Linguistic Research 43(1): 105-140
DOI: 10.17250/khisli.43.1.202603.004

Does AI reflect sound change? A comparison of human and TTS production of Korean stops*

Yein Kang** · **Jonny Jungyun Kim***
(Pusan National University)

Kang, Yein and Jonny Jungyun Kim. 2026. Does AI reflect sound change? A comparison of human and TTS production of Korean stops. *Linguistic Research* 43(1): 105-140. This study presents an acoustic analysis of speech generated by 60 commercial text-to-speech (TTS) voices to examine whether socially indexed phonetic variation is reflected in artificial speech systems. Focusing on cross-generational variation in phrase-initial Korean stops, we investigated the extent to which voices labeled as "younger" across three TTS platforms approximate the ongoing merger in voice onset time (VOT) between aspirated and lenis stops, as well as a compensatory cue-shift toward fundamental frequency (F0). Younger TTS voices generally showed reduced reliance on VOT cues: for example, IP-initial aspirated and lenis stops differed by 12ms in younger TTS voices, compared to 21ms in older TTS voices. However, the magnitude of this merger in younger TTS voices differed substantially from that observed in younger human speakers, who exhibited an almost complete VOT merger (e.g., a 1ms difference, IP-initially). In contrast, F0 patterns in TTS speech broadly aligned with human speech. The results suggest a potential link between socially indexed phonetic realization and naturalness fidelity in AI-generated speech. We further outline future directions for examining AI speech from a sociophonetic perspective, focusing on fundamental differences between human and machine learning mechanisms. **(Pusan National University)**

Keywords TTS systems, speech naturalness, sociophonetic indexicality, speaker agency, Korean stops, sound change

* This work was supported by a Journal Research Promotion (2025) for Graduate Students of Pusan National University. Portions of this work were presented at the 50th Anniversary International Conference of the Linguistic Society of Korea (November 1, 2025). The authors acknowledge the use of ChatGPT (OpenAI) for language editing and R coding.
** First Author
*** Corresponding author

1. Introduction

Recent advancements in AI speech synthesis have achieved remarkable improvements in speech naturalness. The quality of speech generated by text-to-speech (TTS) engines is often evaluated using the Mean Opinion Score (MOS), which is estimated by either human listeners' subjective ratings on predefined scales (Viswanathan and Viswanathan 2005; Minixhofer et al. 2025), or by prediction models (Lo et al. 2019; Kondo et al. 2025). However, the quality of AI speech is now approaching a ceiling, to the extent that its naturalness cannot be suitably evaluated relying on a fixed set of judgment criteria (Le Maguer et al. 2024; Minixhofer et al. 2025; Wang et al. 2025). We have entered an era in which machines produce language so human-like that alternative evaluation paradigms, such as Turing-style tests, are required to test whether we can distinguish them from us (Varadhan et al. 2025; Wang et al. 2025).

Nevertheless, in many cases AI-generated speech remains perceptually distinguishable from human speech in ways that are not adequately captured by either MOS or Turing-style tests. While speech prosody has often been identified as a primary source of this discrepancy (Hu et al. 2024; Chan and Kuang 2025), much less attention has been paid to social aspects of segmental realizations as a contributor to perceived naturalness in TTS speech.

This study focuses on the extent to which sociophonetic indexicality (Eckert 2008) is reflected in commercial TTS systems for Korean. Phonetic forms systematically covary with speakers' social characteristics (Labov 2001). Through repeated interplay between cognitive learning mechanisms and socially structured experience, indexical links between phonetic variants and social information become entrenched in long-term memory over the course of language use (Hay and Drager 2007; Foulkes and Hay 2015). Consequently, socially meaningful phonetic variation plays a central role not only in the perception of speakers' social attributes (Fellowes et al. 1997; Remez et al. 1997; Campbell-Kibler 2007), but also in core speech processing functions, including phonetic cue weighting (So and Kim 2024; Zhang 2025), segmental categorization (Strand and Johnson 1996; Johnson et al. 1999; Hay et al. 2006; Hay and Drager 2010), lexical access (Walker and Hay 2011; Kim 2016; Kim and Drager 2017, 2018), and sentence processing (Yi et al. 2013; Babel and Russell 2015). In this vein, the role of sociophonetically conditioned cognitive links in linguistic processing is widely recognized in models of human speech perception and production (Johnson

1997; Pierrehumbert 2001; Kleinschmidt and Jaeger 2015; Drager and Kirtley 2016; Hay and Foulkes 2016; Hay 2018).

Trained on vast corpora of human speech, TTS models can generate output that closely approximates human-like fluency and intelligibility, even emulating voices across broad social categories, such as age, gender, or regional background. However, it remains unclear whether they can exploit social-phonetic covariation at a more fine-grained level to imitate socially indexed phonetic behaviors embedded in human speech. To what degree, for instance, can TTS reproduce cross-generational variation that underlies the production of familiar-sounding voices in everyday interaction? If TTS systems remain insensitive to these socially meaningful dimensions of variation, this may lead us to a broader question for future research: do essential aspects of human language—those grounded in social experience and indexical meaning—lie beyond what can be captured through large-scale acoustic learning alone?

2. Background

2.1 Speaker agency

There remains a consistent, and perhaps telling, divergence between how humans and machines absorb change-in-progress phenomena. Unlike AI models, language users do not acquire variation merely by tracking statistical distributions. Rather, sound change in human language is closely tied to social meaning and identity. It often begins in specific demographic groups and spreads as individuals orient toward or away from those groups through social learning (Labov 1963, 1972; Trudgill 1972; Eckert 2000; Foulkes and Hay 2015). In this way, sociolinguistic variants serve not just communicative but also indexical functions, either explicitly or implicitly signaling group membership, social stance, or identity (Drager and Kirtley 2016). What is noteworthy here is that human speakers do not passively undergo linguistic change simply because they belong to a social group. Rather, they actively engage in linguistically coded socialization to affiliate themselves with particular groups (or communities of practice) within their social network (Milroy 1978; Bucholtz 1999; Eckert 2008).

With such socially motivated agency, speakers possess and dynamically exploit

the mutability of indexical signs in linguistic forms (Eckert 2008) and actively construct their own social identity by positioning themselves from the interlocutors during the conversation (Bucholtz 1999; Campbell-Kibler 2007; Podesva 2007; Love and Walker 2013; Lev-Ari and Pepperkemp 2014; Sneller 2024). For example, Campbell-Kibler (2007) demonstrated that the English -ing variable exhibits extensive social indexicality: its [ɪn] and [ɪŋ] variants are associated with a wide range of social meanings, including formality, region, urbanity, sexuality, education, and social class. She also showed how speakers use the [ɪn] form in complex, context-sensitive ways during conversation, reflecting their multi-faceted identity.

With such communicatively driven agency, humans often treat AI as socially situated partners, despite the absence of genuine social intentions on the part of the artificial interlocutor. Studies on phonetic accommodation during human-machine interactions (e.g., Zellou et al. 2021; see Székely et al. 2025 for a review) found that humans aligned phonetic forms with those of AI talkers in shadowing and interactive tasks—as they do in human-to–human interaction—even when the interlocutor was explicitly identified as a device. In social interaction theories, this tendency reflects the extent to which people spontaneously apply social-cognitive frameworks developed for human interaction to artificial agents (Nass et al. 1994).

With such cognitively integrated agency, listeners unwittingly interpret acoustic input in accordance with perceived talker information during speech perception. As an example of the sociophonetic variable to be tested in this study (see Section 2.2), So and Kim (2024) found that when hearing Korean stops, listeners shifted their perceptual weight of VOT (used as a primary cue by older speakers) and F0 (used by younger speakers) as a function of talker age. Kim and Drager (2017) further showed that, while hearing a talker with age-neutral voice, stops with F0-based cues (a young guise) or with VOT-based cues (an old guise) facilitated subsequent recognition of age-congruent Korean words (e.g., a young-associated word /k*uldʒɛm/, *꿀잼*; an old-associated word /ʌmʌm/, *어멈*). These results suggest that multiple phonetic cues are organized in line with age-related social information within cognitive representations of phonemes and words.

AI speech models, in contrast, have little access to any of such agency. They do not possess social goals, nor do they try to infer socio-indexical meanings of a phonetic variant. Their phonetic approximations are instead driven purely by statistical covariations between phonetic forms and metadata such as speaker age or gender.

This apparent gap highlights a potential limitation of AI-synthesized speech. It reproduces observable patterns without actually participating in social interactions that drive linguistic changes. This may suggest a fundamental difference in how humans and machines embrace language as it continually evolves.

2.2 Ongoing sound change in Korean stops

This study chose the ongoing phonetic-level change in Korean stops as the key variable for investigation, as it has been extensively documented as a robust age-related pattern across Korean speakers. A substantial body of corpus-based or experimental research has shown systematic variation in the realization of Korean stops as a function of speaker age (Silva 1992, 2002, 2006; Kim et al. 2002; Wright 2007; Kang and Guion 2008; Kang 2010; Oh 2011; Kang 2014; Kim 2014; Bang et al. 2018; Oh et al. 2018; Choi et al. 2020; Kim 2024; see also Lee et al. 2020 for a recent review). Evidence further suggests that this change originated from Seoul Dialect since the 1990s and is now spreading to other regional varieties (Ahn 2017; Schertz et al. 2019; Kang et al. 2024), or is subject to dialect-specific constraints—i.e., the pitch accent system in Gyeongsang Dialect (Lee et al. 2013; Lee and Jongman 2019; Hwang et al. 2019; Lee 2020).

The change is described as a VOT merger between aspirated stops /p^h, t^h, k^h/ and lenis stops /p, t, k/, which is compensated by F0 cues in younger speakers' production. Specifically, VOT traditionally served as the primary phonetic cue to distinguish the two series of stops, and older speakers still tend to preserve the VOT-based distinction. Younger speakers instead rely on the difference in F0 of the following vowel (high for aspirated; low for lenis) while VOT cues are not distinctively used any more.

Just as diachronic phonemic shifts often emerge from certain restricted phonological environment—e.g., see Morgan 1969; Brown 1991, for the *pin-pen* merger that is applied only before nasal—this shift in phonetic cue weighting from VOT to F0 is prosodically conditioned, currently being restricted in phrase-initial position. Note that, in this study, phrase-initial position refers to the initial position of an Accentual Phrase (AP-initial), Intonational Phrase (IP-initial), or the entire utterance (Utt-initial). In Jun's (2000) framework, at least two levels of prosodic phrases are

hierarchically organized in the intonational structure of Korean. Intonational Phrases (IP) are demarcated by a pause and preboundary lengthening (Jang and Katsika 2020; Kim et al. 2024a), and Accentual Phrases (AP)—a low-level phrasal unit embedded in an IP—by a particular pattern of tonal melody at their left and right edges. For example, the left edge of an AP is realized with either HH (when it begins with an aspirated sound) or LH (elsewhere).

The innovative F0-based stop distinction is systematically aligned with these two AP-level left-edge tonal patterns (HH and LH), such that a high tone is assigned to an AP-initial syllable with an onset aspirated stop, whereas a low tone is associated with an AP-initial syllable beginning with a lenis onset. Since higher-level units (e.g., IP or the whole utterance) are composed of at least one AP unit, an AP serves as the basic phrasal unit, and the left edge of the first AP is aligned with that of the embedding IP, following the strict layering hypothesis (Selkirk 1984).

