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.

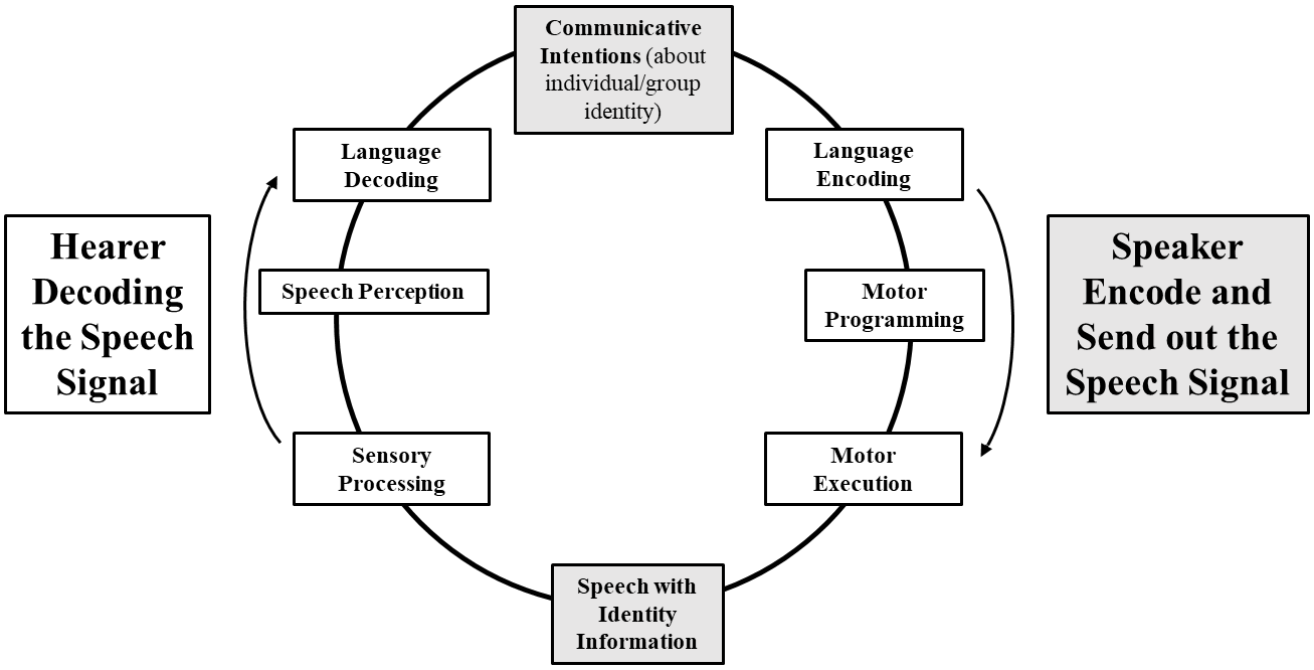

Figure 1. Speaker Identity Encoding and Decoding Model for Verbal Interaction Scenarios

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.

2023 年 7 月 外 国 语 July 2023
第 46 卷第 4 期 *Journal of Foreign Languages* Vol.46 No.4
文章编号: 1004 - 5139(2023)04 - 0068 - 14 中图分类号: H030 文献标识码: A

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[1,2]
(1. 上海外国语大学 语言研究院,上海 201620;
2. 上海外国语大学 语言科学与多语智能应用重点总实验室,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;在引入内/外群体概念后,进一步探讨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。基于以上提出言语互动场景下的言者身份编码及解码模型,并展望三个研究方向。

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

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

CHEN Wenjun[1], *HU Yanbing*[1], *JIANG Xiaoming*[1,2]
(Institute of Linguistics, Shanghai International Studies University, Shanghai 201620, China;
Key Laboratory of Language Science and Multilingual Intelligence Applications,
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 supports 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 in three directions.

Key words: speaker identification; vocal expression; communication intention; social grouping; social psycholinguistics

1. 引言

马看四蹄，人看四相。《红楼梦》中林黛玉能快速通过王熙凤的声音感知到王氏飞扬跋扈的性格和在贾府的显赫地位正是凭借了“音相”。言语交流中，人声不仅传递了语言信息，也包含了言者身份和情绪信息(Belin *et al.* 2004)。听者不仅可以从声音中听出对方“是谁”，也能对其“是什么样的人”形成大致印象。人声和人脸一样承载了身份信息，也被称为“听觉人脸”(Schirmer 2018)。言者身份包括性别、年龄、体型等信息，其由以基频和声道长度为主的语音信号组合编码(Lavan *et al.* 2019c)，而听者对其解码则主要依靠右侧颞上沟(right anterior superior temporal sulcus)(Formisano *et al.* 2008)。言者身份信息和言语中的语言信息乃至表示语用目的的重音强调等共享基频这样的语音信号(Frühholz & Schweinberger 2021; Tang *et al.* 2017)，即言者身份会随着发声的言语任务不断改变。然而，大量的言者身份研究没有考虑言语交互场景下的言者身份编码和解码的动态性，即对社会互动维度的关注远少于对认知心理维度的关注(束定芳、张立飞 2021)，故本文特别探讨了言语交际互动维度下言者身份和语言信息编码和解码之间的交互关系。

社会心理语言学语言观下，言语交际是一种有意识的言语活动，对其具体的语用模式和社会心理的言语机制研究需要整合跨学科的证据(王德春、孙汝建 1992a; 1992b)。束定芳(1992)列举了都柏林的教师可以通过学生言语中的语言线索(如不标准的发音)推断贫民学生家庭的社会地位，进而降低对学生的评价这一现象(即对言者群体身份的判断影响社会互动)。他还指出归因(causal attribution)以及集团特征(group distinctiveness)这两个概念以解释 Giles *et al.* (1991)的适应理论(accommodation theory)下的语言趋同现象(convergence)，进而指出言语交际中编码过程的社会心理作用以及解码过程中的社会心理因素是社会心理语言学的主要研究对象。束定芳(1992)在展望中呼吁语言学界探究语言变化、语言结构、语言风格以及集团语和社会心理之间的关系。因此有关言者身份的研究是一个社会心理语言学视角下的语言跨学科问题，需要综合来自心理语言学、社会语言学、言语交际科学、实验心理学、实验语用学和认知神经科学交叉学科的证据进行探讨。对言者身份编码和解码研究将促进对语言和社会心理之间的关系、人工智能克隆语音的应用、以及语言学习和跨语种加工等问题的理解。因此，本文在语言跨学科的视角下探讨言者个体和群体身份在动态的言语互动中的编码和解码机制，关注口语交际中，语言规则(音系结构和句法结构)和语言风格如何影响听者对言者身份的解码以及后续社会互动方案决策之间的关系。

