Exploring the Speech-to-Song Illusion: A Comparative Study of Standard Korean and Dialects

Haesun Joung[1] **Ahyeon Choi**[1] **Kyogu Lee**[1,2,3]
[1] Department of Intelligence and Information, Seoul National University
[2] Interdisciplinary Program in Artificial Intelligence, Seoul National University
[3] Artificial Intelligence Institute, Seoul National University
{gotjs3841, chah0623, kglee}@snu.ac.kr

Abstract

The Speech-to-Song Illusion (STS) phenomenon, where repeated short speech utterances transform into perceived song, has drawn attention to its underlying mechanisms and cross-linguistic differences. This study examines the STS effects among Korean speakers, comparing standard Korean (non-tonal) and dialects such as Gyeongsang (pitch-accent, tonal) and Jeju (non-tonal but intonation-rich), which exhibit varying levels of linguistic tonal features. Participants (N = 60), evenly divided between standard and dialect users, evaluated 180 auditory stimuli comprising standard Korean, Gyeongsang, and Jeju utterances under controlled repetition conditions. Results revealed significant STS effects across all groups and stimuli, with stronger effects observed for dialectal stimuli, particularly Jeju, compared to standard Korean. Interestingly, differences between standard and dialect speaker groups in STS perception were not statistically significant, suggesting that exposure to diverse linguistic environments, facilitated by modern Korean media, may homogenize perceptual responses to tonal variations. The study highlights the influence of tonal and rhythmic elements in STS perception and underscores the cultural and linguistic uniqueness of Korean as a fertile ground for exploring auditory illusions. This research contributes to understanding the interplay of linguistic and perceptual factors in STS and opens avenues for cross-cultural comparisons and neuroscientific investigations of auditory illusions.

Keywords: Auditory Illusion; Speech-to-Song Illusion; Tonal and Non-Tonal Languages; Korean Dialects;

Introduction

The phenomenon of auditory illusions has fascinated researchers for decades, with a variety of intriguing examples capturing the complexities of auditory perception. Pioneering studies on auditory illusions have significantly advanced our understanding of sound perception. These include research on the Scale Illusion, where the brain reorganizes fragmented tones into continuous musical scales (Deutsch, 1975); Shepard Tones, which create the auditory illusion of a pitch that appears to rise or fall endlessly (Shepard, 1964); the Octave Illusion, where alternating high and low tones are perceived as constant in one ear (Deutsch, 1974); and the Speech-to-Song Illusion, where repeated speech transforms into a song-like perception (Deutsch, Henthorn, & Lapidis, 2011). Among these, the Speech-to-Song (STS) illusion holds a particularly captivating place due to its unique interplay between speech and music perception.

The STS illusion, first systematically investigated by Diana Deutsch (2011), involves the transformation of a spoken phrase into a song-like experience through repetitive playback. This phenomenon underscores how repetition can alter the perceptual context of auditory stimuli. Deutsch and others have explored potential contributors to the STS effect, focusing on factors such as the pitch, rhythm, and phonetic structure of the stimuli. However, these studies have often yielded conflicting results, with some identifying rhythm as a key factor and others highlighting the role of pitch (Tierney, Patel, & Breen, 2018). Moreover, cultural and linguistic differences appear to further modulate the illusion, adding layers of complexity.

A notable line of research has examined how tonal languages, such as Mandarin and Thai, affect the STS illusion. Early work by Jaisin et al. (2016) showed that native speakers of tonal languages tend to experience a diminished STS effect relative to non-tonal language speakers, presumably because pitch variations carry lexical meaning. More recently, however, Kachlicka et al. (2024) reported no reliable difference between tonal and non-tonal language speakers, arguing instead for the cross-cultural universality of the illusion. These conflicting findings leave open the question of whether the STS illusion is governed by universal perceptual mechanisms or modulated by linguistic experience. Korean, which uniquely combines non-tonal Standard speech with pitch-accent or intonation-rich dialects, provides an ideal testing ground for disentangling these possibilities.