In sum, younger speakers' F0-based distinction as a compensation of the VOT merger occurs in phrase-initial positions, including AP-initial, IP-initial, and Utt-initial position. On the other hand, phrase-medial stops in word-initial position (Wd-initial, henceforth) are contrasted by VOT regardless of speaker age (see Kang 2014; Cho and Lee 2016; Choi et al. 2020; Kim 2024, for more details about this prosodic conditioning).

2.3 Predictions

To the best of our knowledge, no previous research has systematically examined whether ongoing phonetic change is reflected in TTS-generated speech, such that voices identified as belonging to different age groups exhibit distinguishable phonetic patterns, each associated with their social identity. In addition, since commercial TTS systems typically do not disclose their internal mechanisms of reproducing phonetic variability in human speech—or whether they specify objective functions to encode certain socially indexed phonetic features—our a priori expectations allow for two equally possible outcomes.

Under the first hypothesis, TTS voices would systematically imitate the age-associated phonetic differences, differentially weighting phrase-initial VOT cues in accordance with speaker age. In this case, it would suggest that TTS systems can

sensitively accommodate sociophonetic indexicality of human speech, closely in line with age-indexed bimodal distribution of fine-grained phonetic forms of a phoneme that is present in the training data.

Under the null hypothesis, TTS systems may fail to capture socially meaningful phonetic variation, instead distinguishing the stops primarily using the VOT dimension in speech data associated with both older and younger voices. If so, it would suggest that data-driven speech synthesis may lack access to the socially grounded mechanisms. Alternatively, the incomplete merger may merely reflect a data-driven averaging process, whereby the merger is attenuated when recordings from different times are combined, or when the phrase-medial primary function of VOT is overgeneralized across prosodic contexts (see Section 5.2 for discussion).

3. Method

3.1 Collection of human data and speech materials

We used sentence reading data drawn from previous research (Kim 2024) as the baseline human speech dataset to be compared with TTS speech. That dataset comprised 12 young native speakers of Korean, but additional speech data were collected from 3 more speakers using the same method. Thus, a total of 15 younger speakers (ages 21–26; 9 females and 6 males) were included in the human speech data. Although they varied in regional background (5 from Seoul, 5 from Jeolla, 3 from Chungcheong, 2 from Gyeongsang), all were fluent in Seoul Korean and consistently exhibited the VOT merger, primarily distinguishing phrase-initial stop consonants through F0 cues (see Results and Kim 2024 for further details).

The speech materials used for both human and TTS speech in the current study consisted of 6 verse-style sentences shown in Table 1. These 6 sentences were selected from the 12 sentence items used in Kim (2024) to systematically compare the two phonation conditions (aspirated, lenis) across three places of articulation (bilabial, alveolar, velar), while controlling for other extrinsic factors that did not show significant effects in Kim (2024). For example, the phonological context was controlled to be /ʌlʌŋ/ after the aspirated or lenis stop. By doing so, the stop segment occurred within a low-frequency mimetic word (or a pseudo-word in case of /k^hʌlʌŋ/ and

/kʌlʌŋ/) to avoid phonetic reduction in high-frequency words (Bybee 2002).

Table 1. Speech materials

Place	Phonation	Phrase 1			IP boundary	Phrase 2
		P1 (Utt)	P2 (Wd)	P3 (AP)		P4 (IP)
bilabial	aspirated	pʰʌlʌŋ	pʰʌlʌŋ	pʰʌlʌŋi		pʰʌlʌŋ kʌɾjʌjo
		Flap-flap, flapping,				*it is flapping.*
	lenis	pʌlʌŋ	pʌlʌŋ	pʌlʌŋi		pʌlʌŋ kʌɾjʌjo
		Flip-flip, flipping is,				*it is flipping.*
alveolar	aspirated	tʰʌlʌŋ	tʰʌlʌŋ	tʰʌlʌŋi		tʰʌlʌŋ kʌɾjʌjo
		Jangle-jangle, jangling,				*it's jangling.*
	lenis	tʌlʌŋ	tʌlʌŋ	tʌlʌŋi		tʌlʌŋ kʌɾjʌjo
		Jingle-jingle, jingling,				*it's jingling.*
velar	aspirated	kʰʌlʌŋ	kʰʌlʌŋ	kʰʌlʌŋi		kʰʌlʌŋ kʌɾjʌjo
		Curly-curly, curly,				*it's curly.*
	lenis	kʌlʌŋ	kʌlʌŋ	kʌlʌŋi		kʌlʌŋ kʌɾjʌjo
		Curly-curly, curly,				*it's curly.*

Importantly, all stop-initial target words were manipulated to occur in specific prosodic contexts. In each sentence, the target word appeared in four prosodic positions, each associated with a different boundary strength: Utt-initial (P1), Wd-initial (P2), AP-initial (P3), and IP-initial (P4). These sentences were read following a singing session in which participants performed the sentences as song lyrics, a method intended to prime the target prosody. During the subsequent reading session, participants were instructed to maintain a similar rhythm. Overall, the utterances were naturally produced with the intended prosody and a fast speech rate (see Kim 2024; Kim et al. 2024b, for further details about this method).

A total of 720 stop tokens were obtained from the human speech data (6 sentences × 4 positions × 2 repetitions × 15 participants). One token in P3 was excluded from the analysis due to an unidentifiable release burst caused by background noise.

3.2 TTS voice selection and speech generation

Three commercial AI-based TTS platforms were used to generate the sentences in Table 1: Naver CLOVA Dubbing (hereafter Naver), Typecast, and WE MAKE VOICE

(WMV). Permission for academic use was obtained from each platform prior to data collection, and all speech data were generated between June 29 and 30, 2025.

To select AI voices appropriate for this study (hereafter referred to as TTS speakers), we first sorted the available voices by user popularity (i.e., usage frequency) as provided by each platform, under the assumption that frequently used voices would be more refined and better optimized. This procedure allowed us to exclude voices that sounded unnatural when the target sentences were provided as input (e.g., voices with unnaturally stable amplitude across the utterance or synthesis errors due to misrecognition of the input text). Through this process, we obtained a total of 115 candidate voices, aiming to include 10 voices for each of the four age (older, younger) × gender (male, female) categories from the platforms where possible. This balancing was fully achieved for Naver and Typecast; however, WMV offered only 8 older male voices and 7 older female voices.[1]

Next, since TTS speakers varied in speech rate at their default settings, we prioritized those whose speech rates closely matched those of the human speech data. To this end, we first generated a pair of sentence items contrasting aspirated and lenis stops (i.e., /p^hʌlʌŋ/ vs. /pʌlʌŋ/), and individually adjusted the speech rate parameter for each TTS speaker. We then retained only naturally sounding voices whose duration of the first IP (i.e., Phrase 1 in Table 1) fell between 1.0 and 1.2 seconds, corresponding to the range observed in the human speakers' productions.

From this subset, 60 TTS speakers were selected, balanced across age and gender. This resulted in five TTS speakers per age/gender group from each platform: old female (OF), old male (OM), young female (YF), and young male (YM). To assess whether these voices were perceived as belonging to the intended age group, we conducted an online rating task with 31 human participants in their 20s. Participants were played the first sentence item in Table 1 (i.e., the one with /p^h/) in each TTS speaker's voice, and then provided subjective ratings for sociolinguistic properties (age, gender, dialect) as well as speech naturalness. These perceptual properties are analyzed in Section 4.1, and a summary of each TTS speaker's rating appears in the Appendix.

---

1 This selection process required platform-specific adjustments due to the differences in speaker categorization across the platforms. For instance, Naver did not explicitly display older speakers, instead using the term 'middle-aged or older speakers'. Additionally, WMV had a limited number of older voices. To address these inconsistencies, older voices were selected from the category named such as 'middle aged', 'older' or containing both of terms, as in the case of Naver.

To elicit the intended prosodic renditions illustrated in Table 1, we consistently applied a fixed spacing and punctuation scheme in the text input (e.g., *펄렁펄렁 펄렁이, 펄렁거려요.*),[2] and further fine-tuned the speech rate for individual tokens when necessary. After generating the TTS outputs, the authors examined prosodic realizations of each token and confirmed that all tokens had the smallest word boundary before P2 and a brief pause before P4 (as an effect of the comma). The juncture before P3 was generally realized as an AP boundary (as an effect of spacing), although this was not perfectly controllable.

A total of 1,440 TTS-generated stop tokens were obtained (6 sentences × 4 positions × 2 ages x 2 genders x 3 platforms x 5 TTS speakers), and sound files were exported for acoustic analysis in WAV format with 16-bit quantization. Sampling rates ranged from 24kHz to 48kHz. One token was excluded due to a synthesis error.[3]

3.3 Measurement

VOT and F0 values were acoustically measured using Praat 6.4.18 (Boersma 2001), consistently with Kim's (2024) procedure. VOT duration was measured from the release burst onset to the onset of periodic voicing. In cases of breathy voice (mostly found in the human speech data), the vowel onset was marked at the beginning of a consistent voice bar. Lenis stops with a weak release in phonetically intervocalic position—i.e., a momentary disruption in an otherwise continuous periodic waveform—were identified as fully voiced tokens and assigned a VOT value shorter than 1ms. Lenis tokens that lacked clear evidence of aspiration but exhibited a sufficiently strong release to disrupt global periodicity were not classified as fully voiced.

F0 values were measured at the midpoint of the rhyme duration (i.e., the [ʌl]

2 This scheme was useful to generate the intended prosodic landmarks: (1) the period induced an utterance-final falling boundary tone and preboundary lengthening; (2) the comma induced a clear IP boundary before P4, accompanied by a short pause and preboundary lengthening; (3) the space induced an AP boundary with an LH right-edge marking before P3; and (4) no space or punctuation marks induced the smallest Wd boundary before P2.

3 This error occurred in an aspirated bilabial stop in P4 produced by a TTS speaker named 코맙소 from Naver CLOVA Dubbing, which was perceptually closer to an alveolar stop. In addition, 5 tokens were perceptually misaligned with their intended phonation category. However, these tokens were retained, as their acoustic ambiguity was driven by VOT or F0 (rather than formant transitions indicating the place of articulation), which was considered part of the variability this study aims to investigate.

portion preceding the lateral release), rather than the vowel onset, avoiding segment-dependent F0 variation primarily due to the local pitch perturbation effect of aspirated and lenis stops (Hanson 2009). In addition, all target syllables had a stop–vowel–liquid structure and coarticulatory effects of the liquid tended to emerge early in the rhyme, making the midpoint of the rhyme a reasonable carrier of the syllable-level pitch target. A Praat script originally written by Mietta Lennes (available at https://lennes.github.io/spect/) was used to automatically extract F0 and VOT values. Tracking errors or missing F0 values were manually corrected. All F0 values were normalized by converting them to z-scores within each speaker.

3.4 Statistical analysis

To statistically test VOT patterns in the human and TTS speech data, linear mixed-effects regression analyses were carried out in three folds, using the *lmerTest* package (Kuznetsova et al. 2017) in R 4.4.1 (R Core Team 2024).

