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

陈文均[1]，胡砚冰[1]，蒋晓鸣[1]

(1. 上海外国语大学 语言研究院，上海 201620)

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

CHEN Wenjun[1], *HU Yanbing*[1], *JIANG Xiaoming*[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

research's attention in three directions.

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

基金项目/Funding：

上海市哲学社会科学规划课题 (2018BYY019)；上海市教育发展基金会和上海市教育委员会"曙光计划" (20SG31)；上海市自然科学基金面上项目（22ZR1460200)；上海外国语大学第五届"导师学术引领计划项目"（2022113001）Shanghai Philosophy and Social Science Planning Project (2018BYY019); Shanghai Education Development Foundation and Shanghai Education Commission "Aurora Project" (20SG31); Shanghai Natural Science Foundation (22ZR1460200); Shanghai International Studies University 5th "Mentor Academic Leadership Programme" (2022113001)

Received: 2022-09-04

Authors:

Chen Wenjun (1999-), male, from Suining, Sichuan, postgraduate student. Research interests: psychology and neurolinguistics, voice encoding and decoding.

Hu Yanbing (1996-), Male, from Tianshui, Gansu, PhD student. Research interests: psychology and neurolinguistics, voice expression decoding and voice production.

Xiaoming Jiang* (Corresponding author; 1983-), Male, from Shanghai, PhD, Professor, Shanghai Shuguang Scholar. Research interests: psychology and neurolinguistics, experimental linguistics, verbal communication and speech disorders, voice encoding and decoding, neuropragmatics.

1. Introduction

In Dream of the Red Chamber, Lin Daiyu is able to 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) (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 purpose (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., much less attention has been paid to the social interaction dimension than to the cognitive-psychological dimension (Shu 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 discourse patterns and psychosocial speech mechanisms requires the integration of the disciplines of sociolinguistics, psycholinguistics, and engineering linguistics (Wang Dechun and Sun Rujian 1992a; 1992b). Shu 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., non-standard 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. (1992), in which he argues that the social-psychological role of encoding and decoding in speech communication is the main object of study in social-psychological linguistics. In his outlook, Shu 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 psychosocial linguistics that 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.

Why is this paper concerned with the relationship between how linguistic rules (phonological and syntactic structures) and linguistic style affect listeners' decoding of speaker identity and subsequent decisions about social interaction schemes in spoken communication? The classical linguistic theory does not consider how the variability of acoustic information in multimodal interaction scenarios, such as spoken communication, affects listeners' perceptions of the individual and group identities of speakers. For example, Chomsky (1969:48-50) argues that language users have an innate ability which 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 have different backgrounds and experiences. Austin's (1975:100) speech Grice (1975) argues that there is more than what is implied by the literal meaning of the speaker and that listeners use the implicit meaning of the conversation to understand the sentence. This theory suggests that context and non-verbal information play a key role in understanding discourse but does not emphasise the internal identity of the individual speaker.

It is clear that 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 that 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' and 'you' 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 structures, such as the dichotomy between SOV and OSV syntactic

structures in German, where listeners expect the speaker to speak simple SOV sentences, but when they actually hear the speaker speak complex OSV sentences, the experimenter observes increased P600 activity in the listener's brain (Kroczek & Gunter 2021). A similar P600 effect was observed in reading experiments under the pragmatic category when the speaker did or did not take a commonly used sarcastic expression (Regel et al. 2010). Thus, from the speaker's perspective, linguistic rules consisting of phonological structure and linguistic style (focusing on communication at the phonological level) influence the speaker's speech production, with differences in output reflected in sophisticated acoustic analyses (e.g., intergroup differences in parameters such as fundamental frequency, vocal tract length parameters, sound intensity, duration, jitter, and shimmer, which can be: confident, neutral, sceptical "sense of knowing" driven speech rhythm differences (Jiang & Pell 2017)). Listeners, on the other hand, will be sensitive to acoustic variation in the speech produced by the speaker. In particular, the fundamental frequency and vocal tract length critically characterise identity differences within the individual speaker and between the individual speaker and other individual speakers; thus, individuals are sensitive to changes in speaker identity in speech. Evidence for the above inference comes from an EEG study exploring listeners' decoding of confidence levels in speech produced by speakers of English with different accents, which found that linguistic structure and speaker identity information, including phonological structure, are processed at an early stage of ERP (Jiang et al. 2020). Therefore, this paper focuses on the individual's understanding of the speaker's produced speech in the context of phonological transmission of information, focusing on how listeners interpret the identity of the speaker and thus influence the listener's vocalisation strategies to engage in social interaction.

Much of the research on speaker identity encoding and decoding has focused on the cognitive processing patterns involved in individual vocalisations under controlled conditions (i.e. experimental stimuli recorded with neutral voice expressions). 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 judicial practice in which suspect identification, where the person listens to an array of speech and then recalls memory to point to the identity of the suspect (Frühholz & Belin 2018; Levi et al. 2019). There are several 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 Xeping (2021) compared the differences in the processing of human voice identity between normal sighted people (2021) compared the brain mechanisms that differ in the processing of human voice identity between normal sighted and blind people. 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, affect speakers' 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 research perspectives based on this.