Our study uniquely investigates the interplay of tonal and non-tonal features within a single linguistic framework, as exemplified by Korean, where both tonal material (e.g., the pitch-accented Gyeongsang dialect) and non-tonal material (e.g., the Jeju dialect and Standard Korean) coexist. Unlike previous studies, which typically relied on participants from different countries speaking tonal and non-tonal languages, our research benefits from the fact that Korean speakers are exposed to and can largely understand both tonal and non-tonal speech within their native language. This distinction allows us to explore the Speech-to-Song Illusion (STS) within a culturally and linguistically unified participant pool.

Furthermore, Korean provides a unique opportunity to examine dialects like Jeju, which, while part of the Korean language, pose significant challenges for complete linguistic comprehension even among native Korean speakers. This inclusion of both comprehensible and less comprehensible dialects as stimuli adds another layer of depth to our investigation. By leveraging these distinctive features of Korean, our

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

study offers unprecedented insights into how tonal and linguistic contexts shape the STS Illusion, advancing beyond the limitations of cross-linguistic studies that require participants from disparate linguistic backgrounds.

Hypotheses

This study aims to explore the Speech-to-Song Illusion (STS) effects among Korean speakers by examining linguistic and tonal influences. The following hypotheses are proposed:

H1. STS effects will be significant for all stimulus groups among Standard Korean speakers.

H2. Dialect speakers will exhibit weaker STS effects compared to standard Korean speakers.

H3. For standard Korean speakers, stimuli in regional dialects (e.g., Gyeongsang, Jeju) will elicit stronger STS effects than stimuli in standard Korean.

H4. Tonal foreign languages will induce stronger STS effects than non-tonal foreign languages for standard Korean speakers.

H5. For standard Korean speakers, unintelligible foreign languages will elicit stronger STS effects than intelligible ones.

Methods

Participants

Characteristics	Standard	Dialect
Participants (count)	30	30
Age (years)	22.87 (2.04)	23.53 (2.85)
Gender (M : F)	12 : 18	10 : 20
Music Edu. (years)	5.57 (3.81)	4.57 (5.28)
Music Major (count)	2	6

Table 1: Participant Characteristics: Demographics and Music Background

The study recruited a total of 60 native Korean speakers, divided into two groups: 30 speakers of Standard Korean and 30 speakers of dialects from the Gyeongsang and Daegu regions; none were native Jeju speakers. Table 1 summarizes the demographic and musical background characteristics of participants, including age, gender distribution, and years of musical training. Participants were aged between 19 and 28 years, ensuring a relatively homogenous young adult demographic. Individuals with a history of hearing disorders were excluded to maintain auditory consistency. Written informed consent was obtained from all participants, and they were informed about their right to withdraw from the study at any time. Ethics approval for the study was granted by the Institutional Review Board of Seoul National University. Participants were compensated financially for their time and participation in the experiment.

Stimuli

The stimuli comprised six distinct language categories: Standard Korean, Gyeongsang dialect, Jeju dialect, English, German, and Mandarin Chinese. Each category included 30 utterances, comprising a total of 180 stimuli. Each utterance was a short phrase or sentence lasting approximately 1–3 seconds.

The preparation process and data sources for each stimulus category are detailed below:

- All datasets provided recordings from single or multi male speakers, with 30 stimuli curated per dataset. Each set was carefully edited and standardized using Logic Pro X to ensure clarity and to preserve the tonal and prosodic characteristics essential for the study.

- For the Gyeongsang dialect, two male speakers from Busan and Pohang, who had lived in their respective regions for over 20 years, were recruited to record Gyeongsang dialect versions of Standard Korean sentences. A set of sentences was assigned to each speaker to ensure regional authenticity and natural intonation. From the collected recordings, 15 distinct utterances per speaker were carefully selected to avoid overlap.

Stimulus Design:

Each stimulus was presented in two stages: (1) Initial Playback, where the utterance was played once to capture its baseline perception, and (2) Repetition Playback, involving eight consecutive repetitions to evaluate its transformation into a song-like experience.