First, a global model was fitted to the entire dataset to evaluate two hypotheses: (1) whether younger TTS speakers exhibited a VOT merger between aspirated and lenis stops in phrase-initial positions (i.e., P1, P3, and P4), deviating from older TTS speakers' VOT-based stop contrast; (2) if so, whether the degree of VOT merger in younger TTS speakers was statistically comparable to, or heterogeneous from, that of younger human speakers. In this model, raw VOT values were predicted by the main effects of, and all two-way and three-way interactions among, three experimental factors: Phonation (asp, lenis; contrast-coded as –0.5 and 0.5, with the underlined level as the reference), Position (P1, P2, P3, P4; dummy-coded), and Group (Human younger, TTS younger, TTS older; dummy-coded).[4][5] In addition, Place (bilabial, alveolar, velar; scaled sum-coded as –0.5 and 0.5) was included as a fixed effect to

4 For consistency in interpreting the coefficients across fixed effects, we initially considered applying scaled sum coding to all predictors. However, the two non-binary factors (Position and Group) were ultimately dummy-coded, allowing for direct comparison of Phonation effects between the reference level (i.e., P2 or TTS younger) and the other levels.

5 We also considered a simpler specification of the fixed effects, including only one 2-way interaction term (Phonation × Position) and one 3-way interaction term (Phonation × Position × Group). However, the two models did not differ in fit, showing virtually identical conditional R^2 values (R^2c=.827 in both models; Nakagawa and Schielzeth 2013). We therefore retained the full model, interpreting the 3-way interaction at the global level rather than breaking down all possible combinations.

control for potential phonetic variability. Random effects included by-speaker intercepts and by-speaker slopes for Phonation and Position. We confirmed via a likelihood ratio test that this structure was best supported by the experimental design and the sampled data (see Barr et al. 2013; Matuschek et al. 2017; Winter 2020).[6]

Second, a post-hoc analysis was conducted to further examine the local statistical significance of VOT differences between aspirated and lenis stops across the experimental conditions. We constructed a series of 12 subset *lmer* models, each fitted to one of the 4 prosodic position x 3 speaker group combinations. Each model included two fixed effect terms (main effects of Phonation and Place) as well as by-speaker random intercepts and random slope for Phonation.

Third, to examine platform-specific differences in VOT realizations, TTS-generated stops were separately modeled to examine the main effects of and interactions among Position (P1, P2, P3, P4), Age (older, younger), and Platform (Naver, Typecast, WMV), along with other control factors that reached significance (i.e., Gender, Place). To avoid an overly complex 4-way interaction of the test variables (i.e., Position x Age x Platform x Phonation), the dependent variable was defined as the VOT difference between aspirated and lenis stops—i.e., ΔVOT(aspirated-lenis)—within each matched pair. To compute these Δ values, we retained only those pairs that contained both aspirated and lenis tokens produced by the same TTS speaker under identical experimental conditions (i.e., Place, Repetition, and Position). For each valid pair, the VOT difference was obtained by subtracting the lenis value from the aspirated value. This procedure yielded 719 usable pairs from the original TTS dataset (N=1,439). The model included by-speaker random intercepts and by-speaker random slopes for the Position effect, representing the maximal random-effects structure justified by the design (Barr et al. 2013).

All sound files, Praat textgrid files, processed data spreadsheet, and R script necessary to reproduce the results are publicly available at the OSF repository (https://osf.io/2jp3c/).

6 Note that by-item random effects were not included to avoid overfitting, as item-level variation was already captured in the model: the six sentence items were uniquely defined by the factorial combination of Phonation and Place.

4. Results

4.1 Perceived social characteristics of TTS speakers

Before examining how patterns of sound change are reflected in TTS speech, this section first reports results from the online rating task, in which human listeners assessed sociolinguistic characteristics of our TTS speakers. Figure 1 compares the relative frequencies of perceived age ratings between the younger and older TTS speaker groups, broken down by TTS speaker gender and platform.

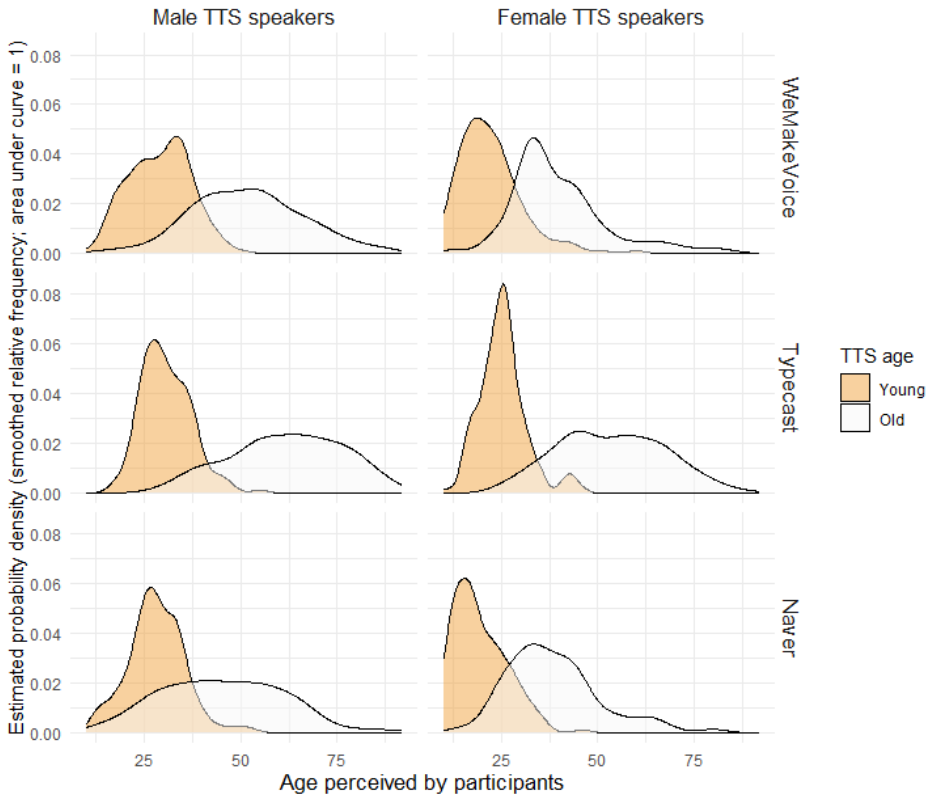

Figure 1. Human participants' (N=31) perceived age rating results by TTS speakers' (N=60) gender and age categories in each platform

Across the six panels, it is clear that TTS voices generally aligned with the age

categories defined by each of the three platforms, regardless of speaker gender. This overall trend was supported by a two-sample independent t-test, which revealed a significant difference between the age groups, t (1858)=-38.220, p<.001. The young TTS group (M=25.795) was rated as younger than the old group (M=48.445). The 95% confidence interval for the difference in means ranged from -23.812 to –21.487. This difference corresponded to a very large effect size (Cohen's d=−1.77, 95% CI [−1.88, −1.66]), indicating a robust perceptual separation between the two TTS age groups.

However, ratings for the older TTS speakers showed wide variability. The overlaps observed in Figure 1 were primarily driven by the distribution for older TTS speakers, which was twice as wide (SD=16.134) as that for younger TTS speakers (SD=8.146). In particular, 17 of the 30 older TTS speakers (11 females, 6 males) were rated as sounding younger than 50, with minimum perceived ages ranging from 11 to 33 (see Appendix).

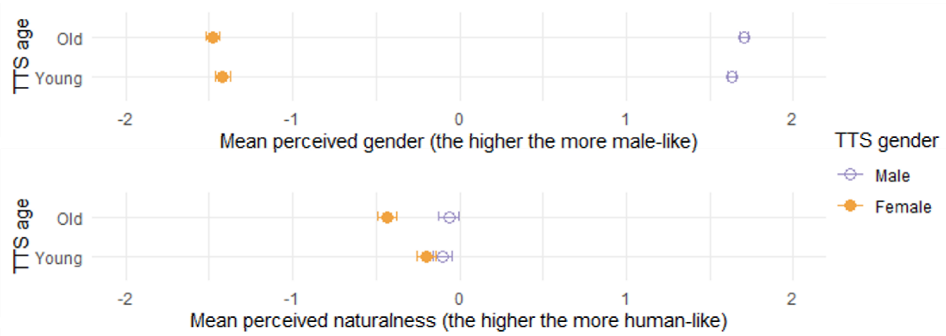

Figure 2. Human participants' rating results for perceived gender (upper panel) and naturalness (lower panel)

Figure 2 presents the results for gender and naturalness ratings. Error bars represent standard error of the mean. Gender ratings closely aligned with each system's intended gender categorization, regardless of age, although male TTS speakers were perceived as slightly more prototypically male (M=1.670) than female speakers were perceived as prototypically female (M=–1.451). As for naturalness, both older (M=–0.249) and younger (M=–0.149) TTS speakers were rated as sounding similarly neutral, falling between machine-like and human-like. However, older female speakers were

perceived as more machine-like (M=–0.434) than the other three categories.

Finally, the overall rate of responses identifying the auditory stimuli as non-Seoul dialect was 13.317%. Specifically, 19 out of the 60 TTS speakers were categorized as non-Seoul speakers by more than 3 out of 31 human listeners. Among them, 6 TTS speakers received a relatively high rate of non-Seoul responses, ranging from 13 (42%) to 22 (71%).

4.2 VOT analysis

This section provides the main analysis of the current study, regarding the degree of VOT merger reflected in TTS speech in comparison to human speech. Figure 3 presents VOT distributions (in raw values) of aspirated (blue empty circles) versus lenis (red filled circles) stops across the four prosodic positions in (a) younger human speakers, (b) younger TTS speakers, and (c) older TTS speakers. β -coefficients and *p*-values of the Phonation effect estimated in the subset analyses are summarized in each of the 12 panels.

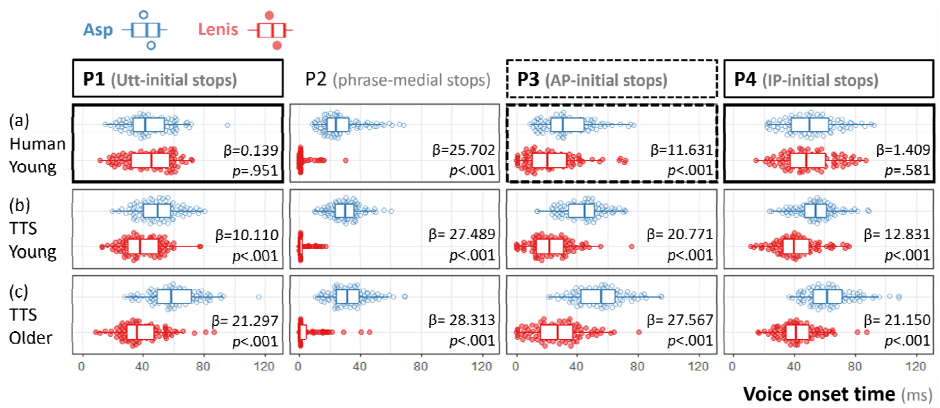

Figure 3. VOT distributions by Phonation, Position, and Group

As shown in Figure 3 (a), younger human speakers showed the expected VOT merger in P1 and P4—the two phrase-initial positions with the largest prosodic boundaries (marked by solid outlines of the panel). By contrast, an effect of Phonation emerged in P2 (p<.001), indicating a VOT-based phonetic distinction between lenis stops and their aspirated counterparts in phrase-medial position. This effect was largely

driven by fully voiced lenis tokens in P2, as evidenced by the dense cluster of red data points around VOT=0.