经典的语言学理论没有考虑多模态互动场景下，例如，口语交流时声学信息的可变性，如何影响听者对言者个体和群体身份的认知。例如，Austin (1975: 100)的言语行为理论认为，言者会采用话语以实现特定的言语行为(如提出要求和下达命令)，听者会根据交际场景中的语境来理解言者的言语行为；此时，言者是谁并不影响听者对言者言语行为的理解。Grice (1975)认为言者字面意思以外还有暗示的更多内容，而听者会利用会话中的隐含意义来理解句子。这一理论表明，语境和非语言信息在理解话语方面起着关键作用，然而仍然没有强调言者个体内部身份。以上语言学理论是以语言本体结构(并未考虑多模态的互动)为中心的观点，即听者对话语的理解不会因为言者是谁而有所改变，大脑似乎不会因为言者身份不同而区别化地加工句子。然而，语音为载体的沟通互动场景下的心理语言学和神经语言学实验却表明，听者对言者的话语的加工会受到言者个体间和个体内身份差异的影响。例如，社会范畴下的地位高低不同的双方在进行社会互动时，大

脑神经活动指标会敏感于对敬语“您”和“你”的使用出现违反的条件(Jiang *et al.* 2013)。句法结构范畴下,德语中 SOV 和 OSV 句法结构的对立研究发现听者预期言者会讲简单的 SOV 句子,但实际听到言者讲复杂的 OSV 句子时,听者大脑会有增大的 P600 活动(Kroczek & Gunter 2021)。语用范畴下,阅读实验中,当言者是否采取常用的讽刺表达,也有类似的 P600 效应(Regel *et al.* 2010)。因此,从言者角度看,由音系结构和语言风格(侧重于语音层次的沟通)组成的语言规则会影响言者的言语产出,产出的差异具体体现在精密的声学分析上(例如,基频、声道长度参数、音强、时长、jitter、和 shimmer 等参数的组间差异上,此类的组可以是:自信、中性和怀疑的“知道感”驱使下的言语韵律差异(Jiang & Pell 2017))。同时听者会对言者产出的语音中的声学变化敏感。其中,基频和声道长度关键表征了言者个体与其他言者个体之间的身份差异;由此,个体敏感于声音中的言者身份变化。以上推论的一个证据来自探索听者解码不同口音英语使用者产出的语音中的自信水平的脑电研究,该研究发现,在 ERP 早期阶段听者就能加工音系结构和言者身份信息(Jiang *et al.* 2020)。因此,本文重点关注了言语沟通背景下个体对言者产出的言语的理解,侧重于听者如何解读言者的身份进而影响听者的发声策略和社会互动。

言者身份编码与解码的研究多关注个体在控制条件下(即用中性嗓音表情录制实验刺激)的发声所涉及的认知加工模式。基于人声线索的身份解码实验主要通过两个范式探究听者对不熟悉声音和熟悉声音的识别机制:说话人辨别(AX discrimination),即听者基于所听的两个句子判断陌生说话者是否为同一个人;再认(speaker identification),此范式源于司法实践中的嫌犯指认,即当事人听一个阵列的语音后调用记忆以指出嫌犯身份(Levi *et al.* 2019)。国内有相关综述:伍可等(2020)从感知的神经机制角度介绍了人声言语、情绪及身份加工所涉及的双通路模型、多阶段模型和整合模型;周爱保等(2021)区分了获得性和发展性人声失认症患者受损脑区的差异性。然而,这些综述还没有探讨基于社会群体之间的人声身份差异与加工机制的问题。事实上,个体还会在中性嗓音表情以外更复杂地发声。个体发声会随着讲话风格、环境和社会场景(例如,模仿他人)、所处认知发展阶段、情绪、生理和心理状态等产生内部变异(Lavan *et al.* 2019b),即听者从声音中感知言者“是谁”和“什么样的人”的结果不是恒定的。

在语言结构和语言风格与发声生理基础共同影响言者发声策略的基础上,本文回顾了言者的个体和群体身份编码与解码相关文献,提出了一个言语互动场景下言者身份编码及解码整合模型,并基于此进行了研究展望。

2. 言者个体身份的生理基础和语言声学编码

语言官能理论(Hauser *et al.* 2002)提示了言者个体身份与语言编码之间的关系。广义语言官能(broad language faculty, FLB)包括了言者个体身份编码的生理基础,声音身份编码和解码是跨物种的普遍能力;狭义语言官能(narrow language faculty, FLN)包括了语言结构的递归性,人类的言者身份编码因为语言的进化更加复杂。言者在言语计划阶段确定语言结构和语言风格后,通过调用发声生理基础执行特定的“发声策略”,基于此产出携带声音身份的语音。其中,语言结构包括音系结构(例如,由言者口音带来的音节特征偏好(Coupland 2007: 173))和句法结构(例如,言者偏好使用 SOV 或 OSV 结构(Kroczek & Gunter 2021));语言风格包括说话风格或特定的语用选择(例如,倾向于说讽刺的话(Regel *et al.* 2010)、风格变异(stylistic variant)带来的 [ɹ] 音卷舌程度(Labov 2006: 40 – 47)或性别二元的说话风格(Hogg 1985)等)。

2.1 声音身份的编码与识别是普遍能力

生命体的体型大小是声音身份编码的关键要素,比如野狼在捕猎或产仔时会通过嚎叫昭示领

地，对同类体型感知是许多物种社会组织中的普遍现象（Harrington & Mech 1979）。身份编码的后果与具有进化意义的繁殖后果密切相关。Reby & McComb（2003）分析了24头雄性红鹿的吼叫声、体重和繁殖成功率数据后发现基于共振峰间距计算的声道长度与红鹿体重正相关，且普通状态下吼叫对应的最大声道长度与繁殖成功率正相关。人类社会中也有类似的发现。Šebesta *et al.*（2019）邀请了84名来自巴西和68名来自捷克的异性恋参与者朗读短句和唱歌，并报告社会化性生活情况，发现女性短演讲中声道长度更短及歌唱中声道长度更长可以预测女性的性行为。由此可知，物种对同类身份的判断是一种普遍能力。

人类从声音中解码对方身份的能力先于语言交流出现。Polka *et al.*（2022）合成了共振峰间距较大的婴儿（F2-F1 = 3761）和间距较小的成人女性（F2-F1 = 2315）对应的元音/i/音频，发现平均年龄220天的婴儿对模拟婴儿发声状态的元音有偏好。而该能力与语言习得有密切关系：Fecher & Johnson（2019）给平均年龄136天的大部分时间仅曾暴露在英语环境下的婴儿呈现了由4名双语者（2名讲英语和波兰语；2名讲英语和西班牙语）女性录制的英语、波兰语和西班牙语句子；以注视时间为因变量、言者（异/同）和语言（母语/非母语）为主要拟合参数的混合线性模型，结果显示言者和语言均无主效应但其间有交互作用，这表明刺激是否为婴儿母语会调节婴儿听同一言者或不同言者音频时的注视时间。

基于其他物种与早期语言习得的研究表明，人类在进化过程中保留了从声音中识别对方的能力，且该能力早于语言交流出现，然而语言这一人类特有的现象使该能力较于其他物种更加复杂化。