This carefully designed set of stimuli allowed the exploration of the Speech-to-Song (STS) illusion across diverse linguistic and prosodic contexts. By incorporating tonal material (Mandarin, Gyeongsang dialect) and non-tonal material (Standard Korean, Jeju dialect, English, German) languages, the stimuli facilitated a cross-linguistic analysis of the factors influencing the STS illusion.

Procedure

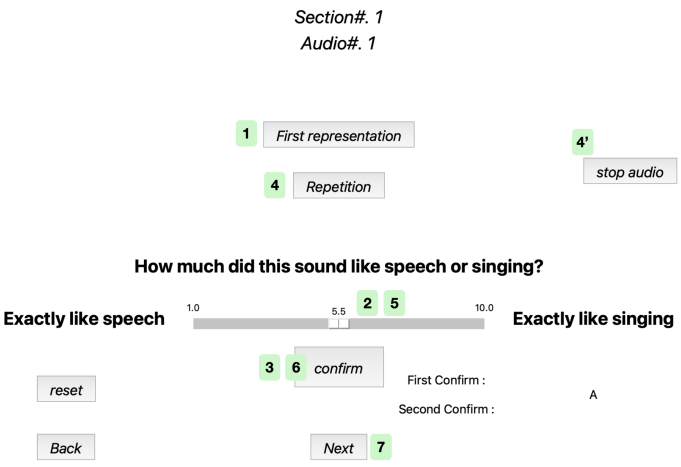

Figure 1: Actual experimental interface viewed by participants, with the experimental procedure steps numbered.

Language	Tonality	Comprehensibility	References	Speaker Characteristics	Sentence Selection Criteria
Standard Korean	Non-Tonal	O	Jo & Lee (2018) ; Selvy AI (2020)	Male, multiple speakers	Neutral, high-quality sentences
Gyeongsang Dialect	Tonal	O	Custom recordings (authors)	Male, 2 speakers (Busan, Pohang)	Standard Korean converted to dialect
Jeju Dialect	Non-Tonal	X	Park (2019)	Male, 1 speaker	pitch-shifting
English	Non-Tonal	O	Park (2017); Zen et al. (2024)	Multiple speakers	Neutral content with natural intonation
German	Non-Tonal	X	LT & Telecooperation Group (2015)	Multiple speakers	Prosodic diversity, neutral content
Mandarin Chinese	Tonal	X	MAGIC DATA (2019); Park (2018)	Multiple speakers	Short phrases preserving tonal features

Table 2: Categorization of Stimuli by Tonal Features, Comprehensibility, References, Speaker Characteristics, and Sentence Selection Criteria

The experiment was conducted in a controlled setting using a computer-based interface to present stimuli and record participant responses. Participants completed the experiment in individual sessions, each lasting approximately 80 minutes, including practice trials and breaks. The experimental procedure was designed as follows:

- **Practice Session:** Each participant first completed a practice session to familiarize themselves with the interface and evaluation process. During this session, a small subset of stimuli (not included in the main experiment) was presented, and participants were guided through the rating procedure.
- **Experimental Blocks:** The experiment was divided into six blocks, corresponding to the six language categories (Standard Korean, Gyeongsang dialect, Jeju dialect, English, German, Mandarin Chinese). The order of blocks was randomized for each participant to counterbalance potential order effects.
- **Stimulus Presentation:** Within each block, participants received 30 stimuli in random order. For each stimulus:
  1. The *first representation* of the stimulus was played once.
  2. Participants rated the extent to which the stimulus sounded like speech or a song using a continuous slider scale ranging from 1.0 (*entirely speech-like*) to 10.0 (*entirely song-like*).
  3. The *repetition* condition followed, where the same stimulus was played eight consecutive times.
  4. After the repetition, participants re-evaluated the stimulus using the same slider scale.

  Participants were encouraged to focus on their perception during the stimulus playback and to provide honest evaluations based on their subjective experience.
- **Final Session:** At the end of the experiment, participants were asked to provide a written response describing their criteria for judging a stimulus as 'song-like'. This open-ended question aimed to capture individual differences in perception and decision-making, revealing subjective factors such as rhythm, pitch variation, repetition, or other musical qualities they associated with the stimuli.