A significant Phonation effect was also observed in P3 (p<.001), but its β-value (11.631) was less than half of that in P2 (25.702), reflecting a markedly smaller effect magnitude in AP-initial position. In other words, the merger associated with an AP boundary underwent prosodically conditioned weakening of phrase-initial VOT merger (marked by dashed outlines of the panel), owing to its smaller boundary strength.[7]

Importantly, neither younger TTS speakers in Figure 3 (b) nor older TTS speakers in (c) clearly exhibited the phrase-initial merger pattern. Instead, aspirated and lenis stops remained acoustically distinct along the VOT dimension across all four prosodic positions (p<.001 in all 8 subset models). However, a comparison between the two TTS age groups reveals systematic differences in effect sizes (i.e., β-values) across the three phrase-initial contexts: younger TTS speakers in (b) consistently fell between younger human speakers in (a) and older TTS speakers in (c), in terms of the degree of VOT merger in P1, P3, and P4. These trends were assessed in the global model, the output of which is summarized in Table 2.

7 As discussed in Kim (2024), the VOT merger weakened in P3 is understood as a function of boundary strength. The significant Phonation effect in P3 is attributed to coarticulatory processes affecting lenis stops after the smaller phrasal boundary (i.e., an AP boundary). In fast speech, gestural overlap between oral release and vocal fold vibration can result in partially or fully voiced tokens. Consequently, despite being in phrase-initial position, the VOT merger in AP-initial P3 was substantially weaker than in the other two phrase-initial positions with larger boundaries (i.e., P1 and P4).

Table 2. Summary of the global analysis

	Est.	S.E.	z-value	*p*-value
(Intercept)	16.441	0.942	17.445	<.001
Phonation=asp	27.489	2.014	13.652	<.001
Position=P1	28.240	1.912	14.771	<.001
Position=P3	16.780	1.625	10.326	<.001
Position=P4	31.053	1.726	17.990	<.001
Group=human young	-1.588	1.489	-1.066	.291
Group=TTS old	2.652	1.333	1.990	.050
Place=velar	12.935	0.547	23.647	<.001
Place=bilabial	-5.671	0.547	-10.363	<.001
Phonation=asp : Position=P1	-17.379	1.894	-9.174	<.001
Phonation=asp : Position=P3	-6.718	1.894	-3.547	<.001
Phonation=asp : Position=P4	-14.658	1.894	-7.738	<.001
Phonation=asp : Group=human young	-1.787	3.220	-0.555	.580
Phonation=asp : Group=TTS old	0.824	2.848	0.289	.773
Position=P1 : Group=human young	1.025	3.173	0.323	.748
Position=P3 : Group=human young	-3.663	2.651	-1.382	.172
Position=P4 : Group=human young	4.578	2.836	1.614	.112
Position=P1 : Group=TTS old	2.591	2.704	0.958	.341
Position=P3 : Group=TTS old	5.209	2.298	2.267	.026
Position=P4 : Group=TTS old	2.956	2.442	1.211	.230
Phonation=asp : Position=P1 : Group=human young	-8.184	2.679	-3.055	.002
Phonation=asp : Position=P3 : Group=human young	-7.363	2.681	-2.746	.006
Phonation=asp : Position=P4 : Group=human young	-9.635	2.679	-3.597	<.001
Phonation=asp : Position=P1 : Group=TTS old	10.363	2.679	3.868	<.001
Phonation=asp : Position=P3 : Group=TTS old	5.972	2.679	2.229	.026
Phonation=asp : Position=P4 : Group=TTS old	7.396	2.681	2.759	.006

Since Position and Group were dummy-coded, the estimated intercept (16.441) represents the mean VOT of younger TTS speakers in P2—i.e., the second panel of Figure 3 (b), averaged across aspirated and lenis stops. Along with the main effects of the predictor terms (Phonation, Position, Group, Place), the three negative β-slope coefficients in the 2-way interaction between Phonation and Position (indicated in the upper box in Table 2) shows that the Phonation effect—that aspirated stops were 27ms longer in VOT than lenis—was attenuated in effect size by about 17ms in P1, 7ms in P3, and 15ms in P4 (p<.001, respectively).

While these 2-way interactions reflect a prosodically conditioned adjustment of VOT cues in the younger TTS group, the negative β-values in the 3-way interaction terms comparing the reference-level younger TTS vs. younger human groups (i.e., the middle box in Table 2) reveal that younger human speakers reduced VOT differences to a greater extent than younger TTS speakers in each phrase-initial position (p<.01 in P1 and P3; p<.001 in P4).

In the lower box, on the other hand, the three β-values are positive, showing that, compared to younger TTS speakers, older TTS speakers weighted VOT cues to a greater degree to distinguish phrase-initial stop categories (p<.001 in P1; p<.05 in P3; p<.01 in P4).

In sum, these results demonstrate that the three groups (TTS younger, TTS older, human younger) differed significantly in their use of VOT as a cue to phrase-initial stop contrasts.

Recalling the wide variability in perceived age ratings among older TTS speakers, one might be concerned that those older TTS speakers perceived as relatively younger may have reduced VOT cues, consistent with the innovative merger pattern. Importantly, the goal of this study is not to model phonetic patterns as a function of listeners' perceived age, but to test whether the age-based voice categories implemented by TTS platforms capture systematic sociophonetic patterns associated with sound change, particularly focusing on the pattern of younger TTS speakers. However, to directly evaluate this possibility, we conducted a supplementary analysis in which perceived age (over 50, under 50) was incorporated into the model.

Specifically, we fit a new model to the subset data of older TTS speakers, including main effects of, and 2-way and 3-way interactions among, Phonation, Position, and Perceived Age. Perceived Age was first coded as a binary factor (over 50 or under 50 in average rating). The model retained the same structure for all other terms,

including Place as a control predictor, as well as identical random-effects specifications and coding schemes.[8]

This analysis revealed no main effect of Perceived Age (β=−0.764, p=.744), no Phonation × Perceived Age interaction (β=−4.868, p=.287), and no Phonation × Position × Perceived Age interactions (β=3.570, p=.352 in P1; β=5.650, p=.141 in P3; β=−7.045, p=.067 in P4). Similar results were obtained when Perceived Age was treated as a continuous predictor (i.e., using the mean perceived age rating for each speaker): neither the main effect of Perceived Age nor any interaction terms reached statistical significance.

These results indicate that variation in perceived age within the older TTS group did not meaningfully modulate VOT cue weighting. In particular, as suggested by the non-significant 3-way interaction, there was no evidence that younger-sounding older TTS voices exhibited reduced VOT contrasts in phrase-initial positions.

4.3 F0 analysis

This section provides a comparison of F0 realizations across the three groups. As shown in Figure 4 (a), younger human speakers showed a clear F0-based distinction in all positions, including P2 where VOT served as the primary cue (see Choi et al. 2020; Kim 2024 for discussion of younger speakers' duplicative use of F0 under focus). By contrast, younger TTS speakers in (b) displayed a substantially larger overlapping distribution in phrase-medial P2.[9] This difference was captured in a supplementary model fit to z-scored F0 (instead of VOT) with all other terms held constant with the VOT model reported in Table 2: a significant Phonation x Group interaction was found (β=1.016, p<.001), indicating a greater Phonation effect on F0 in P2 for the younger human group when compared to the younger TTS group. Except for this difference, younger TTS speakers exhibited clear F0 contrasts across the three

8 Note that this model was fit using a newer version of R (R 4.5.2).

9 The overlap in P2 was primarily driven by exaggerated intonational patterns, which is not relevant to the VOT merger. First, the F0 down-stepping pattern of aspirated stops from P1 to P3 (Kim 2024) spanned a wider pitch range than that of younger human speakers, which in turn lowered the F0 range in P2. Second, lenis stops in P2 were raised in pitch due to an exaggerated left-edge marking of the first AP domain in younger TTS speech. For example, the H tone assigned to the second syllable in /pʌlʌŋ pʌlʌŋ pʌlʌŋ-i .../ was realized so high that the subsequent P2 position was also produced with high pitch.

phrase-initial positions (P1, P3, P4) that closely resembled those of human speakers.

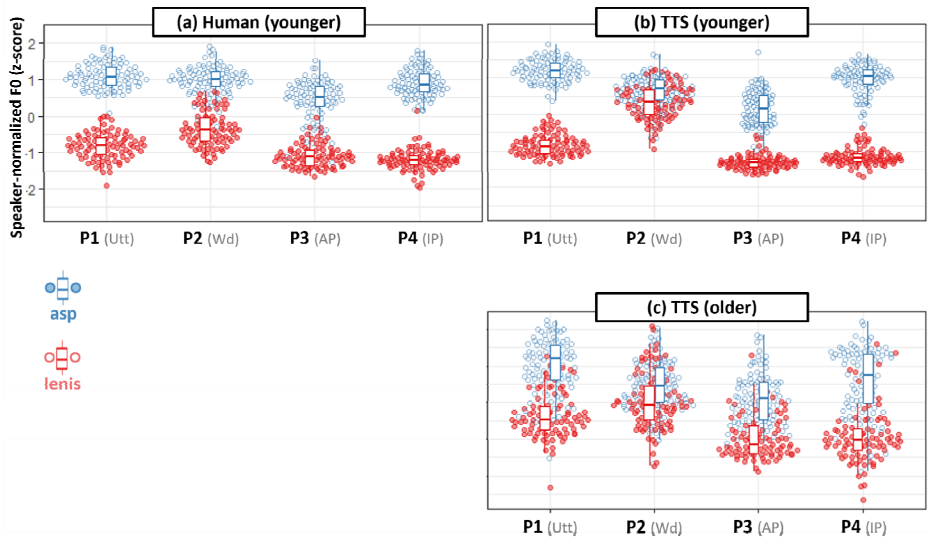

Figure 4. F0 distributions by Phonation, Position, and Group

Older TTS speakers showed far more overlapped F0 distributions in phrase-initial positions. The supplementary model confirmed this pattern: the Phonation x Position x Group interaction indicated that the use of F0 cues was significantly attenuated in magnitude in the older TTS group (when compared to the younger TTS group) in all phrase-initial positions (β=-7.534 in P1, β=-4.928 in P3, β=-6.955 in P4; p<.001 in all three positions). In line with previous research on human speech data (e.g., Kang 2014), this suggests that F0 is less likely used primarily in older TTS speech, in the presence of their more pronounced VOT-based distinction, than the TTS younger group, which was shown in Section 4.2.

4.4 VOT patterns across the platforms

To summarize the results above, TTS systems adjusted VOT and F0 cues in accordance with age to some extent. However, compared with human younger speakers, stops produced by younger TTS speakers exhibited divergent behavior, primarily in the VOT dimension rather than in F0. Specifically, younger TTS speakers displayed a greater

VOT contrast across all phrase-initial positions (P1, P3, P4) relative to human younger speakers.

