368-375.
[86] Walker, M., Perry, C. It's the words you use and how you say them: Electrophysiological correlates of the perception of imitated masculine speech [J]. *Language, Cognition and Neuroscience*, 2022(1): 1-21.
[87] Winters, S.J., Levi, S.V., Pisoni, D.B. Identification and discrimination of bilingual talkers across languages [J]. *The Journal of the Acoustical Society of America*, 2008(6): 4524-4538.
[88] Xu, H., Armony, J.L. Influence of emotional prosody, content, and repetition on memory recognition of speaker identity [J]. *Quarterly Journal of Experimental Psychology*, 2021(7): 1185-1201.
[89] Xu, M., Homae, F., Hashimoto, R.-i., et al. Acoustic cues for the recognition of self-voice and other-voice [J]. *Frontiers in psychology*, 2013: 735.
[90] Zhang, Y., Ding, Y., Huang, J., et al. Hierarchical cortical networks of "voice patches" for processing voices in human brain [J]. *Proceedings of the National Academy of Sciences*, 2021(52): e2113887118.
[91] 陈忠敏. 语音感知的特点及其解剖生理机制 [J]. *中国语音学报*, 2021(1): 8-24.
[92] 蒋晓鸣. 文化互鉴视角下非言语表情的嗓音编码和解码 [J]. *《 同济大学学报》(社会科学版)*, 2020(1): 116-124.
[93] 明莉莉, 胡学平. 人类嗓音加工的神经机制——来自正常视力者和盲人的脑神经证据 [J]. *心理科学进展*, 2021(12): 2147.
[94] 束定芳. 《语言与社会心理学》评介——兼论社会心理语言学的研究对象、目标及方法 [J]. *外国语(上海外国语学院学报)*, 1992(03): 10-14.
[95] 束定芳, 张立飞. 后"经典"认知语言学：社会转向和实证转向 [J]. *现代外语*, 2021(03): 420-429.
[96] 王德春, 孙汝建. 社会心理语言学的理论和方法论基础 [J]. *外国语(上海外国语学院学报)*, 1992a(04): 3-7+82.
[97] 王德春, 孙汝建. 社会心理语言学的学科性质和研究对象 [J]. *外国语(上海外国语学院学报)*, 1992b(03): 3-9+82.
[98] 伍可, 陈杰, 李雯婕, et al. 人声加工的神经机制 [J]. *心理科学进展*, 2020(5): 752-765.
[99] 夏志华, 马秋武. *同济博士论丛：汉语对话中韵律趋同的实验研究* [M]. 上海: 同济大学出版社, 2019.
[100] 周爱保, 胡砚冰, 周滢鑫, et al. 听而不"闻"? 人声失认症的神经机制 [J]. *心理科学进展*, 2021(3): 414.