Participants were permitted to take short breaks between blocks to reduce fatigue and sustain attentiveness throughout the session. During data collection, participants provided ratings for both the initial presentation (first representation) and the repeated exposure (repetition) of each stimulus. The primary measure of the Speech-to-Song (STS) illusion was the difference in ratings between these two conditions.

Technical Setup: The experiment was implemented using Python with libraries such as `tkinter` for the graphical user interface and `simpleaudio` for audio playback. Participants interacted with the interface on a laptop or desktop computer. Each participant adjusted the audio volume to a comfortable level before starting the experiment. Responses were saved automatically in a CSV file for subsequent statistical analysis.

This procedure ensured a consistent and repeatable framework for assessing the STS illusion across all participants and language categories, allowing for a detailed comparison of linguistic and prosodic influences on the phenomenon.

Data Analysis

All statistical tests were planned around the five preregistered hypotheses (H1–H5).

Assumption checks. Shapiro-Wilk tests indicated that rating distributions deviated from normality ($p < .001$), so only non-parametric procedures were used. Hereafter, we denote the STS magnitude for each item as $\Delta = (\textit{last rating} - \textit{first rating})$.

Within-stimulus STS effect (H1). For each of the six stimulus categories, we compared the ratings *first* (single presentation) and *last* (after eight repetitions) with Wilcoxon signed-rank tests.

Between-group comparison (H2). To examine whether Dialect speakers show a different STS magnitude from Standard-Korean speakers, we contrasted their distributions of individual STS magnitudes using a Wilcoxon rank-sum test.

Dialect vs. Standard stimuli within listeners (H3). Within Standard-Korean listeners, pairwise Wilcoxon signed-rank tests compared the STS magnitudes elicited by Gyeongsang and Jeju stimuli against those elicited by Standard-Korean stimuli. A Bonferroni correction was applied for the two comparisons.

Tonal vs. non-tonal foreign languages (H4). To evaluate a possible tonal-language advantage, Wilcoxon signed-rank tests compared Mandarin with each non-tonal foreign language (English, German), with $\alpha_{adj} = .025$ after Bonferroni correction.

Intelligibility contrast (H5). Stimuli were additionally coded as *intelligible* (Standard Korean, Gyeongsang, English) or *unintelligible* (Jeju, German, Mandarin). A paired Wilcoxon test compared the average STS magnitudes between these two super-categories for Standard-Korean listeners.

Results

This study investigates the Speech-to-Song (STS) Illusion among speakers of Standard Korean and regional dialects, focusing on the effects of linguistic tonal features and cultural familiarity. Across all six language stimulus groups—Standard Korean, Gyeongsang dialect, Jeju dialect, English, German, and Chinese—the STS effect was found to be significant. While dialectal stimuli (Gyeongsang and Jeju) elicited stronger STS effects than Standard Korean, there was no statistically significant difference in STS perception between Standard Korean speakers and dialect speakers (p = 0.081).

Participant Group Analysis

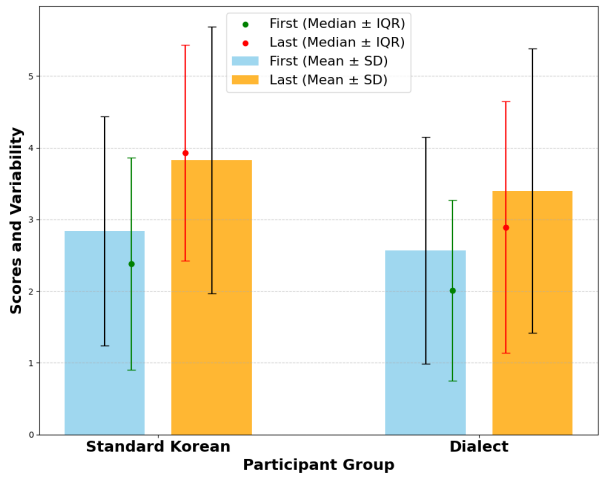

Figure 2: Comparison of STS effect between Standard Korean and Dialect speakers. Bars represent the mean scores (± SD) for first and last ratings, while dots and error bars indicate the median (± IQR).