This section takes a closer look at the TTS data to examine whether the greater VOT contrast reflects a general tendency shared across the platforms, or whether the platforms differ in how closely they approximate the VOT merger. In Figure 5, VOT differences (ΔVOT) between aspirated and lenis stops are plotted across the four prosodic positions (P1–P4). It compares the effects of TTS speaker age (young=filled orange; old=open purple) across platforms: (a) WMV, (b) Typecast, and (c) Naver. Data from human younger speakers are also shown in (d) for reference.

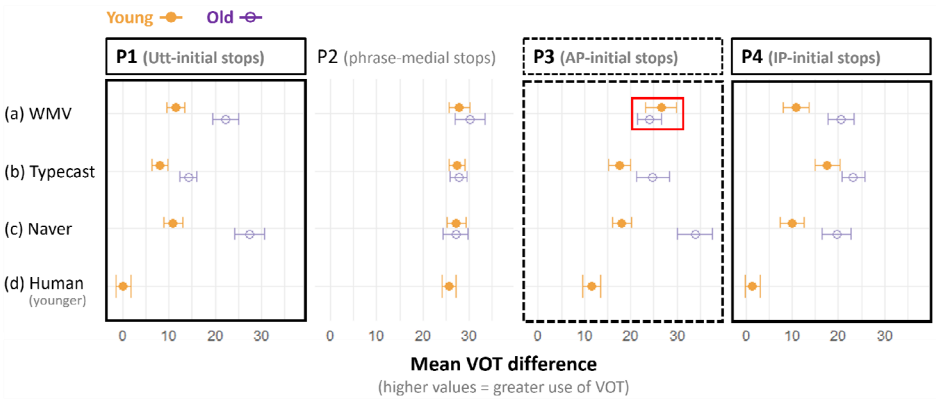

Figure 5. Mean VOT differences (i.e., ΔVOT (aspirated-lenis)) by Position, Age, and Platform (a, b, c) in comparison to human younger speakers (d)

As shown in Figure 5, phrase-medial stops in P2 showed a consistent VOT-based distinction (regardless of TTS speaker age), with minimal variation in ΔVOT values across all platforms and human younger speakers (mean=25.9, SD=12.7). In contrast, in phrase-initial positions where merger was expected—namely P1, P3, and P4—TTS platforms generally exhibited larger ΔVOT values than human younger speakers. This pattern held even for younger TTS speakers, as confirmed also by the analyses reported above.

A comparison of age-related differences across the TTS systems further illustrates how the platforms differentially exploited phrase-initial VOT cues as a function of speaker age. Overall, younger TTS speakers tended to rely less on VOT than their older counterparts, reflecting the age-related variation within TTS-generated speech.

The sole exception to this general tendency was observed for AP-initial stops in P3 produced by the WMV system (highlighted by a red box in Figure 5), where the error bars for younger and older speakers overlapped.

This pattern was captured in the third fold of the mixed-effects model fitted to the TTS data (see Section 3.4 for details). Specifically, a significant Position × Age interaction (β=−15.985, p=.010) was found, indicating that, within Naver—the reference platform—ΔVOT was reduced to a greater extent (by about 16ms) from P2 to P3 when produced by younger speakers relative to older speakers. Crucially, this age-related reduction effect on ΔVOT in P3 was neutralized for WMV, as reflected in the significant 3-way interaction between Position, Age, and Platform (β=20.827, p=.017). In other words, unlike the other TTS systems, WMV showed little differentiation between younger and older voices in AP-initial contexts, resulting in a convergence of VOT patterns across age groups.

5. Discussion

5.1 Age-related VOT and F0 cues realized in TTS

Through the acoustic analyses on three TTS systems, this study found some evidence that age-indexed phonetic cues are reflected in TTS systems in the context of the ongoing sound change in Korean stops. The age-related effects in TTS-generated speech were observed in both dimensions of the phonetic cues involved in the sound change. With respect to VOT (Section 4.2), younger TTS voices showed reduced reliance on phrase-initial VOT values compared to older voices. Although further analyses revealed platform-specific differences (Section 4.4), the only clear deviation from this overall pattern was observed in AP-initial stops produced by WMV. In the F0 dimension (Section 4.3), older TTS voices exhibited substantial overlap between aspirated and lenis stops, indicating a weaker use of F0 as a contrastive cue. This pattern contrasts with that of younger TTS voices and aligns with previous findings from human speech (e.g., Kang 2014).

While perceived age ratings varied widely among older TTS speakers (Section 4.1), additional analyses demonstrated that perceived age did not significantly influence the VOT patterns summarized above. Thus, the observed differences across younger

and older TTS speaker groups do not appear to be contaminated by the variation in perceived age. Instead, our results indicate that AI voices labeled as young or old by TTS providers are capable of approximating age-related phonetic patterns associated with the ongoing sound change, at least at a coarse level.

Notably, however, younger TTS speakers' VOT distribution did not fully converge with that of younger human speakers. While human speakers showed a complete merger of VOT in prosodic positions with the largest boundaries (P1, P4),[10] TTS systems retained significant VOT differences in the same prosodic contexts. Moreover, a similar discrepancy was observed before a smaller phrasal boundary (P3): while younger human speakers exhibited reduced VOT differentiation at the AP-initial position (P3) relative to the phrase-medial context (P2), the corresponding reduction in the younger TTS group was markedly weaker. The implications of these differences for sociophonetic representation in AI-generated speech are discussed below.

5.2 Socio-cognitive agency and the perceived talker

As noted earlier, socio-indexical phonetic cues play a central role in shaping person perception, and their multilayered social meanings are dynamically deployed in human communication (Bucholtz 1999; Campbell-Kibler 2007; Podesva 2007; Eckert 2008; Love and Walker 2013; Lev-Ari and Pepperkemp 2014; Sneller 2024). In sociophonetic literature, this communicative flexibility is understood to be internalized in the cognitive structure. Sociophonetic indexicality emerges from low-level associations, giving rise to indexical fields of social meaning that operate at a largely implicit level of cognition (Eckert 2008). Relatedly, Labov (1972) asserted that phonetic variation is pervasive within speech communities, but only a small subset of such variation becomes overtly recognized or discussed as social or linguistic stereotypes. Drager and Kirtley (2016) emphasized that all naturally produced speech inherently carries associations with social properties, and that such associations are organized as storage information and accessed automatically, even in the absence of conscious awareness. Such integration is explicitly specified within exemplar-based frameworks (e.g., Johnson 1997; Pierrehumbert 2001), in which repeated exposure to covariation between

10 Note that this complete merger does not appear to be idiosyncratic to the present study, given that Choi et al. (2020) provided comparable findings using a different experimental task and distinct lexical and sentential contexts.

phonetic and social attributes shapes richly structured memory representations that play a central role in modulating linguistic and social behaviors.

Encoded at a fine-grained phonetic level, the sound change of Korean stops appears to function as an implicit sociolinguistic marker among Korean speakers. In the lexical recognition experiment using age-neutral talkers mentioned above (Kim and Drager 2017), about half of the listeners reported no systematic difference in age rating across the talker guises. Nevertheless, these listeners exhibited priming effects comparable in magnitude to those who did report such awareness (see Kim 2018 for this additional analyses). This suggests that experienced covariation is robustly represented across social, phonetic, and lexical levels, and gives rise to automatized perceptual routines for integrating social and linguistic information—even if their connections do not emerge consciously.

This account for socially shaped representations of Korean stops directly associated with linguistic processing echoes with a large body of experimental work demonstrating how speech perception is informed by talker-related information, whether it is provided explicitly (Strand and Johnson 1996; Walker and Hay 2011; Yi et al. 2013; Babel and Russell 2015; Kim 2016; Kim and Drager 2018; So and Kim 2024; Zhang 2025) or implicitly (Johnson et al. 1999; Hay et al. 2006; Hay and Drager 2010). Collectively, these studies highlight humans' communicatively oriented cognitive flexibility in interpreting physically variable speech signals as a finite set of linguistically meaningful categories.

From this cognitive perspective, we turn to the question of how such sociophonetic structure might be captured in artificial speech systems. Despite its subtlety, sociophonetic indices are statistically straightforward. From both cognitive models such as Exemplar models and computational algorithms underlying TTS systems, sociophonetic variation can be conceptualized as clusters of stored utterances forming multimodal distributions over continuous phonetic dimensions. In principle, these clusters may be explicitly indexed with social labels in TTS systems. However, contemporary deep learning models may also approximate such structure implicitly, relying solely on statistical correlations between phonetic patterns and voice characteristics present in the training data. This perspective helps explain why the TTS systems examined here were able to imitate the ongoing sound change to a certain extent.

As noted in our initial predictions (Section 2.3), the attenuated merger pattern

in younger TTS speakers may also stem from non-social technical factors, such as heterogeneous training corpora or generalization across prosodic contexts. In such cases, large-scale models might fully reflect apparent-time variations through continued statistical learning alone, without encoding socially meaningful variation per se.

However, in the absence of direct evidence regarding the sources of the discrepancy, our discussion necessarily remains speculative. We therefore raise the possibility that the incomplete imitation observed in this study—rather than socially motivated participation in sound change—may arise from fundamental differences between human and machine learning mechanisms. Humans implicitly adjust phonetic forms, reflecting interactive agency as a socially motivated behavior (Eckert 2008; Podesva 2007; Foulkes and Hay 2015). In contrast, current data-driven systems appears to lack such socio-cognitive grounding. It remains an open empirical question whether future AI systems will develop sensitivity to socially conditioned phonetic cue adjustments used in negotiating social identity and guiding speech processing. If limitations in sociophonetic realization contribute to experiences of the "uncanny valley" (Mori 1970), such differences may in turn affect how readily humans engage with synthetic voices as socially interactive interlocutors (Zellou et al. 2021; Székely et al. 2025).

5.3 Future directions

Building on the incipient findings reported here, further investigation is warranted to probe sociophonetic aspects of AI-synthesized speech. First, the present results should be replicated using a different set of TTS voices. In this regard, it is worth noting several design-related limitations of the current study: (1) the human speech data were collected using verse-style utterances, which may limit generalizability to more typical contexts; and (2) since the data lacked older human speakers, a complete age-based comparison across human and TTS speech was not possible.

Second, future research could examine additional sociophonetic variables in Korean, as well as variables in other languages. In particular, phonetic variables that are not directly tied to diachronic change or prosodic structure may help clarify the sources of the discrepancies observed here.

Third, perception experiments on phonemic categorization can be applied to

automatic speech recognition systems (cf. So and Kim 2024, for a matched-guise social priming paradigm conducted with human listeners). Such approaches would provide more direct evidence of whether implicit knowledge of sociophonetic variation can emerge in AI perceptual systems.

Fourth, talker perception studies using human-produced Korean stops could shed light on whether perceived talker age, or other social meanings associated with the merger, are similarly represented by human listeners and AI models with speech recognition systems.

Finally, one promising direction for future research is to move beyond cross-speaker comparisons and apply dynamically adapting speaker agency within individual human speakers. For example, phonetic convergence and divergence patterns (Giles et al. 1991; Pardo 2006; Lev-Ari and Peperkamp 2014), as well as register shifting and persona construction (Bucholtz 1999; Podesva 2007), could be investigated in conversational interactions between human and AI interlocutors.