2.2 言者个体身份编码与解码的特异性：语言音系规则的制约

听者能够在互动中出于交际功能的目的整合言者的身份信息（即，社会目的，social goal）以及话语中的内容信息，从而形成语言目的（linguistic goal）（Kuhl 2011），而正是听者对语言目的的解码使得人类言者识别不同于普通动物识别同类。Perrachione *et al.*（2011）发现损伤了英语音系规则的、由临床诊断判断为阅读障碍的群体①，在识别用母语英语和完全陌生的中文编码的言者身份时准确率相当，且均远低于健康人群对照组。从音系学的角度看，如果个体完全不懂一门语言，那么在准确听辨，并产出该语言的声音（sounds）和声音模式（sound patterns）时会表现出困难，因此将缺乏对该语言特定音系规则的了解（Goldsmith *et al.* 2014：319）。英语单语阅读障碍患者缺失对中文音系规则的了解，阅读障碍又使他们损伤了英语的音系规则，这导致其对母语的音系规则知识仅处于陌生语言水平，此损伤使他们在母语条件下言者身份识别的准确率和陌生语言水平的条件下相似。人类言者身份解码高度依赖听者的音系规则知识，正是对母语有更多音系知识导致了“语言熟悉效应（language familiarity effect）”：就算是听倒放而无语义通达的句子，单语听者会在母语条件下更准确识别言者身份（Fleming *et al.* 2014）。值得注意的是，Orena *et al.*（2015）发现加拿大蒙特利尔的英语单语成年人比美国康涅狄格州英语单语者能更快更准确地学习和识别法语使用者的言者身份，这表明被暴露于特定音系规则的使用场景也有助于语言熟悉效应的出现。

以上证据提示，人类听者对言者话语中身份和语音信息的整合加工与普通动物识别同类可能具有加工机制上的差异。

2.3 言者个体身份编码与解码的特异性：句法结构和语言风格的绑定

听者会将言者身份与特定句法结构绑定。Kroczek & Gunter（2021）首先训练被试暴露在不同

① 阅读障碍是学习障碍的一种，发生可能与遗传因素和大脑处理语言的区域出现个体差异有关；典型症状包括说话发育迟缓，记忆或命名字母、拼写错误，数学能力低下和学习困难等；其中，掌握语言音系规则是儿童学习识字拼读的重要基础。

句法结构分布的特定说话人的实验条件下(比如言者 A 讲 70% OSV 和 30% SOV 句子,而言者 B 相反),被试由此建立了对特定言者是"OSV 言者"或"OSV 言者"的预期;测试时被试听到"SOV 言者"讲 OSV 句子时,脑电成分中表现为大脑中后部的 P600 活动增强—表征了对特定言者句法结构预期的重分析或修复。类似的实验也发现听者能对言者高/低句法依附模式(syntactic attachment style)建立预期(Kamide 2012)。

更多研究也表明听者会将言者身份和特定的语言风格绑定。例如,Regel *et al.* (2010)采用与以上类似的实验设计让读者形成对两个人使用讽刺/字面不同话语风格的预期,当读者读到字面风格者的讽刺话语时,也会发现有类似的 P600 活动增强。

值得注意的是,Walker & Perry (2022)操纵了男女特定的韵律模式(如一名女性使用习惯的韵律 vs 模仿男性韵律)和语言风格(男/女性化的词汇使用),被试听到该女性用男性韵律讲话时,脑电活动诱发了反映表征语义不一致或不符合预期的 N400 活动增强,言者韵律和语言风格之间也存在交互作用,即当女性言者身份及女性韵律和女性语言风格一致情况,相比不一致情况会表现出不同的脑电活动。该研究表明社会范畴(男女二分的韵律)和语言本体范畴(性别特异的词汇)会共同影响言语产出的结果,而听者将基于两个范畴的加工与对言者身份的解码联系起来。

由上可推知,个体能在严格控制的实验室条件下学习不同言者句法结构和语言风格的使用倾向,而这一能力的产生与个体长期在自然互动场景下,不断将言者语言使用的特征与身份进行组合的语言交流实践密切相关。

2.4 基于言者间变异参数的个体身份编码与解码

语言是人类物种高级进化的产物,因此言者身份编码相较于其他物种更复杂。但人声身份的表征仍然依赖其进化过程中更底层的发声基础。Winters *et al.* (2008)招募了英语单语被试,在熟悉阶段给他们呈现了由英语 - 德语双语者录制的德语元 - 辅 - 元单词,听者需要将德语单词背后的身份和对应的名字联系起来;经由 8 轮次的熟悉 - 再熟悉 - 再认,被试能够将德语音频后的 10 个身份和相应的名字关联;最后在测试与泛化阶段给他们呈现由双语者录制的另一套英语单词,并要求他们指出该音频背后的言者身份;结果发现听德语单词熟悉言者身份后,听者能远高于机会水平分辨出双语言者英语发声背后的身份。这说明言语中有独立于音系结构之外的声学线索稳定地编码了人声身份,即基频(Xu *et al.* 2013)以及表征声道长度的共振峰间距(Johnson 2020),基频和声道长度交互地影响音色(timbre)并编码言者身份(von Kriegstein *et al.* 2006)。

关于发声生理基础,发声时气流自呼吸系统(肺部)经过气管传至发声系统(声带),而后经喉部到达构音系统(由口腔、咽腔和鼻腔组成的调音区,即声道),最终产出语音(Nakagawa *et al.* 1995:75 - 83)。首先,声带震动频率被表征为听觉可感的音高(pitch),即基频。基频是个体间的声门脉冲率(glottal-pulse rate),由于其声带构造差异造成的不同声学表征,尽管其与个体身材大小关系不紧密,但具有明显的性别差异:成人男性声带比女性的长约 60% 且更宽更粗,导致男性声门脉冲普遍比女性低,因此男性基频一般比女性低一个八度(Titze 1989)。其次,共振峰之间的距离与声道长度有统计关系,共振峰间距越小,声道长度越长,且声道长度与个体体型有直接关系(Johnson 2020),个体出生时声道长度约为 8 厘米,成年人的声道长度从 13 到 20 厘米不等(Lammert & Narayanan 2015)。同时,基频和声道长度交互影响言者身份编码(Lavan *et al.* 2019c)。

除了基于共振峰计算的基频和声道长度这样的频谱信息,还有更多参数也表征了言者身份差异。Chen *et al.* (2022)对男女各半的 300 名来自白人、黑人、亚洲人和拉美裔群体的言者发出的 10 秒长的"Aaaaah"就以下参数进行了分析:反映时间信息的均方根能量(RMS energy),反映频谱信息的频谱质心(spectral centroid)与声谱衰减(spectral roll-off),反映频谱包络形状的梅尔频率倒谱

系数(MFCCs),以及反映信号中信息量的熵相关值 - 谱熵(spectral entropy)、概率密度函数(PDF entropy)、排列熵(permutation entropy)、以及奇异值分解(SVD entropy)。结果发现种族间在以上参数上的差异不显著,然而男女(不区分种族)间在多个指标上有显著差异。