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

Language faculty theory (Hauser et al. 2002) suggests a relationship between individual speaker identity and language coding. Broad language faculty (FLB) encompasses the physiological basis of individual speaker identity coding, and sound identity coding, and decoding is a universal ability across species; narrow language faculty (FLN) encompasses the recursive nature of language structure, and human speaker identity coding has become more complex as a result of language evolution. The human speaker identity code has become more complex as a result of language evolution. After determining language structure and language style at the speech planning stage, speakers implement specific 'vocal strategies' by invoking the physiological bases of vocalisation, based on which they produce speech sounds that carry their vocal identity. These include phonological structure (e.g., syllabic feature

preferences resulting from the speaker's accent (Coupland 2007:173)) and syntactic structure (e.g., the speaker's preference for SOV or OSV structure (Kroczek & Gunter 2021)); and linguistic style, which includes speaking style or specific, pragmatic choices (e.g., a tendency to speak sarcastically (Regel et al. 2010), the degree of [r]-sound curl due to stylistic variant (Labov 2006:40-47), or gender-binary speaking styles (Hogg 1985), among others).

2.1 Encoding and recognition of voice identity as a universal competence

The body size of living organisms is a key element of vocal identity coding, for example, wolves howl to indicate territory to their mates when hunting or calving, and the perception of body size is common in the social organisation of many species (Harrington & Mech 1979). Reby & 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; Šebesta et al. (2019) invited 84 heterosexual participants from Brazil and 68 heterosexual participants from the Czech Republic to read short sentences and sing aloud and to report on socialised sexuality and found that shorter vocal tract length in short speeches and longer vocal tract length in singing predicted female sexual behaviour. This suggests that a species judgement of identity is a universal ability.

The human ability to decode each other's identity from sounds preceded the emergence of verbal communication: Polka et al. (2022) synthesised vowels/i/audio from infants with widely spaced resonance peaks (F2-F1 = 3761) and adult females with less spacing (F2-F1 = 2315) and found that infants with an average age of 220 days had a preference for vowels that mimicked infant vocalisation states. And this ability was strongly related to language acquisition: infants who had been exposed only to English for most of their mean age 136 days were presented by Fecher & Johnson (2019) with English, Polish and Spanish sentences recorded by four bilingual (two English and Polish speakers; two English and Spanish speakers) females; with gaze duration as the dependent variable, speaker (A mixed linear model with the duration of gaze as the dependent variable and speaker (different/same) and language (native/non-native) as the main fitting parameters showed no main effects for either speaker or language but an interaction between them, suggesting that whether the stimuli were in the infant's native language modulated the infant's duration of gaze when listening to audio from the same or different speakers.

Studies based on other species and early language acquisition suggest that humans have retained the ability to recognise each other from sound during evolution and that this ability predates verbal

communication, but that language is a uniquely human phenomenon that complicates this ability compared to other species.

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

Listeners are able to integrate information about the identity of the speaker (i.e., 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 the phonological memory and phonological awareness subtests of the comprehensive test of phonological processing (CTOPP) impaired the phonological rules of English. The group with impaired phonological rules (as characterised by scores on the CTOPP subtests of phonological memory and phonological awareness) and clinically diagnosed as dyslexic[1] had comparable accuracy in recognising the identity of speakers coded in their native English and in a completely unfamiliar Chinese language, both of which were much lower than those of healthy controls. From a phonological perspective, individuals who are completely ignorant of a language will show difficulty in accurately hearing and producing the sounds and sound patterns of that language and will therefore lack knowledge of the phonological rules specific to that language (Goldsmith et al. 2014:319). People with English monolingual dyslexia lack knowledge of Chinese phonological rules, and their dyslexia impairs English phonological rules, resulting in knowledge of the phonological rules of their native language at an unfamiliar language level, an impairment that 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 & 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 faster and more accurately than English monolinguals in Connecticut, USA,

[1] Dyslexia is a developmental reading and spelling disorder caused by the brain's inability to coordinate the processing of visual and auditory information, mainly in childhood. Dyslexia is often characterised by intellectual abnormalities, with special attention paid to dyslexia due to low intelligence and those with very high IQs. It is characterised by functional abnormalities in literacy, spelling and reading, and the acquisition of the phonological rules of language is an important foundation for children learning to read and spell.

suggesting that being exposed to usage scenarios with specific phonological rules can also contribute to the emergence of language familiarity effects.