6. Conclusion

Given the increasing social nature of human–AI interaction in contemporary society, this study examined the Korean VOT merger as realized in AI-based synthetic speech. The results suggest that, although commercial TTS systems can approximate surface-level phonetic patterns associated with an ongoing sound change, they do not fully converge on the innovative phonetic realizations exhibited by younger human speakers.

This pattern was interpreted through the lens of speaker agency, a notion developed in sociolinguistic theory to explain socially motivated participation of speakers in language change. The gap between human and TTS speakers leads us to further explore the nature of contemporary AI speech synthesis: it is fundamentally descriptive rather than agentive. That is, contemporary TTS systems reproduce statistical regularities present in the data but do not actively participate in the socially motivated processes that drive linguistic change.

In this respect, this study contributes not only to evaluating TTS fidelity but also to the discussion of an understudied factor that may help distinguish human language from AI-based speech. Continued research may further clarify how social cognition

emerges through language use, thereby shedding light on the social foundations of linguistic evolution.

References

Ahn, Mee-Jin. 2017. Prosodic effects on acoustic cues for the Korean stop contrast: Evidence from Daejeon Korean. *The Journal of Linguistic Science* 82: 177–198.

Babel, Molly and Jamie Russell. 2015. Expectations and speech intelligibility. *The Journal of the Acoustical Society of America* 137(5): 2823–2833.

Bang, Hye-Young, Morgan Sonderegger, Yoonjung Kang, Meghan Clayards, and Tae-jin Yoon. 2018. The emergence, progress, and impact of sound change in progress in Seoul Korean: Implications for mechanisms of tonogenesis. *Journal of Phonetics* 66: 120–144.

Barr, Dale J., Roger Levy, Christoph Scheepers, and Harry J. Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* 68(3): 255–278.

Boersma, Paul and Vincent van Heuven. 2001. Speak and unSpeak with PRAAT. *Glot International* 5(9/10): 341–347.

Brown, Vivian. 1991. Evolution of the merger of /ɪ/ and /ɛ/ before nasals in Tennessee. *American Speech* 66(3): 303–315.

Bucholtz, Mary. 1999. "Why be normal?": Language and identity practices in a community of nerd girls. *Language in Society* 28(2): 203–223.

Bybee, Joan. 2002. Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Language Variation and Change* 14(3): 261–290.

Campbell-Kibler, Kathryn. 2007. Accent, (ING) and the social logic of listener perceptions. *American Speech* 82(1): 32–64.

Chan, Cedric and Jianjing Kuang. 2025. Toward objective and interpretable prosody evaluation in text-to-speech: A linguistically motivated approach. arXiv preprint, arXiv:2511.02104.

Cho, Sunghye and Yong-Cheol Lee. 2016. The effect of the consonant-induced pitch on Seoul Korean intonation. *Linguistic Research* 33(2): 299–317.

Choi, Jiyoun, Sahyang Kim, and Taehong Cho. 2020. An apparent-time study of an ongoing sound change in Seoul Korean: A prosodic account. *PLoS ONE* 15(10): e0240682.

Drager, Katie and Megan J. Kirtley. 2016. Awareness, salience, and stereotypes in exemplar-based models of speech production and perception. In Anna Babel (ed.), *Awareness and control in sociolinguistic research*, 1–24. Cambridge: Cambridge University Press.

Eckert, Penelope. 2000. *Linguistic variation as social practice: The linguistic construction of identity in Belten High*. Oxford: Blackwell Publishing.

Eckert, Penelope. 2008. Variation and the indexical field. *Journal of Sociolinguistics* 12(4): 453–

476.
Fellowes, Jennifer M., Robert E. Remez, and Philip E. Rubin. 1997. Perceiving the sex and identity of a talker without natural vocal timbre. *Perception & Psychophysics* 59(6): 839–849.
Foulkes, Paul and Jennifer Hay. 2015. The emergence of sociophonetic structure. In Brian MacWhinney and William O'Grady (eds.), *The handbook of language emergence*, 292–313. Malden, MA: Wiley-Blackwell.
Giles, Howard, Justine Coupland, and Nikolas Coupland. 1991. *Contexts of accommodation: Developments in applied sociolinguistics*. Cambridge: Cambridge University Press.
Hanson, Helen M. 2009. Effects of obstruent consonants on fundamental frequency at vowel onset in English. *The Journal of the Acoustical Society of America* 125(1): 425-441.
Hay, Jennifer. 2018. Sociophonetics: The role of words, the role of context, and the role of words in context. *Topics in Cognitive Science* 10(4): 696–706.
Hay, Jennifer and Katie Drager. 2007. Sociophonetics. *Annual Review of Anthropology* 36(1): 89–103.
Hay, Jennifer and Katie Drager. 2010. Stuffed toys and speech perception. *Linguistics* 48(4): 865–892.
Hay, Jennifer and Paul Foulkes. 2016. The evolution of medial /t/ over real and remembered time. *Language* 92(2): 298–330.
Hay, Jennifer, Aaron Nolan, and Katie Drager. 2006. From fush to feesh: Exemplar priming in speech perception. *The Linguistic Review* 23(3): 351–379.
Hu, Na, Jiseung Kim, Riccardo Orrico, Stella Gryllia, and Amalia Arvaniti. 2024. Can OpenAI's TTS model convey information status using intonation like humans? Presented at *Speech Prosody 2024 (SP2024)*. Leiden: Leiden University. 2–5th July.
Hwang, Young, Samson Lotven, and Kelly Berkson. 2019. Pitch accent and the three-way laryngeal contrast in North Kyungsang Korean. In Sasha Calhoun, Paola Escudero, Marija Tabain, and Paul Warren (eds.), *Proceedings of the 19th International Congress of Phonetic Sciences (ICPhS 2019)*, 2976–2980. Melbourne: The Australasian Speech Science and Technology Association.
Jang, Jiyoung and Argyro Katsika. 2020. The amount and scope of phrase-final lengthening in Seoul Korean. In Nobuaki Minematsu, Mariko Kondo, Takayuki Arai, and Ryoko Hayashi (eds.), *Proceedings of Speech Prosody 2020*, 270–274. Tokyo: International Symposium on Computer Architecture.
Johnson, Keith. 1997. Speech perception without speaker normalization: An exemplar model. In Keith Johnson and John W. Mullennix (eds.), *Talker variability in speech processing*, 145–165. San Diego, CA: Academic Press.
Johnson, Keith, Elizabeth A. Strand, and Mariapaola D'Imperio. 1999. Auditory-visual integration of talker gender in vowel perception. *Journal of Phonetics* 27(4): 359–384.
Kang, Kyoung-Ho. 2010. Generational differences in the perception of Korean stops. *Phonetics and Speech Sciences* 2(3): 3–10.

Kang, Kyoung-Ho and Susan G. Guion. 2008. Clear speech production of Korean stops: Changing phonetic targets and enhancement strategies. *The Journal of the Acoustical Society of America* 124(6): 3909–3917.

Kang, Yoonjung. 2014. Voice onset time merger and development of tonal contrast in Seoul Korean stops: A corpus study. *Journal of Phonetics* 45: 76–90.

Kang, Yoonjung, Suyeon Yun, and Na-young Ryu. 2024. Talker- and listener-conditioned use of height-dependent vowel duration cue under sound change in progress: /o/ to /u/ raising in Daejeon Korean. In Taehong Cho, Sahyang Kim, Jeff Holiday, and Sang-Im Lee-Kim (eds.), *Proceedings of the 19th Conference on Laboratory Phonology*, 243–244. Seoul: Hanyang University Press.

Kim, Jonny Jungyun. 2016. Perceptual associations between words and speaker age. *Laboratory Phonology* 7(1): 18.

Kim, Jonny Jungyun. 2018. *Socially-conditioned links between words and phonetic realizations*. PhD Dissertation. University of Hawai'i at Mānoa.

Kim, Jonny Jungyun. 2024. Prosodically conditioned VOT merger of Korean stops examined in verse-style broad-focus speech. *Journal of Language Sciences* 31(1): 111–134.

Kim, Jonny Jungyun and Katie Drager. 2017. Sociophonetic realizations guide subsequent lexical access. In Francisco Lacerda (ed.), *Proceedings of Interspeech 2017*, 621–625. Stockholm: International Symposium on Computer Architecture.

Kim, Jonny Jungyun and Katie Drager. 2018. Rapid influence of word-talker associations on lexical access. *Topics in Cognitive Science* 10(4): 775–786.

Kim, Jonny Jungyun, Sahyang Kim, and Taehong Cho. 2024a. Preboundary lengthening and articulatory strengthening in Korean as an edge-prominence language. *Laboratory Phonology* 15(1): 1–36.

Kim, Jonny Jungyun, Hyunjung So, Ahjin Ko, Jiyea Heo, and Seoyeong Ahn. 2024b. Ongoing VOT merger unmerged in a singing context. Poster presented at *19th Conference on Laboratory Phonology (LabPhon 19)*. Seoul: Hanyang University. June 2024.

Kim, Mi-Ryoung. 2014. Ongoing sound change in the stop system of Korean: A three-to two-way categorization. *Studies in Phonetics, Phonology and Morphology* 20(1): 51–82.

Kim, Mi-Ryoung, Patrice Speeter Beddor, and Julie Horrocks. 2002. The contribution of consonantal and vocalic information to the perception of Korean initial stops. *Journal of Phonetics* 30(1): 77–100.

Kleinschmidt, Dave F. and T. Florian Jaeger. 2015. Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review* 122(2): 148–203.

Kondo, Risa, Ayu Teramen, Reon Kajikawa, Koki Horiguchi, Tomoyuki Kajiwara, Takashi Ninomiya, Hideaki Hayashi, Yuta Nakashima, and Hajime Nagahara. 2025. Text normalization for sentiment analysis in Japanese social media. In JinYeong Bak, Rob van der Goot, Hyeju Jang, Weerayut Buaphet, Alan Ramponi, Wei Xu, and Alan Ritter (eds.), *Proceedings*