总的来说,生理结构导致的个体间基频和声道长度参数差异主要表征了言者间身份,且更广泛的时间、频谱、倒谱以及熵相关参数也可以反映性别群体身份。

2.5 基于言者内变异参数的个体身份编码与解码

人声身份识别领域多采用言者中性嗓音下的发声作为刺激以探索听者对声音身份的识别机制(Perrachione *et al.* 2011; Fleming *et al.* 2014)。然而所使用的语种、不同场景下的副语言信息表达的需要(比如,言者的自信和怀疑时的情态)使得言者个体内部有较大变异。例如,Voigt *et al.* (2016)分析了25名德法、20名德意双语者在轻松话题采访时的录音,发现德法双语女性说法语时使用的基频平均值更高,德意双语女性讲意大利语基频更低。在针对展现言者不同水平"知道感"的陈述语的分析中发现,当观察句首和句中成分时,自信的基频相较于不自信高,然而观察句尾成分时不自信的基频却更高(Jiang & Pell 2017),这对言者身份仅通过表征言者间变异的声学参数进行编码的观点形成了挑战,即听者解码言者身份时依赖的线索更复杂。

研究发现,听者具有适应言者身份内部变异的能力。不同言者有基于不同的原型的声音身份,这些身份分布在多维声音空间中(Latinus & Belin 2011)。基于此,Lavan *et al.* (2019c)通过调整半音的方式操纵了以基频和声道长度为二维空间上的身份,通过在以声道长度为横轴、基频为纵轴的空间中,向上/下移动1.6个半音的基频和向左/右移动2.36个半音的声道长度建立了4个独立于源声的声音身份(左下角的声音因不自然而被排除)。围绕3个新声音原型的中点,在声道长度左右2.25半音和基频上下3.6半音范围内,各自新建了距离原型较近的16个内周点以及较远的18个外周点;内/外周点反映了3个声音的内部变异。听者在训练阶段仅听过外周点上的声音,然而他们在测试阶段判断声音"新/旧(old/new)"时却报告曾听到过内周点的声音。这表明,听者对言者声音身份有原型的认识并可以适应言者身份的内部变异。

然而,听者适应言者身份内部变异的能力是有限的。Lavan *et al.* (2019a)将来自美剧《绝命毒师》的两位男主角的1.2秒到4秒长的情绪发声峰值标准化至0.400帕并在10kHz处进行低通滤波后,要求听者对经合成操作后音频背后的言者身份进行拖拽分类;结果发现,即使是熟悉该剧集的听者仍然很难将音频准确地归类为两位言者,而是归类为更多位言者。Xu & Armony (2021)准备了12名言者分别产出的4个句子(2种情绪韵律:害怕/中性* 2种语义内容),在听者熟悉阶段给他们呈现了其中6名言者的中性/害怕韵律下的各1个句子,而后在测试阶段播放所有12名录音者的48个句子并要求听者判断该声音身份是否出现过,结果发现听相同内容的句子时,如果韵律相同,被试准确率普遍高于80%,然而若韵律不匹配时其准确率仅在机会水平附近。因此,听者对言者内部变异的适应仅限于特定阈限内。

以上研究表明,言者身份会随使用的语言或具体情态而产生内部变异,而听者能够在一定阈限内适应这样的变异,将声音身份有所改变的言者身份归一化地识别为同一个人。

3. 言者群体身份渗透意图调节发声策略

社会心理学从原型理论的角度解释了社会群体的划分,即态度、行为、习俗等个体相关属性的模糊集合会在交流个体的心智表征中形成一个原型化的人类群体观念。该原型所代表的属性可最大限度地提高群体的实体性进而导致刻板印象的产生。其中,语言和言语风格是个体所处群体的身份象征之一(Hogg 2016),人们据此划分社会互动对象为内/外群体成员(Jiang *et al.* 2020);且

社会群体划分具有可渗透性，即成员可以改变其身份表征(Hogg 2016)。下文回顾了言者群体身份如何被编码及个体如何通过调节发声进行群体渗透，而后提出了一个基于群体互动视角的言者身份表征解码的整合性框架。

3.1 言者群体身份解码

言者所用语言是听者划分内/外群体的标准之一。肯尼亚人认为，使用斯瓦希里语和吉里阿马语不同地定义了语言使用者的自我、权利、权益及宗教(Kinzler 2021)。有关婴儿的实证研究为母语者的内群体偏好提供了印证。例如，12 个月大的婴儿更倾向于取用母语内群体者递过的食物(Shutts *et al.* 2009)；10 个月大的英语母语的婴儿会优先选择英语内群体讲解员展示过的玩具，且 2.5 岁左右的英/法单语儿童也更多把物品交给内群体母语者以进行游戏互动(Kinzler *et al.* 2012)。Begus *et al.* (2016) 给 11 个月左右的婴儿呈现了其母语内群体(英语)和外群体(西班牙语)女性言者指着婴儿不熟悉物品进行名词教学的视频，观察了婴儿的 3 – 5Hzθ 频段活动(θ 频段的神经振荡通常用以表征信息处理和学习，成年人的 θ 频段为 4 – 8Hz)，发现婴儿在内群体言者条件下有更强烈活动的 θ 振荡。这都表明听者对特定语种的语言结构敏感并会由此判断言者的群体身份，进而与不同言者群体差异化地开展社会互动。

口音规则是语言的一部分，言者的口音被听者用于群体划分。Rubin (1992) 给北美大学生播放了标准南方美音的授课音频并同时匹配了亚洲或白人面孔图片，发现当标准南方美音和亚洲面孔建立联系时，学生认为该言者有更重的非标准口音、更差的教学资质、讲课内容更难懂。Jiang *et al.* (2020) 招募了 44 名有相当法语能力且对澳大利亚英语口音了解的来自加拿大魁北克省的英语母语听者，被试在听了由加拿大英语口音、澳大利亚英语口音和带魁北克法语口音者录制的，带有自信或怀疑嗓音表情的音频后，进行了可信度评分，发现在自信条件下，加拿大英语口音和澳大利亚英语口音相比法语口音的音频听上去更可信；然而当句子带有怀疑嗓音时，加拿大英语口音相比澳大利亚英语或法语口音者听起来更不可信。Tamagawa *et al.* (2011) 提供了真人言者以外的证据：使用新西兰口音的被试听了基于英国、美国、和新西兰口音训练的、介绍同一款血压计的机器人的合成语音后，认为美国口音相比新西兰口音的合成声更像机器声，且比带有新西兰合成口音的机器人性能更差。此实验中机器人的口音属于语言结构中对不同的音系结构的表现。以上三项实验表明，言者对音系结构的选择关键地编码了他们被听者感知到的群体身份；而基于不同言者身份的感知听者会不同程度地改变互动决策。

性别二分态的言语风格也标志群体身份。男性言语的典型特征是使用粗俗语、讲话更加直率、声音低沉、咄咄逼人和显得更有权威，而女性的特征是速率和基频变化幅度大、讲话更加温和、开放、自我揭露和情绪化(Giles *et al.* 1983; Hogg 1985)。Martin & Slepian (2021) 提出的“大二”(big two) 模型认为，个体对他人特质判断和评价遵循了两个与性别角色有很大重叠的维度：a. 代理/男性气质，即自信、竞争、支配、独立、自我利益、目标追求；b. 社区/女性气质，即养育、温暖、表达、关注他人、社会导向。其中男性气质与他人感知到的“能力”密切相关，例如，男性化的面部特征使个体看起来更有能力(Oh *et al.* 2019)，声音更低的即更加男性化的女性被认为更具主导性，而女性化的声音则被与幼稚及性不成熟联系起来(Borkowska & Pawlowski 2011)。因此，听者基于男女言语风格差异区分其群体身份并将此划分结果与特定刻板印象联系起来。