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 & Gunter (2021) first trained subjects to be exposed to experimental conditions with specific speakers of different syntactic structure distributions (e.g., Speaker A spoke 70% OSV and 30% SOV sentences and Speaker B the opposite), from which subjects built up an expected representation of the specific speaker as an "OSV speaker" or "OSV speaker"; when the subject heard the "SOV speaker" speak OSV sentences, the EEG component of the test showed increased P600 activity in the postcentral part of the brain - characterising the expectation of the syntactic structure of the particular speaker. The subjects showed increased P600 activity in the posterior part of the brain when they heard the "SOV speaker" speak OSV sentences during the test - indicating an expected reanalysis or repair of the syntactic structure of the particular speaker. Similar experimental manipulations have also found that listeners can build anticipation of the speaker's high/low syntactic attachment style (Kamide 2012).

More research has also shown that listeners bind the identity of the speaker to a particular linguistic style. For example, Regel et al. (2010) used an experimental design similar to the above to allow readers to form expectations of two people using sarcastic/literal different discourse styles and similar enhanced P600 activity was found when readers read the sarcastic discourse of the literal stylist.

Notably, Walker & Perry (2022) manipulated male- and female-specific rhyme patterns (e.g. a female using habitual rhymes vs. imitating male rhymes) and language style (masculine/feminine lexical use), and when subjects heard the female speaking in male rhymes, EEG activity induced enhanced N400 activity reflecting representational semantic inconsistencies or inconsistencies as expected, and speaker There was also an interaction between rhyme and language style, i.e. when female speaker identity and female rhyme were consistent with female language style, different EEG activity was shown compared to inconsistent situations. This study suggests that the social category (dichotomous rhymes) and the linguistic ontology category (gender-specific vocabulary) jointly influence the outcome of speech production and that listeners associate processing based on both categories with the decoding of speaker identity.

From the above, it can be deduced that the ability of individuals to learn the tendency to use syntactic structures and language styles of different speakers in a strictly controlled laboratory is closely related to the linguistic communication practice of individuals who constantly combine features of speaker language use with their identity in natural interaction scenarios over time.

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

Language is a product of the higher evolution of the human species, which makes 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 identities behind the German words with their corresponding names; after eight rounds, of familiarisation-refamiliarisation- After eight rounds of familiarisation-refamiliarisation-recognition, the subjects were able to associate the 10 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 bilinguals well above the chance level. This suggests that there are acoustic cues in the speech that steadily encode vocal identity independent of phonological structure, namely the fundamental frequency (Matsumoto et al. 1973; Xu et al. 2013) and the resonance peak spacing that characterises vocal tract length (Ghazanfar & Rendall 2008. Johnson 2020), with fundamental frequency and vocal tract length interacting to influence timbre and encode 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. 1995:75-83). 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 there is a direct relationship between vocal tract length and 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 & 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 & 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. Parameters: root mean square energy (RMS energy) reflecting temporal information, spectral centroid and spectral roll-off reflecting spectral information, Mel frequency cepstrum coefficients (MFCCs) reflecting the shape of the spectral envelope, and entropy correlation values-spectral entropy reflecting the amount of information in the signal spectral entropy, probability density function (PDF entropy), permutation entropy, and singular value decomposition (SVD entropy). No significant differences were found between races for the above parameters; however, significant differences were found between males and females (regardless of race) for several indicators.

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

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

The field of vocal identity mostly uses vocalisations in speakers' neutral voices as stimuli to explore listeners' recognition mechanisms of vocal identity (Perrachione et al. 2011; Fleming et al. 2014). However, the language used, and the need to express paralinguistic information in different contexts (e.g., the speaker's assertiveness and doubtfulness) allow for a high degree of intra-speaker variability. For example, Voigt et al. (2016) analysed recordings of 25 German-French and 20 German-Italian bilinguals during interviews on lighter topics and found that German-French bilingual women used higher mean basal frequencies when speaking French, and German-Italian bilingual women used lower basal frequencies when speaking Italian. In an analysis of declarative showing different levels of speakers' 'sense of knowing', it was found that the basal frequency of assertiveness was higher when looking at sentence-initial and mid-sentence components compared to assertiveness, yet the basal

frequency of assertiveness was higher when looking at sentence-final components (Jiang & Pell 2017), which has implications for speaker identity only through acoustic parameters that characterise inter-speaker variation. This challenges the idea that speaker identity is encoded solely through acoustic parameters that characterise inter-speaker variation, i.e., listeners rely on more complex cues when decoding speaker identity.

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

However, listeners' ability to adapt to internal variations in speaker identity is limited. Lavan et al. (2019a) normalised 1.2 to 4-second-long emotional vocal spikes from the two male leads of the American drama series Desperado to 0.400 Pa and, after low-pass filtering at 10 kHz, asked listeners to drag and drop classify the speaker identities behind the synthesised manipulated audio; the results It was found that even listeners familiar with the episode still had difficulty in accurately classifying the audio as two speakers, but rather as more than one speaker. Xu & Armony (2021) prepared 4 sentences (2 emotional rhythms: fear/neutral* 2 semantic contents) from each of the 12 speakers and presented them with 6 of the speakers under the neutral/fear rhythm during the listener familiarisation phase In 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 accurate than 80% 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 permeates intentions to regulate 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 of 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). 2012). (Begus et al. 2016) presented infants 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 for noun instruction and observed infants' 3-5 Hz theta band activity (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, as a result, determine the group

identity of the speaker and thus interact socially with different speaker groups in a differentiated manner.