Standard Korean Users vs. Dialect Users Participants were divided into two groups: 30 Standard Korean speakers and 30 speakers of regional dialects (Gyeongsang and Jeju). Both groups exhibited significant STS effects, with higher post-repetition ratings ("last" scores) compared to initial ratings ("first" scores). However, the magnitude of the STS effect did not differ significantly between the groups (Wilcoxon Signed-Rank test: $Z = -1.75, p = 0.081$). Across both groups, "last" ratings were more strongly correlated with the perceived song-like quality of stimuli compared to "first" ratings (Spearman correlation: $r = 0.47, p < 0.001$).

Stimuli-Specific Analysis

This section investigates the differences in STS effects across various linguistic stimuli, focusing on both Korean and foreign language categories to uncover patterns in perceptual responses. The results across six language categories, including median changes and variability, are summarized visually in Figure 3. Among foreign languages, English yielded the largest median STS effect, while Mandarin exceeded German but not English (pairwise Wilcoxon tests reported in the following).

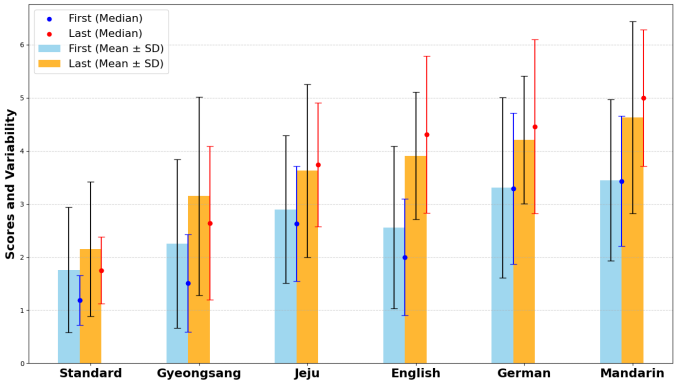

Figure 3: Stimuli-specific analysis showing the first and last ratings (median ± IQR and mean ± SD) across six language categories.

Stimulus Language	Z-score	p-value
Standard Korean	4.33	<.001
Gyeongsang Dialect	5.05	<.001
Jeju Dialect	5.57	<.001
English	6.28	<.001
German	5.53	<.001
Mandarin	5.51	<.001

Table 3: Z-Score and P-Value for Stimulus Languages

Comparison Across Korean Stimuli Among the three Korean stimuli categories—Standard Korean, Gyeongsang dialect, and Jeju dialect—all elicited significant STS effects ($p < 0.001$; Table 3). However, the magnitude of the STS

effect varied between these categories. Standard Korean stimuli resulted in the weakest STS effect, as indicated by lower mean differences between the first and last ratings. In contrast, the dialectal stimuli, particularly Jeju dialect, showed stronger STS effects.

Interestingly, while Jeju stimuli elicited slightly higher ratings than Gyeongsang stimuli, this difference was not statistically significant ($Z = 1.02, p = 0.31$).

Comparison Across Foreign Language Stimuli The foreign language stimuli—English, German, and Chinese—also demonstrated significant STS effects ($p < 0.001$; Table 3). Among these, English produced the largest Δ, Mandarin an intermediate Δ, and German the smallest; thus the pattern does not follow a simple tonal advantage.

Hypothesis-Driven Results

H1: Significant STS Effects for Standard Korean Speakers Each of the six stimulus categories produced a significant STS effect (Wilcoxon Z 4.33, $p < .001$; Table 3). For the three foreign languages, Mandarin ($Z = 5.51$), English ($Z = 6.28$), and German ($Z = 5.53$) all showed robust illusions.

H2: Dialect Users Will Exhibit Weaker STS Effects This hypothesis was not supported. While slight differences in STS magnitude were observed, they were not statistically significant ($p = 0.081$).