言语交际中所使用的具体语言种类、所展现的口音以及男性化或女性化的程度都编码了言者的群体身份，听者基于此将互动对方识别为内群体或外群体成员并差异化地调整互动方案。值得注意的是，大量现有文献已经关注听者对言者的口音感知如何影响社会互动，但大多是基于真人之间的言语沟通，少有研究关注到近年新出现的人工智能克隆人声与语音合成技术。

3.2 言者身份编码的内/外群体渗透机制

言者的群体身份可以通过其基于意图调整发声策略而得到改变。交际调节理论(communication accommodation theory，CAT)提出，言者会在交际过程中为了缩短交际双方的社会距离、促进相互理解并提高交际效率而在口音、语速、音量、停顿与语言内容使用上与对方趋同(Coupland *et al*. 1988；Bernhold & Giles 2020)，这样的韵律层次的趋同现象在汉语自然会话场景下也被发现(夏志华、马秋武 2019)。典型的社会范畴下的趋同的一个例子是高低地位等级的双方互相采用对方常用的言语模式(束定芳 1992)。Sorokowski *et al*.(2019)录制了 27 名男性和 24 名女性在大学工作的科学家谈论日常话题(问路)和权威性话题“如何成为科学家且这样值得吗”的音频，发现男女言者在提供专业建议时基频都降低了，且女性(33Hz)比男性(14Hz)做出了更多降低的努力。Harrington *et al*.(2000)调查英女王伊丽莎白二世在 20 世纪 50 到 80 年代讲话音频中的元音后，发现其发声方式有向年轻群体和平民靠近的趋势。以上案例说明，言者为了显得亲切或更专业会策略性地改变对语言风格的选择，而这样的改变不仅体现在对发声基础的调控努力程度上，也体现于对文体变异(stylistic variant)方式的计算和执行。

Pisanski *et al*.(2021)指出，人类有声语言的发声复杂性(vocalic complexity)可能起源于动物界中这样一个普遍现象：物种会降低声道共鸣(即降低共振峰)以达到声音体型夸张。人类言语场景中，言者咄咄逼人时声道长度相较于中性发声更长(Pisanski *et al*. 2022)，言者开心时发声的声道长度比愤怒、悲伤或中性时更短(Kim *et al*. 2020)，且言者产出自信相比不自信录音时的基频更低(Jiang & Pell 2017)。这样的声学参数提示了，言者通过拉长或缩短声道以及其他调控生理基础的方式来编码自己身份；而这些发声策略的改变将直接影响听者感知到的言者身份。也就是说，人类在进化出语言系统后仍然保留了通过拉长/缩短声道、提高/降低基频改变声音体型的能力，并在具体语用场景中潜意识地进行了声音体型改变(如，咄咄逼人时使自己听起来体型更大)。由上可知，更具短时、动态性的副语言信息传递可能和长期稳定的语言结构、语言风格的形成有关。

本节表明，言者会基于特定交际目的，遵循特定的语言规则改变自己的言语发声方式进而使自己听起来属于特定群体。这一机制可能起源于动物界中普遍存在的体型夸张化现象，因而具有进化意义。然而，对此类发声调控的精细观察以及语言规则是否具有跨文化间一致性(考虑到语种多样性)仍然有待回答。同时，对语言规则习得或使用异常群体如何编码和解码言者身份的研究将助于理解声音中身份、语言、副语言信息间的关系。

3.3 言语社会互动场景下的言者身份编码和解码的理论框架

以上综述表明，言者会在交流意图下调整发声策略以影响听者得到的“言者是谁以及是什么样的人”的印象。互动场景下，听者接下话轮时将基于所整合的关于言者的身份和语言的信息调整言语产出。因此，本文也将整合三个有关模型：1) Belin *et al*.(2004)提出的人声加工模型，即人脑中的三条神经通路会在识别声音为人声后被分别激活，用以分别精细化地加工言语、情绪和身份信息；2) Braber *et al*.(2015:335)从听者言者话轮循环角度提出的人类言语沟通循环模型；3) Jiang *et al*.(2020)从声音身份和情绪加工时间进程角度提出的嗓音表情的认知处理模型，其中听者在人声处理的早期阶段会加工言者语音中的嗓音信息结构(vocal structure)，包括嗓音身份信息和嗓音言语结构(包括语言结构)。最终整合后形成社会互动场景下言者身份编码和解码框架(图 1)。

框架中两个基础元素是：(1)情绪信息，包括惊讶、高兴、生气、悲伤、害怕、厌恶、中性这样的基础情绪，也有表现“知道感”的自信嗓音信号(蒋晓鸣 2020)；(2)身份信息，由性别、年龄、受教育程度、吸引力、能力、群族等变量表征(Frühholz & Belin 2018)。

话轮伊始，在渗透到对方群体身份的交际意图的驱动下，言者选择听者特有的音系和句法结

构完成语言编码(如是否采用男性化的言语风格);在言语运动编码阶段,言者基于语言信息(如特定的句法结构与音节结构特征(Labov 2006:40–47))和副语言信息的声学表达规则(如是否降低基频以显得自己更自信(Jiang & Pell 2017))完成如何调用舌头、嘴唇和声带等发声基础的计划;在言语运动执行阶段,言者脑中发出神经信号控制发声的解剖基础完成言语发声,传出声波。

听者的听觉系统将机械波震动转为神经信号传递至听觉中枢,完成听觉信息接受。在言语感知阶段,听者在100毫秒左右进行了嗓音结构分析,同步加工了嗓音身份信息、情绪信息以及话语内容信息(理解表示句子功能的句法信息结构);在200毫秒左右,听者进行了嗓音重要性侦测(比较言者语气、与听者自身的口音相似度等)以决定注意力分配的多少;在250毫秒往后,听者进入了语言理解阶段,对模糊的语义再确认/消歧、基于身份信息进行语用推理并将身份信息整合到语境中。而后言者接过话轮并基于意图策略性地发声,话轮周而复始。