Accent rules are part of language, and the speaker's accent is used by listeners to classify groups; Rubin (1992) played audio lectures in a standard Southern American accent to North American university students and matched them with 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 credentials, 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. The subjects were given credibility ratings after listening to audio recorded with a Canadian English accent, an Australian English accent, and a person with a Quebec French accent with a confident or sceptical voice expression, and found that in the confident condition, Tamagawa et al. (2011) provided evidence from outside of real speakers: subjects with New Zealand accents listened to audio based on British, American, and New Zealand accents that introduced the same product. After listening to a synthetic speech from a robot that introduced the same blood pressure monitor based on accents trained in the UK, US, and New Zealand, it was concluded that the US accent was more machine-like than the synthetic voice of the New Zealand accent, and performed worse than the robot with the New Zealand synthetic accent. The accents of the robots in this experiment belonged to the representation of different phonological structures in the language structure. These three experiments suggest that speakers' choice of phonological structure critically encodes their group identity as perceived by listeners; and that listeners will vary their interaction decisions based on their perception of different speaker identities.

Speech styles based on gender dichotomies also mark group identity. Men's speech styles are typically characterised by the use of slang (vulgarity), more blunt speech, lower voices, aggressiveness, and appearing more authoritative, whereas women's are characterised by greater variation in rate and fundamental frequency, gentler speech, openness, self-disclosure, and appearing more emotional (Giles et al. 1983; Hogg 1985). Slepian's (2021) "big two" model suggests that individuals' judgments and evaluations of others' traits follow two dimensions that overlap significantly with gender roles: a. agency/masculinity, which is assertive, competitive, dominant, independent, self-interested, and goal-seeking; and b. community/femininity, which is nurturing, warm, expressive, and emotional. i.e. nurturing, warmth, expression, concern for others, and social orientation. Of these, masculinity is strongly associated with the perceived 'competence' of others; for example, masculine facial features

make individuals appear more competent (Oh et al. 2019), women with lower, i.e. more masculine, voices are perceived as more dominant, and feminine voices are associated with naivety and sexual immaturity (Borkowska & 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 & Giles 2020). Bernhold & Giles 2020), and such rhyme-level convergence has also been found in natural conversational contexts in Chinese (Xia Zhihua & Ma Qiowu 2019). An example of convergence under typical social categories is the adoption of each other's common speech patterns by parties of high and low-status hierarchies (Shu 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 & 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 universities 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 male and female speakers had lower fundamental frequencies when giving professional advice and that women (Harrington et al. (2000) investigated the vowels in the audio of Queen Elizabeth II's speeches from the 1950s to the 1980s and found a tendency to move towards a younger demographic and a commoner approach to her vocalisation. The above examples suggest that speakers strategically change their choice of speech style in order to appear friendly or more professional and that such changes are reflected not only in the level of effort put into modulating

the vocal base but also in the calculation and execution of stylistic variant.

On the other hand, Pisanski et al. (2021) suggest that vocalic complexity in human-voiced speech may have its origins in a common phenomenon in the animal kingdom: species lower vocal tract resonance (i.e. lower resonance peaks) to achieve vocal body exaggeration, a phenomenon that exists in humans with complex language systems and is also common in other groups that do not have It is also common in other animal groups that do not have a similar human language system. In human speech scenarios, the vocal tract length is longer when the speaker is aggressive compared to neutral vocalisations Pisanski et al. (2022), and the vocal tract length is shorter when the speaker is happy compared to angry, sad or neutral vocalisations (Kim et al. 2020), and the speaker produces lower fundamental frequencies for confident compared to unconfident recordings (Jiang & Pell 2017), such acoustic parameters suggest that speakers encode their identity by lengthening or shortening their vocal tracts and other ways of modulating their physiological underpinnings; and that changes in these vocal strategies will directly affect the listener's perception of the speaker's identity. That is, humans have evolved a language system that retains the ability to change the body shape of the voice by lengthening/shortening the vocal tract, raising/lowering the fundamental frequency, and subconsciously changing the body shape of the voice in specific speech situations (e.g., speakers make themselves sound larger when they are aggressive). From the above, it is clear that more short-lived, dynamic paralinguistic messaging may be related to the formation of stable language structures and language styles in humans over time, 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 production based on the integrated information about the speaker's identity and language. Therefore, this paper integrates (1) the vocal 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 Braber et al. (2015:335) from the perspective of the listener-speaker discourse wheel cycle; and (3) Jiang et al. cycle; (3) 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 the vocal information structure (vocal structure) in speaker speech, including voice identity information and voice speech structure (including language structure) in the early stage of human voice processing; and finally proposed a social The final framework for encoding and decoding speaker identity in interactive scenarios is proposed (Figure 1).