of the Tenth Workshop on Noisy and User-generated Text (W-NUT 2025), 149-157.
Kuznetsova, Alexandra, Per B. Brockhoff, and Rune H. B. Christensen. 2017. lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software* 82(13): 1–26.
Labov, William. 1963. The social motivation of a sound change. *Word* 19(3): 273–309.
Labov, William. 1972. *Sociolinguistic patterns*. Philadelphia, PA: University of Pennsylvania Press.
Labov, William. 2001. *Principles of linguistic change, vol. 2: social factors*. Malden; Oxford: Blackwell.
Le Maguer, Sebastien, Simon King, and Naomi Harte. 2024. The limits of the mean opinion score for speech synthesis evaluation. *Computer Speech & Language* 84: 101577.
Lee, Hyunjung. 2020. Cross-generational acoustic comparisons of tonal Kyungsang Korean stops. *The Journal of the Acoustical Society of America* 148(2): EL172–EL178.
Lee, Hyunjung and Allard Jongman. 2019. Effects of sound change on the weighting of acoustic cues to the three-way laryngeal stop contrast in Korean: Diachronic and dialectal comparisons. *Language and Speech* 62(3): 509–530.
Lee, Hyunjung, Jeffrey J. Holliday, and Eun Jong Kong. 2020. Diachronic change and synchronic variation in the Korean stop laryngeal contrast. *Language and Linguistics Compass* 14(7): e12374.
Lee, Hyunjung, Stephen Politzer-Ahles, and Allard Jongman. 2013. Speakers of tonal and non-tonal Korean dialects use different cue weightings in the perception of the three-way laryngeal stop contrast. *Journal of Phonetics* 41(2): 117–132.
Lev-Ari, Shiri and Sharon Peperkamp. 2014. An experimental study of the role of social factors in sound change. *Laboratory Phonology* 5(3): 379–401.
Lo, Chen-Chou, Szu-Wei Fu, Wen-Chin Huang, Xin Wang, Junichi Yamagishi, Yu Tsao, and Hsin-Min Wang. 2019. MOSNet: Deep learning based objective assessment for voice conversion. In Gernot Kubin and Zdravko Kačič (eds.), *Proceedings of Interspeech 2019*, 1541–1545. Graz: International Symposium on Computer Architecture.
Love, Jessica and Abby Walker. 2013. Football versus football: Effect of topic on /r/ realization in American and English sports fans. *Language and Speech* 56(4): 443–460.
Matuschek, Hannes, Reinhold Kliegl, Shravan Vasishth, Harald Baayen, and Douglas Bates. 2017. Balancing Type I error and power in linear mixed models. *Journal of Memory and Language* 94: 305–315.
Milroy, James and Lesley Milroy. 1978. Belfast: Change and variation in an urban vernacular. In Peter Trudgill (ed.), *Sociolinguistic patterns in British English*, 19–36. London: Edward Arnold Publisher.
Minixhofer, Christoph, Ondrej Klejch, and Peter Bell. 2025. TTSDS2: Resources and benchmark for objective evaluation of text-to-speech. In Sébastien Le Maguer, Matt Coler, Joakim Gustaffson, and Juraj Šimko (eds.), *Proceedings of the 13th Speech Synthesis Workshop (SSW 2025)*, 68–75. Baixas: International Symposium on Computer Architecture.

Morgan, Lucia C. 1969. North Carolina accents. *Southern Speech Journal* 34(3): 223–229.

Mori, Masahiro. 1970/2012. The uncanny valley (Karl F. MacDorman and Norri Kageki, Trans.). *IEEE Robotics & Automation Magazine* 19(2): 98–100.

Nass, Clifford, Jonathan Steuer, and Ellen R. Tauber. 1994. Computers are social actors. In Beth Adelson, Susan Dumais, and Judith Olson (eds.), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 72–78. New York, NY: ACM Publications.

Oh, Eunjin. 2011. Effects of speaker gender on voice onset time in Korean stops. *Journal of Phonetics* 39(1): 59–67.

Oh, Eunjin, Kaori Idemaru, and Boram Kim. 2018. The use of a voice onset time cue in the perception of Seoul Korean stops as a function of listener gender. *The Korean Journal of Linguistics* 43(4): 761–780.

Pardo, Jennifer S. 2006. On phonetic convergence during conversational interaction. *Journal of the Acoustical Society of America* 119(4): 2382–2393.

Pierrehumbert, Janet B. 2001. Exemplar dynamics: Word frequency, lenition and contrast. In Joan L. Bybee and Paul J. Hopper (eds.), *Frequency effects and the emergence of linguistic structure*, 137–158. Amsterdam: John Benjamins Publishing Company.

Podesva, Robert J. 2007. Phonation type as a stylistic variable: The use of falsetto in constructing a persona. *Journal of Sociolinguistics* 11(4): 478–504.

R Core Team. 2024. *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Remez, Robert E., Jennifer M. Fellowes, and Phillip E. Rubin. 1997. Talker identification based on phonetic information. *Journal of Experimental Psychology: Human Perception and Performance* 23(3): 651–666.

Schertz, Jessamyn, Yoonjung Kang, and Sungwoo Han. 2019. Sources of variability in phonetic perception: The joint influence of listener and talker characteristics on perception of the Korean stop contrast. *Laboratory Phonology* 10(1): 1–32.

Selkirk, Elisabeth. 1984. *Phonology and syntax: The relation between sound and structure*. Cambridge, MA: MIT Press.

Silva, David James. 1992. *The phonetics and phonology of stop lenition in Korean*. PhD Dissertation. Cornell University.

Silva, David James. 2002. Consonant aspiration in Korean: A retrospective. In Sang-Oak Lee and Gregory K. Iverson (eds.), *Pathways into Korean language and culture: Essays in honor of Young-Key Kim-Renaud*, 447–469. Seoul: Pagijong Press.

Silva, David James. 2006. Acoustic evidence for the emergence of tonal contrast in contemporary Korean. *Phonology* 23(2): 287–308.

Sneller, Betsy. 2024. Gritty Philadelphia: Orientation to local ideology as a predictor of sound change. *Language Variation and Change* 36(3): 333–354.

So, Hyunjung and Jonny Jungyun Kim. 2024. Age-indexed perceptual cue-shifting of Korean stops in various prosodic positions. *Linguistic Research* 41(3): 367–390.

Strand, Elizabeth A. and Keith Johnson. 1996. Gradient and visual speaker normalization in the perception of fricatives. In Dafydd Gibbon (ed.), *Natural language processing and speech technology: Results of the 3rd KONVENS conference*, 14–26. Berlin; Boston: De Gruyter Mouton.

Székely, Éva, Jura Miniota, and Míša Hejná. 2025. Will AI shape the way we speak? The emerging sociolinguistic influence of synthetic voices. In Maria Ines Torres, Yuki Matsuda, Zoraida Callejas, Arantza del Pozo, and Luis Fernando D'Haro (eds.), *Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology*, 335–340. Bilbao: Association for Computational Linguistics.

Trudgill, Peter. 1972. Sex, covert prestige and linguistic change in the urban British English of Norwich. *Language in Society* 1(2): 179–195.

Varadhan, Praveen Srinivasa, Sherry Thomas, M. S. Sai Teja, Suvrat Bhooshan, and Mitesh M. Khapra. 2025. The state of TTS: A case study with human fooling rates. In Odette Scharenborg, Catharine Oertel, Khiet Truong (eds.), *Proceedings of Interspeech 2025*, 2285–2289. Baixas: International Symposium on Computer Architecture.

Viswanathan, Mahesh and Madhubalan Viswanathan. 2005. Measuring speech quality for text-to-speech systems: Development and assessment of a modified mean opinion score (MOS) scale. *Computer Speech & Language* 19(1): 55–83.

Walker, Abby and Jennifer Hay. 2011. Congruence between `word age' and `voice age' facilitates lexical access. *Laboratory Phonology* 2(1): 219–237.

Wang, Hualei, Na Li, Chuke Wang, Shu Wu, Zhifeng Li, and Dong Yu. 2025. Vox-Evaluator: Enhancing stability and fidelity for zero-shot TTS with a multi-level evaluator. arXiv preprint, arXiv:2510.20210.

Winter, Bodo. 2020. *Statistics for linguists: An introduction using R*. New York, NY: Routledge.

Wright, Jonathan D. 2007. *Laryngeal contrast in Seoul Korean*. PhD Dissertation. University of Pennsylvania.

Yi, Han-Gyol, Jasmine E. B. Phelps, Rajka Smiljanic, and Bharath Chandrasekaran. 2013. Reduced efficiency of audiovisual integration for nonnative speech. *The Journal of the Acoustical Society of America* 134(5): EL387–EL393.

Zellou, Georgia, Michelle Cohn, and Tyler Kline. 2021. The influence of conversational role on phonetic alignment toward voice-AI and human interlocutors. *Language, Cognition and Neuroscience* 36(10): 1298–1312.

Zhang, Yinjun. 2025. *Dynamic adaptation in speech perception: Native transfer and social cue integration in L2 English*. Seoul: The Korea Institute for Humanities and Social Sciences (KIHSS).

Appendix

In Table 3, each of the 60 TTS speakers is identified by their 'Voice ID' and 'Platform'. To measure perceived ages, participants were instructed to provide a specific number (e.g., 19, 20, 21), rather than an age range (e.g., 20s, about 20, or 19-21). Gender was rated on a 5-point scale from –2 (prototypically female), through 0 (completely neutral), to 2 (prototypically male). Naturalness was rated similarly from –2 (completely machine-like) to 2 (completely human-like). For dialect, participants were first asked whether each voice sounded like *saturi* (*사투리*; i.e., any non-Seoul dialect). If so, they were prompted to specify the dialect region from Chungcheong, Jeolla, or Gyeongsang. Table 3 presents the number and percentage of responses indicating non-Seoul dialects.

Table 3. TTS speaker information and sociolinguistic rating results

	Plat-form	Voice ID	Rating results (mean, max, min, sd)			
			Age	Gender	Naturalness	Non-Seoul responses
OF	Naver	유진	(**31.52**, 46.00, 23.00, 6.93)	(**-1.66**, -1.00, -2.00, 0.48)	(**-0.68**, 1.00, -2.00, 1.12)	4 (13%)
		진아	(**36.65**, 61.00, 17.00, 9.38)	(**-1.76**, -1.00, -2.00, 0.43)	(**-0.29**, 2.00, -2.00, 1.17)	3 (10%)
		희그리다	(**37.68**, 52.00, 25.00, 6.68)	(**-1.55**, 0.00, -2.00, 0.56)	(**-0.90**, 1.00, -2.00, 1.12)	1 (3%)
		선경	(**37.94**, 65.00, 23.00, 9.78)	(**-1.79**, -1.00, -2.00, 0.41)	(**-0.52**, 2.00, -2.00, 1.19)	1 (3%)
		선희	(**48.77**, 80.00, 13.00, 17.81)	(**-0.03**, 2.00, -2.00, 1.33)	(**-0.61**, 2.00, -2.00, 1.29)	3 (10%)
	Type cast	선영	(**40.94**, 67.00, 25.00, 9.55)	(**-1.86**, -1.00, -2.00, 0.34)	(**-0.23**, 2.00, -2.00, 1.21)	0 (0%)
		주하	(**46.48**, 66.00, 31.00, 9.65)	(**-1.66**, 0.00, -2.00, 0.54)	(**0.13**, 2.00, -2.00, 1.10)	1 (3%)
		인선	(**53.13**, 86.00, 26.00, 13.66)	(**-1.48**, 0.00, -2.00, 0.68)	(**-0.10**, 2.00, -2.00, 1.25)	6 (19%)
		순이	(**62.71**, 87.00, 43.00, 10.47)	(**-1.72**, 0.00, -2.00, 0.52)	(**0.03**, 2.00, -2.00, 1.20)	22 (71%)
		정순	(**64.03**, 80.00, 44.00, 8.88)	(**-1.76**, -1.00, -2.00, 0.43)	(**-0.03**, 2.00, -2.00, 1.28)	20 (65%)
	We Make	안아림	(**36.19**, 46.00, 22.00, 6.21)	(**-1.86**, -1.00, -2.00, 0.34)	(**-0.61**, 2.00, -2.00, 1.34)	0 (0%)