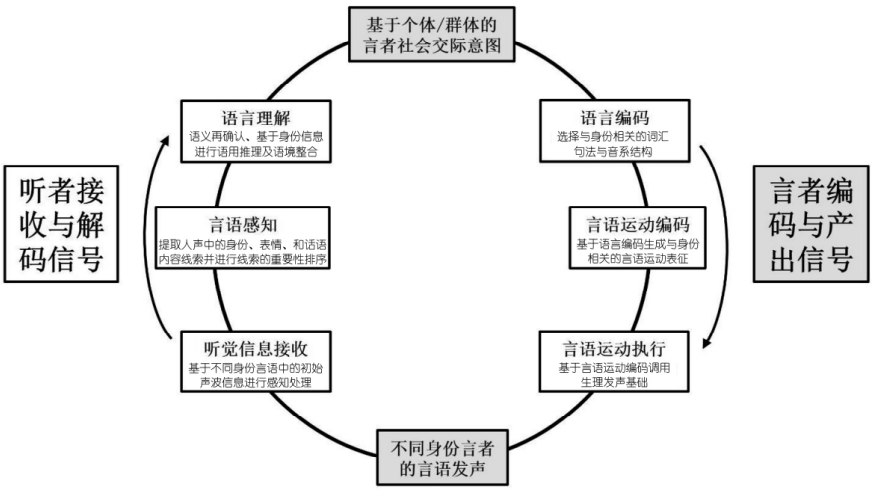

图1　言语互动场景下的言者身份编码及解码模型

4. 研究展望

基于以上回顾可知,言者个体和群体身份的编码与解码和语言和社会的具体维度交互影响。前文主要探讨了多个因素如何影响言者身份的表征,将来研究可以逆向地探索言者身份的表征和感知如何影响社会认知,例如,听者对言者情绪的编码和解码是否及如何受到言者身份的调节。除此之外,未来研究可以探讨1)什么指标可以表征言者调整发声运动以达到群体渗透目的时的努力;2)听者对人工智能克隆人声加工的脑机制及情态韵律线索对跨群体发声解码的调节作用;3)语言规则缺损者编码和解码言者身份的内部变异。

第一,个体如何基于语言规则调控发声策略尚需探讨。社会规范与行为习惯是个体通过言语的和非言语的社会交互习得的,其具有跨文化的共性,因此互动参与者可以跨文化范畴地理解言者的文化符号(蒋晓鸣2020)。那么,特定语用功能指导下的言语产出,比如自信、支配、顺从等社会态度的发声,是否可以通过对言者音频的声学参数和成像技术对发声模态进行跨文化的比较?即是否可以观察到自信发声拉长声道而怀疑缩减声道长度,而这样的模式是否有文化间一致性?可能的研究手段有声学参数分析、声门生理运动测量技术和磁共振成像技术。

第二,人类对克隆人声的认知加工机制亟待探索。首先,技术已经可以基于几秒长的音频克隆真人的声音模型,并以此伪造个体的言者身份(Jia *et al.* 2018),那么听者在感知真人声源和其对应的克隆声时,将如何定义克隆声的内/外群体身份并进行社会互动决策呢?例如,Pernet *et al.* (2015)发现

颞叶声音区有三个斑块选择性地对人声敏感，且 Zhang *et al.* (2021) 通过皮层脑电发现，癫痫病人左侧前颞叶的特定电极点仅对母语人声有反应，更值得注意的是，Di Cesare *et al.* (2022) 给听者呈现了传递社会意图的词汇“hello”后，真人声较于中性嗓音的合成声特异地激活了听者的背侧－中央脑岛区域。那么 是否有特定神经相关物表征个体不同地加工真人和其对应的克隆声呢？其次，语音合成技术可以产出富含哭、笑、打哈欠等表达性发声的音频(Kharitonov *et al.* 2022)，甚至让两个合成声模型进行自发而实时的、带有自然的重叠和停顿的闲聊(Kreuk *et al.* 2021)；如果以上技术和克隆人声相结合，即克隆声变得更加“拟人化”时，听者对克隆声的群体归类会受到改变吗？将来的研究可结合电生理和磁共振成像从时间进程和空间维度上探索听者对不同条件下言者身份的差异化感知。同时，经典的语音适应范式(adaptation paradigm) 的理论基础是听者特定神经元反应会随暴露在同一类刺激下的时间增加而减弱，若刺激特征改变，神经元反应将更加强烈，因此该范式可被用于进一步探索语音韵律如何调节人类听众对真人和克隆声的认知(Grill-Spector *et al.* 2006)。再次，声纹识别技术将如何应对克隆人声带来的言者身份危机呢？尽管已知升级声道长度归一化算法可提升人声身份识别产品的准确率(Tan 2021)，但前文提及的更广泛的频谱、倒谱等参数在改变识别准确率上的作用尚不清晰。此外，在人工智能语音服务声音在听感上更加“拟人化”的基础上，其人机交互时所产出的言语内容在 ChatGPT 这样的大语言模型支持下可能会更加“专家化”和“定制化”，此时用户个体对智能服务的感知和态度是否会受到言语内容的调节？

第三，语言规则缺损者的言者产出和听者感知机制也需要研究。首先，跨性别者为了使发声达到渗透至与生理性别对立的性别群体，需要通过(辅以视觉的) 口腔共鸣语音治疗习得对应规则或接受环甲切除术(Neumann & Welzel 2004; Hardy *et al.* 2016; Dahl & Mahler 2020)，这些干预措施会影响语音规则的表征与应用，这将如何启示对性别认同产生危机群体的矫正方案实施？其次，在社会交流/互动方面存在持续性障碍的自闭症谱系人群的“社交脑”存在异常，这些病人通常表现出语用规则运用的障碍，这与他们的额下回(inferior frontal gyrus，IFG)、颞上回(superior temporal gyrus，STG) 和杏仁核(amygdala) 过度激活密切关联(Peng *et al.* 2020)。然而颞上回(STG) 参与了陌生嗓音身份加工，额下回(IFG) 通过与前颞上沟(anterior superior temporal sulcus，anterior STS) 的功能连接参与了熟悉嗓音身份加工(伍可等 2020)，且双侧中、后部颞上沟(posterior STS/ superior STS) 表征了个体对声音中情感韵律信息的解码(Leipold *et al.* 2022)；由此，自闭症谱系群体在整合真人和克隆言者身份与典型被试相比是否存在差异？引入韵律信息这一言者内发声变量后将如何调节行为结果和相对应的神经相关物表征？规则缺损导致的行为后果差异将进一步检验语言规则在言者身份编解码中的因果机制。

参考文献:

[1] Austin，J. L. *How to Do Things with Words* [M]. Oxford: Oxford university press，1975.

[2] Begus，K.，Gliga，T. & V. Southgate. Infants' preferences for native speakers are associated with an expectation of information [J]. *Proceedings of the National Academy of Sciences*，2016，113(44): 12397－12402.

[3] Belin，P.，Fecteau，S. & C. Bedard. Thinking the voice: neural correlates of voice perception [J]. *Trends in Cognitive Sciences*，2004，8(3): 129－135.

[4] Bernhold，Q. S. & H. Giles. Vocal accommodation and mimicry [J]. *Journal of Nonverbal Behavior*，2020，44(1): 41－62.

[5] Borkowska，B. & B. Pawlowski. Female voice frequency in the context of dominance and attractiveness perception [J]. *Animal Behaviour*，2011，82(1): 55－59.