The two basic elements of the framework are: (1) emotional information with basic emotions such as surprise, happiness, anger, sadness, fear, disgust, and neutrality, as well as assertive voice signals that express a "sense of knowing" (Jiang 2020); (2) identity information with variables such as gender, age, education, attractiveness, ability, and group ethnicity (Frühholz & Belin 2018).

At the beginning of the discourse round, the speaker completes linguistic encoding by selecting the listener-specific phonological and syntactic structures driven by communicative intentions that permeate the identity of the other group (e.g., whether to adopt a typically masculine speech style); during the speech-motor encoding phase, the speaker completes a plan of how to invoke vocal foundations such as the tongue, lips, and vocal folds based on linguistic information (e.g., specific syntactic and syllabic structural features (Labov 2006:40-47) and acoustic rules of expression for paralinguistic information (e.g., whether to lengthen the vocal tract, lower the fundamental frequency to appear more confident (Jiang & Pell 2017)) to complete a plan of how to invoke the vocal base of the tongue, lips, and vocal folds; in the speech motor execution phase, the anatomical basis of vocalisation is controlled by neural signals from the speaker's brain to complete speech production and transmit sound waves.

The listener's auditory system converts mechanical wave vibrations into neural signals that are transmitted to the auditory centre, completing the reception of auditory information. During the speech perception phase, the listener performs a structural analysis of the voice at around 100 ms, simultaneously processing voice identity, 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.

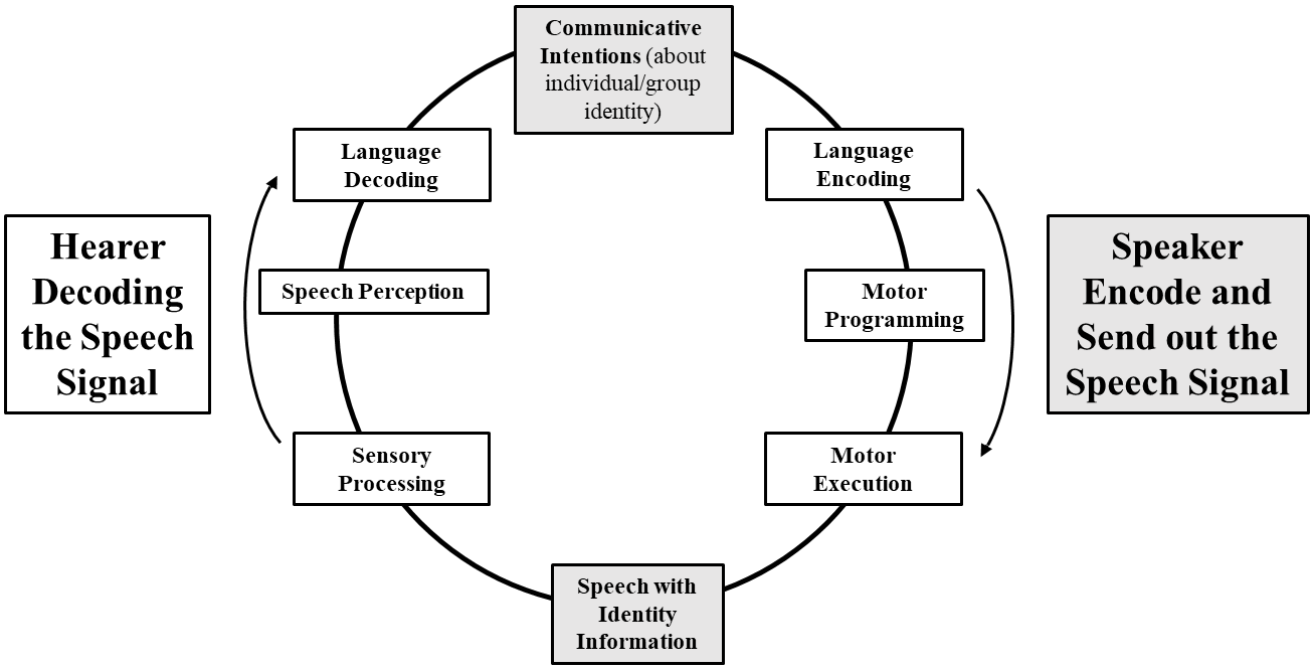

Figure 1. Speaker Identity Encoding and Decoding Model for Verbal Interaction Scenarios