	Voice	최민서	(**36.48**, 60.00, 25.00, 7.41)	(**-1.55**, -1.00, -2.00, 0.50)	(**-0.94**, 2.00, -2.00, 1.13)	0 (0%)
		박하은	(**38.10**, 52.00, 25.00, 7.77)	(**-1.72**, -1.00, -2.00, 0.45)	(**-0.29**, 2.00, -2.00, 1.25)	1 (3%)
		박채은	(**38.35**, 75.00, 20.00, 10.96)	(**-1.52**, -1.00, -2.00, 0.50)	(**-0.90**, 2.00, -2.00, 1.00)	2 (6%)
		박예숙	(**50.26**, 85.00, 11.00, 20.01)	(**-0.31**, 2.00, -2.00, 1.26)	(**-0.58**, 1.00, -2.00, 1.10)	4 (13%)
	OF mean		(**43.95**, 67.20, 24.87, 10.34)	(**-1.48**, -0.33, -2.00, 0.58)	(**-0.43**, 1.80, -2.00, 1.18)	4.53 (13.4%)
OM	Naver	코맙소	(**36.39**, 87.00, 18.00, 13.73)	(**1.69**, 2.00, 1.00, 0.46)	(**-0.29**, 2.00, -2.00, 1.49)	8 (26%)
		김종익	(**38.19**, 79.00, 11.00, 18.07)	(**1.38**, 2.00, 0.00, 0.67)	(**-0.32**, 2.00, -2.00, 1.40)	9 (29%)
		수아아빠	(**47.42**, 80.00, 23.00, 13.95)	(**1.62**, 2.00, 1.00, 0.49)	(**0.39**, 2.00, -2.00, 1.41)	12 (39%)
		이션	(**47.87**, 67.00, 25.00, 11.04)	(**1.59**, 2.00, -2.00, 0.81)	(**-0.35**, 2.00, -2.00, 1.36)	3 (10%)
		원탁	(**55.74**, 89.00, 16.00, 13.86)	(**1.76**, 2.00, 1.00, 0.43)	(**0.16**, 2.00, -2.00, 1.14)	3 (10%)
	Type cast	중현	(**51.71**, 75.00, 27.00, 10.83)	(**1.79**, 2.00, 1.00, 0.41)	(**0.13**, 2.00, -2.00, 1.13)	2 (6%)
		덕환	(**57.58**, 80.00, 33.00, 13.64)	(**1.62**, 2.00, -2.00, 0.81)	(**-0.26**, 2.00, -2.00, 1.52)	16 (52%)
		창배	(**59.71**, 89.00, 34.00, 16.34)	(**1.76**, 2.00, 0.00, 0.50)	(**0.03**, 2.00, -2.00, 1.23)	7 (23%)
		영길	(**66.52**, 81.00, 35.00, 10.19)	(**1.86**, 2.00, 1.00, 0.34)	(**-0.06**, 2.00, -2.00, 1.24)	13 (42%)
		덕춘	(**74.03**, 92.00, 50.00, 10.09)	(**1.83**, 2.00, 1.00, 0.38)	(**0.06**, 2.00, -2.00, 1.29)	15 (48%)
	We Make Voice	강윤우	(**45.58**, 72.00, 33.00, 10.55)	(**1.62**, 2.00, 0.00, 0.55)	(**-0.65**, 2.00, -2.00, 1.23)	10 (32%)
		김정우	(**45.81**, 60.00, 23.00, 9.55)	(**1.83**, 2.00, 1.00, 0.38)	(**-0.42**, 2.00, -2.00, 1.21)	2 (6%)
		노영식	(**50.29**, 87.00, 13.00, 18.68)	(**1.62**, 2.00, 1.00, 0.49)	(**-0.03**, 2.00, -2.00, 1.31)	9 (29%)
		박춘배	(**58.32**, 80.00, 35.00, 11.28)	(**1.86**, 2.00, 1.00, 0.34)	(**0.55**, 2.00, -2.00, 1.04)	3 (10%)
		정태식	(**58.97**, 89.00, 21.00, 14.83)	(**1.72**, 2.00, 1.00, 0.45)	(**0.10**, 2.00, -2.00, 1.25)	14 (45%)
	OM mean		(**52.94**, 80.47, 26.47, 13.11)	(**1,70**, 2.00, 0.40, 0.50)	(**-0.06**, 2.00, -2.00, 1.28)	8.4 (27.13%)
YF	Naver	지요	(**15.42**, 32.00,	(**-1.69**, 0.00,	(**0.26**, 2.00,	2 (6%)

			10.00, 4.90)	-2.00, 0.53)	-2.00, 1.14)	
		유나	(**16.42**, 29.00, 10.00, 5.40)	(**-0.59**, 2.00, -2.00, 1.19)	(**-0.26**, 2.00, -2.00, 1.19)	1 (3%)
		기서	(**20.39**, 46.00, 11.00, 7.36)	(**-0.86**, 2.00, -2.00, 1.14)	(**-0.58**, 2.00, -2.00, 1.13)	5 (16%)
		소현	(**20.71**, 32.00, 13.00, 5.88)	(**-1.31**, 0.00, -2.00, 0.83)	(**-0.13**, 2.00, -2.00, 1.13)	2 (6%)
		민영	(**24.68**, 36.00, 13.00, 6.32)	(**-1.69**, 0.00, -2.00, 0.59)	(**-0.32**, 2.00, -2.00, 1.42)	1 (3%)
	Type cast	예린	(**21.77**, 31.00, 11.00, 4.96)	(**-1.72**, -1.00, -2.00, 0.45)	(**0.03**, 2.00, -2.00, 1.03)	0 (0%)
		유라	(**25.77**, 35.00, 17.00, 4.58)	(**-1.62**, -1.00, -2.00, 0.49)	(**-0.16**, 2.00, -2.00, 1.08)	1 (3%)
		시연	(**25.84**, 46.00, 17.00, 6.84)	(**-1.69**, 0.00, -2.00, 0.59)	(**0.06**, 2.00, -2.00, 1.08)	2 (6%)
		세희	(**26.65**, 44.00, 20.00, 5.62)	(**-1.45**, 0.00, -2.00, 0.72)	(**0.03**, 2.00, -2.00, 1.20)	3 (10%)
		이나	(**26.68**, 42.00, 15.00, 7.03)	(**-1.41**, 0.00, -2.00, 0.67)	(**-0.19**, 2.00, -2.00, 1.15)	5 (16%)
	We Make Voice	강유나	(**17.29**, 26.00, 10.00, 4.20)	(**-1.31**, 1.00, -2.00, 1.05)	(**0.00**, 2.00, -2.00, 1.14)	1 (3%)
		김지원	(**19.52**, 45.00, 11.00, 7.13)	(**-0.93**, 1.00, -2.00, 1.08)	(**-0.35**, 2.00, -2.00, 1.06)	3 (10%)
		김채원	(**20.94**, 30.00, 13.00, 4.64)	(**-1.31**, 0.00, -2.00, 0.83)	(**-0.29**, 2.00, -2.00, 1.14)	3 (10%)
		한여름	(**21.13**, 43.00, 10.00, 6.95)	(**-1.86**, -1.00, -2.00, 0.34)	(**-0.35**, 2.00, -2.00, 1.36)	4 (13%)
		한선희	(**31.97**, 60.00, 20.00, 8.43)	(**-1.83**, 0.00, -2.00, 0.46)	(**-0.74**, 2.00, -2.00, 1.19)	1 (3%)
	YF mean		(**22.35**, 38.47, 13.40, 6.02)	(**-1.42**, 0.20, -2.00, 0.73)	(**-0.20**, 2.00, -2.00, 1.16)	2.27 (7.2%)
YM	Naver	종혁	(**26.74**, 50.00, 11.00, 8.60)	(**1.52**, 2.00, 1.00, 0.50)	(**-0.52**, 2.00, -2.00, 1.10)	1 (3%)
		이안	(**27.06**, 40.00, 13.00, 6.29)	(**1.34**, 2.00, 0.00, 0.60)	(**-0.35**, 2.00, -2.00, 1.00)	1 (3%)
		동현	(**28.35**, 49.00, 13.00, 7.86)	(**1.55**, 2.00, 0.00, 0.56)	(**-0.45**, 1.00, -2.00, 0.87)	1 (3%)
		대성	(**28.48**, 41.00, 15.00, 6.02)	(**1.59**, 2.00, 1.00, 0.49)	(**-0.55**, 1.00, -2.00, 1.10)	1 (3%)
		최무비	(**29.77**, 53.00, 13.00, 7.89)	(**1.41**, 2.00, 1.00, 0.49)	(**-0.58**, 1.00, -2.00, 1.04)	2 (6%)
	Type cast	지훈	(**28.74**, 45.00, 19.00, 6.03)	(**1.72**, 2.00, 1.00, 0.45)	(**-0.03**, 2.00, -2.00, 1.18)	3 (10%)

		진우	(**29.52**, 55.00, 17.00, 7.40)	(**1.62**, 2.00, 0.00, 0.55)	(**0.03**, 2.00, -2.00, 1.15)	1 (3%)
		서준	(**29.81**, 38.00, 23.00, 4.75)	(**1.83**, 2.00, 1.00, 0.38)	(**0.26**, 2.00, -2.00, 1.05)	0 (0%)
		김건	(**32.39**, 46.00, 21.00, 6.06)	(**1.86**, 2.00, 1.00, 0.34)	(**0.29**, 2.00, -2.00, 1.14)	1 (3%)
		현우	(**32.84**, 48.00, 20.00, 6.97)	(**1.76**, 2.00, 1.00, 0.43)	(**0.65**, 2.00, -2.00, 1.09)	1 (3%)
	We Make Voice	유상진	(**22.97**, 38.00, 14.00, 6.34)	(**1.69**, 2.00, 1.00, 0.46)	(**0.48**, 2.00, -2.00, 1.27)	2 (6%)
		임도윤	(**26.16**, 39.00, 11.00, 7.84)	(**1.41**, 2.00, 1.00, 0.49)	(**-0.23**, 2.00, -2.00, 1.21)	4 (13%)
		오민준	(**30.87**, 45.00, 18.00, 7.90)	(**1.72**, 2.00, 0.00, 0.52)	(**-0.39**, 2.00, -2.00, 1.36)	1 (3%)
		이익준	(**32.42**, 45.00, 16.00, 6.42)	(**1.83**, 2.00, 1.00, 0.38)	(**-0.52**, 2.00, -2.00, 1.36)	1 (3%)
		남하진	(**32.58**, 50.00, 21.00, 5.88)	(**1.69**, 2.00, 0.00, 0.53)	(**0.42**, 2.00, -2.00, 1.24)	1 (3%)
	YM mean		(**29.25**, 45.47, 16.33, 6.82)	(**1.63**, 2.00, 0.60, 0.48)	(**-0.10**, 1.80, -2.00, 1.14)	1.4 (4.33%)

Yein Kang
Master's Student
Department of English Language and Literature
Pusan National University
2, Busandaehak-ro 63beon-gil, Geumjeong-gu,
Busan, 46241 Korea
E-mail: zaqs1578@gmail.com

Jonny Jungyun Kim
Associate Professor
Department of English Language and Literature
Pusan National University
2, Busandaehak-ro 63beon-gil, Geumjeong-gu,
Busan, 46241 Korea
E-mail: jonnykkim@gmail.com

Received: 2025. 01. 05.
Revised: 2026. 01. 27.
Accepted: 2026. 02. 12.