[6] Braber，N.，Cummings，L. & L. Morrish. *Exploring Language and Linguistics* [M]. Cambridge: Cambridge

University Press, 2015.
[7] Chen, X., Li, Z., Setlur, S. & W. Xu. Exploring racial and gender disparities in voice biometrics [J]. *Scientific Reports*, 2022, 12(1): 1 – 12.
[8] Coupland, N. *Style: Language Variation and Identity* [M]. Cambridge: Cambridge University Press, 2007.
[9] Dahl, K. L. & L. A. Mahler. Acoustic features of transfeminine voices and perceptions of voice femininity [J]. *Journal of Voice*, 2020, 34(6): 961 – e919.
[10] Di Cesare, G., Cuccio, V., Marchi, M., Sciutti, A. & G. Rizzolatti. Communicative and affective components in processing auditory vitality forms: An fMRI study [J]. *Cerebral Cortex*, 2022, 32(5): 909 – 918.
[11] Fecher, N. & E. K. Johnson. By 4.5 months, linguistic experience already affects infants' talker processing abilities [J]. *Child Development*, 2019, 90(5): 1535 – 1543.
[12] Fleming, D., Giordano, B. L., Caldara, R. & P. Belin. A language-familiarity effect for speaker discrimination without comprehension [J]. *Proceedings of the National Academy of Sciences*, 2014, 111(38): 13795 – 13798.
[13] Formisano, E., De Martino, F., Bonte, M. & R. Goebel. "Who" is saying" what"? Brain-based decoding of human voice and speech [J]. *Science*, 2008, 322(5903): 970 – 973.
[14] Frühholz, S. & P. Belin. *The Oxford Handbook of Voice Perception* [M]. Oxford: Oxford University Press, 2018.
[15] Frühholz, S. & S. R. Schweinberger. Nonverbal auditory communication-evidence for integrated neural systems for voice signal production and perception [J]. *Progress in Neurobiology*, 2021, 199: 101948.
[16] Giles, H., Coupland, J., Coupland, N. & K. Oatley. *Contexts of accommodation: Developments in applied sociolinguistics*: Cambridge University Press, 1991.
[17] Giles, H., Scholes, J. & L. Young. Stereotypes of male and female speech: A British study [J]. *Central States Speech Journal*, 1983, (4): 255 – 256.
[18] Goldsmith, J. A., Riggle, J. & C. L. Alan. *The Handbook of Phonological Theory* [M]. New York: John Wiley & Sons, 2014.
[19] Grice, H. P. Logic and conversation [M] //Peter Cole, Morgan, J. L., *Speech acts*. New York: Academic Press, 1975, 41 – 58.
[20] Grill-Spector, K., Henson, R. & A. Martin. Repetition and the brain: neural models of stimulus-specific effects [J]. *Trends in Cognitive Sciences*, 2006, 10(1): 14 – 23.
[21] Hardy, T. L. D., Boliek, C. A., Wells, K., Dearden, C., Zalmanowitz, C. & J. M. Rieger. 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, 25(2): 125 – 137.
[22] Harrington, F. H. & L. D. Mech. Wolf howling and its role in territory maintenance [J]. *Behaviour*, 1979, 68(3 – 4): 207 – 249.
[23] Harrington, J., Palethorpe, S. & C. I. Watson. Does the Queen speak the Queen's English? [J]. *Nature*, 2000, 408(6815): 927 – 928.
[24] Hauser, M. D., Chomsky, N. & W. T. Fitch. The faculty of language: what is it, who has it, and how did it evolve? [J]. *Science*, 2002, 298(5598): 1569 – 1579.
[25] 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, 4(2): 99 – 112.
[26] 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, 2016, 3 – 17.
[27] Jia, Y., Zhang, Y., Weiss, R., Wang, Q., Shen, J., Ren, F., Nguyen, P., Pang, R., Lopez Moreno, I. & Y. Wu. Transfer learning from speaker verification to multispeaker text-to-speech synthesis [J]. *Advances in Neural Information Processing Systems*, 2018, 31.

[28] Jiang, X., Gossack-Keenan, K. & M. D. Pell. 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, 73(1): 55 – 79.

[29] Jiang, X., Li, Y. & X. Zhou. 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, 51(11): 2210 – 2223.

[30] Jiang, X. & M. D. Pell. The sound of confidence and doubt [J]. *Speech Communication*, 2017, 88: 106 – 126.

[31] Johnson, K. The ΔF method of vocal tract length normalization for vowels [J]. *Laboratory Phonology*, 2020, 11(1): 10.

[32] Kamide, Y. Learning individual talkers' structural preferences [J]. *Cognition*, 2012, 124(1): 66 – 71.

[33] Kharitonov, E., Copet, J., Lakhotia, K., Nguyen, T. A., Tomasello, P., Lee, A., Elkahky, A., Hsu, W. – N., Mohamed, A. & E. Dupoux. textless-lib: a Library for Textless Spoken Language Processing [J]. 2022, *arXiv preprint arXiv*: 2202.07359.

[34] Kim, J., Toutios, A., Lee, S. & S. S. Narayanan. Vocal tract shaping of emotional speech [J]. *Computer Speech & Language*, 2020, 64: 101100.

[35] Kinzler, K. D. Language as a social cue [J]. *Annual Review of Psychology*, 2021, 72: 241 – 264.

[36] Kinzler, K. D., Dupoux, E. & E. S. Spelke. 'Native' objects and collaborators: Infants' object choices and acts of giving reflect favor for native over foreign speakers [J]. *Journal of Cognition Development*, 2012, 13(1): 67 – 81.

[37] Kreuk, F., Polyak, A., Copet, J., Kharitonov, E., Nguyen, T. –A., Rivière, M., Hsu, W. –N., Mohamed, A., Dupoux, E. & Y. Adi. Textless speech emotion conversion using decomposed and discrete representations [J]. 2021, *arXiv preprint arXiv*: 2111.07402.

[38] Kroczek, L. O. H. & T. C. Gunter. The time course of speaker-specific language processing [J]. *Cortex*, 2021, 141, 311 – 321.

[39] Kuhl, P. K. Who's talking? [J]. *Science*, 2011, 333(6042): 529 – 530.

[40] Labov, W. *The Social Stratification of English in New York City* [M]. Cambridge: Cambridge University Press, 2006.

[41] Lammert, A. C. & S. S. Narayanan. On short-time estimation of vocal tract length from formant frequencies [J]. *PloS One*, 2015, 10(7): e0132193.

[42] Latinus, M. & P. Belin. Anti-voice adaptation suggests prototype-based coding of voice identity [J]. *Frontiers in Psychology*, 2011, 2, 175.

[43] Lavan, N., Burston, L. F., Ladwa, P., Merriman, S. E., Knight, S. & C. McGettigan. Breaking voice identity perception: Expressive voices are more confusable for listeners [J]. *Quarterly Journal of Experimental Psychology*, 2019a, 72(9): 2240 – 2248.

[44] Lavan, N., Burton, A. M., Scott, S. K. & C. McGettigan. Flexible voices: Identity perception from variable vocal signals [J]. *Psychonomic Bulletin Review*, 2019b, 26(1): 90 – 102.

[45] Lavan, N., Knight, S. & C. McGettigan. Listeners form average-based representations of individual voice identities [J]. *Nature Communications*, 2019c, 10(1): 1 – 9.

[46] Leipold, S., Abrams, D. A., Karraker, S. & V. Menon. Neural decoding of emotional prosody in voice-sensitive auditory cortex predicts social communication abilities in children [J]. *Cerebral Cortex*, 2023, 33(3): 709 – 728.