4. Research outlook

Based on the above, it is clear that the encoding and decoding of individual and group identities of speakers interact with specific dimensions of language and society. Whereas the previous paper focused on how several factors affect the representation of speaker identity, future research could invert this by exploring how the representation and perception of speaker identity affect social cognition, for example, whether and how listeners' encoding and decoding of the speaker's emotions are moderated by the speaker's identity. Beyond this, future research could explore (1) what indicators characterise speakers' efforts to adjust vocal movements for group penetration purposes; (2) the brain mechanisms of listeners' processing of artificially intelligent cloned human voices and the moderating effects of affective rhythmic cues on cross-group vocal decoding; and (3) internal variation in the encoding and decoding of speaker identity 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 productions guided 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 that specific electrode sites in the left anterior temporal lobe of epileptic patients responded only to native human voices via cortical EEG ECoG, and more notably, Di Cesare et al. (2022) presented listeners with More notably, Di Cesare et al. (2022) presented listeners with the word 'hello', which conveys social intent, and the real voice specifically activated the listener's dorsal-central insula compared to the synthetic voice of a neutral voice. So are there specific neural correlates that characterise the way in which individuals differentially process real voices and their cloned counterparts? 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", would listeners' group categorisation of cloned voices be altered? Future research could combine electrophysiological and magnetic resonance imaging to explore 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 real and cloned voices (Belin & Co. amp;

Zatorre 2003; Grill-Spector et al. 2006). Once again, how will vocal recognition technology respond to the speaker identity crisis posed by cloned human voices? Although upgrading vocal length normalisation algorithms is known to improve the accuracy of human voice identity products (Tan 2021), the role of the broader spectrum, cepstrum and other parameters mentioned earlier in altering recognition accuracy is unclear. In addition, as AI voice services become more "anthropomorphic" in sound, the speech produced during human-computer interaction may become more "expert" and "customised" with the support of large language models such as ChatGPT. "Will individual users' perceptions and attitudes towards intelligent services be moderated by the verbal content?

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 individuals 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 & Welzel 2004; Hardy et al. 2016; Dahl & Mahler 2020), interventions that affect the representation and application of phonological rules, and how will this inform the implementation of corrective programmes for groups with gender identity crises? 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) is 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, are there differences in the integration of real and cloned speaker identities in autism spectrum groups compared to typical subjects? How does the introduction of rhythmic information as an intra-speaker vocal variable moderate behavioural outcomes and corresponding neural correlate representations? Differences in behavioural consequences due to 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* [M]. Oxford: 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. &Xu, W. Exploring racial and gender disparities in voice biometrics [J]. *Scientific Reports*, 2022(1）: 1-12.

[11] Chomsky, N. *Aspects of the Theory of Syntax* [M]. Cambridge, MA: MIT press, 1969.

[12] Coupland, N. *Style: Language variation and identity* [M]. Cambridge: Cambridge University Press, 2007.

[13] Coupland, N., Coupland, J., Giles, H. &Henwood, K. Accommodating the elderly: Invoking and extending a theory [J]. *Language in Society*, 1988(1）: 1-41.

[14] Dahl, K.L. &Mahler, L.A. Acoustic features of transfeminine voices and perceptions of voice femininity [J]. *Journal of Voice*, 2020(6）: 961-e919.

[15] Di Cesare, G., Cuccio, V., Marchi, M., Sciutti, A. &Rizzolatti, G. Communicative and affective components in processing auditory vitality forms: An fMRI study [J]. *Cerebral Cortex*, 2022(5）: 909-918.

[16] Fecher, N. &Johnson, E.K. By 4.5 months, linguistic experience already affects infants' talker processing abilities [J]. *Child Development*, 2019(5）: 1535-1543.

[17] Fleming, D., Giordano, B.L., Caldara, R. &Belin, P. A language-familiarity effect for speaker discrimination without comprehension [J]. *Proceedings of the National Academy of Sciences*, 2014(38）: 13795-13798.

[18] Formisano, E., De Martino, F., Bonte, M. &Goebel, R. " Who" is saying" what"? Brain-based decoding of human voice and speech [J]. *Science*, 2008(5903）: 970-973.

[19] Frühholz, S. &Belin, P. *The Oxford handbook of voice perception* [M]. Oxford: Oxford University Press, 2018.

[20] 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.

[21] Ghazanfar, A.A. &Rendall, D. Evolution of human vocal production [J]. *Current Biology*, 2008(11）: R457-R460.

[22] Giles, H., Coupland, J., Coupland, N. &Oatley, K. *Contexts of accommodation: Developments in applied sociolinguistics*: Cambridge University Press, 1991.

[23] Giles, H., Scholes, J. &Young, L. Stereotypes of male and female speech: A British study [J]. *Central States Speech Journal* 1983(4）.

[24] Goggin, J.P., Thompson, C.P., Strube, G. &Simental, L.R. The role of language familiarity in voice identification [J]. *Memory Cognition*, 1991(5）: 448-458.

[25] Goldsmith, J.A., Riggle, J. &Alan, C.L. *The handbook of phonological theory* [M]. New York: John Wiley & Sons, 2014.

[26] Grice, H.P. Logic and conversation [M] //Peter Cole, Morgan, J. L., *Speech acts*. New York: Academic

Press; 41-58, 1975.
[27] 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.
[28] Hardy, T.L.D., Boliek, C.A., Wells, K., Dearden, C., Zalmanowitz, C. &Rieger, J.M. 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.
[29] Harrington, F.H. &Mech, L.D. Wolf howling and its role in territory maintenance [J]. *Behaviour*, 1979(3-4）: 207-249.
[30] Harrington, J., Palethorpe, S. &Watson, C.I. Does the Queen speak the Queen's English? [J]. *Nature*, 2000(6815）: 927-928.
[31] 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.
[32] 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.
[33] 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.
[34] 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.
[35] Jia, Y., Zhang, Y., Weiss, R., Wang, Q., Shen, J., Ren, F., Nguyen, P., Pang, R., Lopez Moreno, I. &Wu, Y. Transfer learning from speaker verification to multispeaker text-to-speech synthesis [J]. *Advances in Neural Information Processing Systems*, 2018.
[36] 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.
[37] 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.
[38] Jiang, X. &Pell, M.D. The sound of confidence and doubt [J]. *Speech Communication*, 2017: 106-126.
[39] Johnson, K. The ΔF method of vocal tract length normalisation for vowels [J]. *Laboratory Phonology*, 2020(1）.
[40] Kamide, Y. Learning individual talkers' structural preferences [J]. *Cognition*, 2012(1）: 66-71.
[41] Kharitonov, E., Copet, J., Lakhotia, K., Nguyen, T.A., Tomasello, P., Lee, A., Elkahky, A., Hsu, W.-N., Mohamed, A. &Dupoux, E. textless-lib: a Library for Textless Spoken Language Processing [J]. *arXiv preprint arXiv:2202.07359*, 2022.
[42] Kim, J., Toutios, A., Lee, S. &Narayanan, S.S. Vocal tract shaping of emotional speech [J]. *Computer Speech & Language*, 2020: 101100.
[43] Kinzler, K.D. Language as a social cue [J]. *Annual Review of Psychology*, 2021: 241-264.
[44] 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.
[45] Kreuk, F., Polyak, A., Copet, J., Kharitonov, E., Nguyen, T.-A., Rivière, M., Hsu, W.-N., Mohamed, A., Dupoux, E. &Adi, Y. Textless speech emotion conversion using decomposed and discrete representations [J]. *arXiv preprint arXiv:2111.07402*, 2021.
[46] Kroczek, L.O.H. &Gunter, T.C. The time course of speaker-specific language processing [J]. *Cortex*, 2021: 311-321.
[47] Kuhl, P.K. Who's talking? [J]. *Science*, 2011(6042）: 529-530.
[48] Künzel, H.J. How well does average fundamental frequency correlate with speaker height and weight? [J]. *Phonetica*, 1989(1-3）: 117-125.
[49] Labov, W. *The social stratification of English in New York city* [M]. Cambridge: Cambridge University Press, 2006.
[50] Lammert, A.C. &Narayanan, S.S. On short-time estimation of vocal tract length from formant frequencies

[J]. *PloS One*, 2015(7）: e0132193.
[51] Latinus, M. &Belin, P. Anti-voice adaptation suggests prototype-based coding of voice identity [J]. *Frontiers in Psychology*, 2011: 175.
[52] Lavan, N., Burston, L.F., Ladwa, P., Merriman, S.E., Knight, S. &McGettigan, C. Breaking voice identity perception: Expressive voices are more confusable for listeners [J]. *Quarterly Journal of Experimental Psychology*, 2019a(9）: 2240-2248.
[53] Lavan, N., Burton, A.M., Scott, S.K. &McGettigan, C. Flexible voices: Identity perception from variable vocal signals [J]. *Psychonomic Bulletin Review*, 2019b(1）: 90-102.
[54] Lavan, N., Knight, S. &McGettigan, C. Listeners form average-based representations of individual voice identities [J]. *Nature Communications*, 2019c(1）: 1-9.
[55] 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.
[56] Leipold, S., Abrams, D.A., Karraker, S. &Menon, V. Neural decoding of emotional prosody in voice-sensitive auditory cortex predicts social communication abilities in children [J]. *Cerebral Cortex*, 2022.
[57] Levi, S.V., Harel, D. &Schwartz, R.G. Language ability and the familiar talker advantage: Generalising to unfamiliar talkers is what matters [J]. *Journal of Speech, Language, Hearing Research*, 2019(5）: 1427-1436.
[58] 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.
[59] Matsumoto, H., Hiki, S., Sone, T. &Nimura, T. Multidimensional representation of personal quality of vowels and its acoustical correlates [J]. *IEEE Transactions on Audio Electroacoustics*, 1973(5）: 428-436.
[60] Nakagawa, S., Shikano, K. &Tohkura, Y.i. *Speech, hearing and neural network models* [M]. Amsterdam: IOS Press, 1995.
[61] Neumann, K. &Welzel, C. The importance of the voice in male-to-female transsexualism [J]. *Journal of Voice*, 2004(1）: 153-167.
[62] Oh, D., Buck, E.A. &Todorov, A. Revealing hidden gender biases in competence impressions of faces [J]. *Psychological Science*, 2019(1）: 65-79.
[63] 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.
[64] Peng, Z., Chen, J., Jin, L., Han, H., Dong, C., Guo, Y., Kong, X., Wan, G. &Wei, Z. 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.
[65] Pernet, C.R., McAleer, P., Latinus, M., Gorgolewski, K.J., Charest, I., Bestelmeyer, P.E.G., Watson, R.H., Fleming, D., Crabbe, F. &Valdes-Sosa, M. The human voice areas: Spatial organisation and inter-individual variability in temporal and extra-temporal cortices [J]. *Neuroimage*, 2015: 164-174.
[66] Perrachione, T.K., Del Tufo, S.N. &Gabrieli, J.D. Human voice recognition depends on language ability [J]. *Science*, 2011(6042）: 595-595.
[67] Perrachione, T.K. &Wong, P.C. Increased lefthemisphere contribution to native-versus foreign-language talker identification revealed by dichotic listening [Z]. *Poster presented at the 16th Meeting of the International Congress of Phonetic Sciences, Saarbrücken, Germany*. Citeseer, 2007.
[68] Pisanski, K., Anikin, A. &Reby, D. Static and dynamic formant scaling conveys body size and aggression [J]. *Royal Society Open Science*, 2021(1）: 211496.
[69] 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.
[70] 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.
[71] 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.
[72] 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.
[73] Rubin, D.L. Nonlanguage factors affecting undergraduates' judgments of nonnative English-speaking