[47] Levi, S. V., Harel, D. & R. G. Schwartz. Language ability and the familiar talker advantage: Generalizing to unfamiliar talkers is what matters [J]. *Journal of Speech, Language, Hearing Research*, 2019, 62(5): 1427 – 1436.

[48] Martin, A. E. & M. L. Slepian. The primacy of gender: Gendered cognition underlies the Big Two dimensions of social cognition [J]. *Perspectives on Psychological Science*, 2021, 16(6): 1143 – 1158.

[49] Nakagawa, S., Shikano, K. & Y. I. Tohkura. *Speech, Hearing and Neural Network Models* [M]. Amsterdam:

IOS Press, 1995.

[50] Neumann, K. & C. Welzel. The importance of the voice in male-to-female transsexualism [J]. *Journal of Voice*, 2004, 18(1): 153 – 167.

[51] Oh, D., Buck, E. A. & A. Todorov. Revealing hidden gender biases in competence impressions of faces [J]. *Psychological Science*, 2019, 30(1): 65 – 79.

[52] Orena, A. J., Theodore, R. M. & L. Polka. Language exposure facilitates talker learning prior to language comprehension, even in adults [J]. *Cognition*, 2015, 143: 36 – 40.

[53] Peng, Z., Chen, J., Jin, L., Han, H., Dong, C., Guo, Y., Kong, X., Wan, G. & Z. Wei. 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, 298: 111063.

[54] Pernet, C. R., McAleer, P., Latinus, M., Gorgolewski, K. J., Charest, I., Bestelmeyer, P. E. G., Watson, R. H., Fleming, D., Crabbe, F. & M. Valdes-Sosa. The human voice areas: Spatial organization and inter-individual variability in temporal and extra-temporal cortices [J]. *Neuroimage*, 2015, 119: 164 – 174.

[55] Perrachione, T. K., Del Tufo, S. N. & J. D. Gabrieli. Human voice recognition depends on language ability [J]. *Science*, 2011, 333(6042): 595 – 595.

[56] Pisanski, K., Anikin, A. & D. Reby. Static and dynamic formant scaling conveys body size and aggression [J]. *Royal Society Open Science*, 2021, 9(1): 211496.

[57] Pisanski, K., Anikin, A. & D. Reby. Vocal size exaggeration may have contributed to the origins of vocalic complexity [J]. *Philosophical Transactions of the Royal Society B*, 2022, 377(1841): 20200401.

[58] Polka, L., Masapollo, M. & L. Ménard. Setting the stage for speech production: Infants prefer listening to speech sounds with infant vocal resonances [J]. *Journal of Speech, Language*, 2022, 65(1): 109 – 120.

[59] Reby, D. & K. McComb. Anatomical constraints generate honesty: acoustic cues to age and weight in the roars of red deer stags [J]. *Animal Behaviour*, 2003, 65(3): 519 – 530.

[60] Regel, S., Coulson, S. & T. C. Gunter. The communicative style of a speaker can affect language comprehension? ERP evidence from the comprehension of irony [J]. *Brain Research*, 2010, 1311: 121 – 135.

[61] Rubin, D. L. Nonlanguage factors affecting undergraduates' judgments of nonnative English-speaking teaching assistants [J]. *Research in Higher Education*, 1992, 33(4): 511 – 531.

[62] 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, 13(1): 1 – 13.

[63] Šebesta, P., Mendes, F. D. C. & K. J. Pereira. 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, 10: 2029.

[64] Shutts, K., Kinzler, K. D., McKee, C. B. & E. S. Spelke. Social information guides infants' selection of foods [J]. *Journal of Cognition Development*, 2009, 10(1 – 2): 1 – 17.

[65] Sorokowski, P., Puts, D., Johnson, J., Żółkiewicz, O., Oleszkiewicz, A., Sorokowska, A., Kowal, M., Borkowska, B. & K. Pisanski. Voice of authority: professionals lower their vocal frequencies when giving expert advice [J]. *Journal of Nonverbal Behavior*, 2019, 43(2): 257 – 269.

[66] Tamagawa, R., Watson, C. I., Kuo, I. H., MacDonald, B. A. & E. Broadbent. The effects of synthesized voice accents on user perceptions of robots [J]. *International Journal of Social Robotics*, 2011, (3): 253 – 262.

[67] Tan, Z. -H. Vocal tract length perturbation for text-dependent speaker verification with autoregressive prediction coding [J]. *IEEE Signal Processing Letters*, 2021, 28: 364 – 368.

[68] Tang, C., Hamilton, L. S. & E. F. Chang. Intonational speech prosody encoding in the human auditory cortex [J]. *Science*, 2017, 357(6353): 797 – 801.

[69] Titze, I. R. Physiologic and acoustic differences between male and female voices [J]. *The Journal of the Acoustical Society of America*, 1989, 85(4): 1699 – 1707.

[70] Voigt, R., Jurafsky, D. & M. Sumner. Between-and within-speaker effects of bilingualism on F0 variation [Z]. *Interspeech*. San Francisco, The United States, 2016: 1122 – 1126.

[71] von Kriegstein, K., Warren, J. D., Ives, D. T., Patterson, R. D. & T. D. Griffiths. Processing the acoustic effect of size in speech sounds [J]. *Neuroimage*, 2006, 32(1): 368 – 375.

[72] Walker, M. & C. Perry. 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, 37(1): 1 – 21.

[73] Winters, S. J., Levi, S. V. & D. B. Pisoni. Identification and discrimination of bilingual talkers across languages [J]. *The Journal of the Acoustical Society of America*, 2008, 123(6): 4524 – 4538.

[74] Xu, H. & J. L. Armony. Influence of emotional prosody, content, and repetition on memory recognition of speaker identity [J]. *Quarterly Journal of Experimental Psychology*, 2021, 74(7): 1185 – 1201.

[75] Xu, M., Homae, F., Hashimoto, R.-i. & H. Hagiwara. Acoustic cues for the recognition of self-voice and other-voice [J]. *Frontiers in Psychology*, 2013, (4): 735.

[76] Zhang, Y., Ding, Y., Huang, J., Zhou, W., Ling, Z., Hong, B. & X. Wang. Hierarchical cortical networks of "voice patches" for processing voices in human brain [J]. *Proceedings of the National Academy of Sciences*, 2021, 118(52): e2113887118.

[77] 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, 31(1): 116 – 124.

[78] 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, (3): 10 – 14.

[79] 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, (3): 420 – 429.

[80] 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, (4): 3 – 7.

[81] 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, (3): 3 – 9.

[82] 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, 28(5): 752 – 765.

[83] 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.

[84] 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, 29(3): 414.

基金项目: 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)

收稿日期: 2022 – 09 – 04

作者简介: 陈文均(1999 –),男,四川遂宁人,硕士研究生。研究方向: 心理与神经语言学、嗓音编码与解码。

胡砚冰(1996 –),男,甘肃天水人,博士研究生。研究方向: 心理与神经语言学、嗓音表情解码与嗓音产出。

蒋晓鸣(通讯作者)(1983 –),男,上海人,博士,教授,上海市曙光学者。研究方向: 心理与神经语言学、实验语言学、言语交际与言语障碍、嗓音编码与解码、神经语用学。