teaching assistants [J]. *Research in Higher Education*, 1992(4）: 511-531.
[74] Schirmer, A. Is the voice an auditory face? An ALE meta-analysis comparing vocal and facial emotion processing [J]. *Social Cognitive and Affective Neuroscience*, 2018(1）: 1-13.
[75] 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.
[76] Shutts, K., Kinzler, K.D., McKee, C.B. &Spelke, E.S. Social information guides infants' selection of foods [J]. *Journal of Cognition Development*, 2009(1-2）: 1-17.
[77] 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.
[78] Soderstrom, M., Blossom, M., Foygel, R. &Morgan, J.L. Acoustical cues and grammatical units in speech to two preverbal infants [J]. *Journal of Child Language*, 2008(4）: 869-902.
[79] Sorokowski, P., Puts, D., Johnson, J., Żółkiewicz, O., Oleszkiewicz, A., Sorokowska, A., Kowal, M., Borkowska, B. &Pisanski, K. Voice of authority: professionals lower their vocal frequencies when giving expert advice [J]. *Journal of Nonverbal Behavior*, 2019(2）: 257-269.
[80] Stern, D.N., Spieker, S. &MacKain, K. Intonation contours as signals in maternal speech to prelinguistic infants [J]. *Developmental Psychology*, 1982(5）: 727.
[81] Tamagawa, R., Watson, C.I., Kuo, I.H., MacDonald, B.A. &Broadbent, E. The effects of synthesised voice accents on user perceptions of robots [J]. *International Journal of Social Robotics*, 2011(3）: 253-262.
[82] Tan, Z.-H. Vocal tract length perturbation for text-dependent speaker verification with autoregressive prediction coding [J]. *IEEE Signal Processing Letters*, 2021: 364-368.
[83] Tang, C., Hamilton, L.S. &Chang, E.F. Intonational speech prosody encoding in the human auditory cortex [J]. *Science*, 2017(6353）: 797-801.
[84] 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.
[85] Voigt, R., Jurafsky, D. &Sumner, M. Between-and within-speaker effects of bilingualism on F0 variation [Z]. *Interspeech*. San Francisco, The United States, 2016:1122-1126.
[86] von Kriegstein, K., Warren, J.D., Ives, D.T., Patterson, R.D. &Griffiths, T.D. Processing the acoustic effect of size in speech sounds [J]. *Neuroimage*, 2006(1）: 368-375.
[87] 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.
[88] 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.
[89] 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.
[90] Xu, M., Homae, F., Hashimoto, R.-i. &Hagiwara, H. Acoustic cues for the recognition of self-voice and other-voice [J]. *Frontiers in Psychology*, 2013: 735.
[91] Zhang, Y., Ding, Y., Huang, J., Zhou, W., Ling, Z., Hong, B. &Wang, X. Hierarchical cortical networks of "voice patches" for processing voices in human brain [J]. *Proceedings of the National Academy of Sciences*, 2021(52）: e2113887118.
[92] 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.
[93] 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.
[94] 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.
[95] 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.
[96] 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.
[97] 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.

[98] 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.
[99] 伍可, 陈杰, 李雯婕, 陈洁佳, 刘雷, 刘翠红. 人声加工的神经机制 [J]. 心理科学进展, 2020(5）: 752-765.
[100] 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.
[101] 周爱保, 胡砚冰, 周滢鑫, 李玉, 李文一, 张号博, 郭彦麟, 胡国庆. 听而不 “闻”? 人声失认症的神经机制 [J]. 心理科学进展, 2021(3）: 414.