H3: Dialectal Stimuli Elicit Stronger STS Effects for Standard Korean Speakers This hypothesis was supported. For Standard Korean speakers, both Gyeongsang and Jeju dialect stimuli elicited stronger STS effects compared to Standard Korean stimuli ($p < 0.001$).

H4: Tonal foreign languages will elicit stronger STS effects than non-tonal foreign languages. Pairwise tests on Δ scores yielded a mixed pattern: Mandarin exceeded German ($Z = -2.62$, $p = .009$) but was outperformed by English ($Z = 2.52$, $p = .012$). Accordingly, H4 receives only *partial support*: a tonal advantage appears relative to German but not to English.

H5: Unintelligible foreign languages will elicit stronger STS effects than intelligible ones. H5 was not supported by the analysis. There was no statistically significant difference in STS effects between unintelligible languages (e.g., German) and intelligible ones (e.g., English) ($Z = -1.43, p = 0.1525$). These results suggest that the intelligibility of linguistic content plays a minimal role in influencing the STS effect.

Open-Ended Responses

Participants provided written descriptions of their criteria for perceiving speech as song, offering valuable insights into the perceptual and cognitive mechanisms underlying the STS illusion. A recurring theme in their responses was the importance of rhythm, melody, and repetition in shaping song-like perception. Many participants emphasized that a steady rhythmic structure or naturally emerging beat made speech feel more like a song, while others highlighted pitch variation and melodic contours as key determinants. Repetition was frequently mentioned as a factor that enhanced the perception of rhythm and melody, reinforcing prior findings on its role in STS. Additionally, some participants noted that linguistic meaning could interfere with the song-like perception, particularly for intelligible stimuli, whereas dialect accents with distinct pitch patterns were sometimes perceived as more musical. These responses illustrate the diverse and subjective nature of song perception, shaped by both acoustic features and linguistic context.

Discussion

This section addresses three key aspects of the STS illusion explored in this study. First, participants' subjective criteria for defining songs are analyzed to better understand the cognitive and perceptual basis of the illusion. Second, the results are compared with findings from previous studies to identify consistencies and discrepancies, shedding light on broader implications for linguistic and musical processing. Finally, the cognitive and cultural implications of the findings are discussed, emphasizing the role of linguistic and cultural contexts in shaping the STS illusion.

Participants' Responses on the Criteria for Songs

Participants' descriptions of their criteria for identifying speech as song provide valuable insights into the cognitive and perceptual processes underlying the STS illusion. A recurring theme in their responses was the role of rhythm, melody, and repetition in transforming speech into song-like perception.

Rhythm as a Key Factor: Many participants highlighted rhythm as a crucial determinant. One participant stated, *"It feels like a song when a beat is naturally imagined and it becomes hum-able"*. Another elaborated, *"There must be a steady rhythm upon which melody can build for something to be perceived as a song"*.

Pitch and Melodic Variation: Pitch and its variation were frequently mentioned as critical elements. For example, one participant noted, *"The criteria for songs involve perceiving a melody of highs and lows and smooth articulation"*. Another remarked, *"When words or sentences are connected smoothly to form a melody, they are perceived as songs"*.

The Role of Repetition: Repetition was also emphasized as a factor enhancing the perception of rhythm and melody. One participant explained, *"Repetition naturally creates rhythm, which makes it sound more like a song"*. This supports findings that repetition amplifies melodic and rhythmic salience in speech.

The Influence of Linguistic Meaning: Some participants indicated that linguistic meaning interfered with their perception of speech as song. For example, one stated, *"Dialect accents have unique pitch variations, but when meaning is clear, it doesn't feel like music"*. Another noted, *"Korean sentences are clear in meaning, but dialect accents might sound like songs to foreigners"*.

Variability in Song Criteria: Participants' criteria for songs varied widely, reflecting subjective and contextual factors. Some described rhythmic patterns and melodic sequences as primary indicators, while others focused on factors such as speech rate, phrasing, or familiarity. For instance, one participant stated, *"When stresses occur at regular intervals, it feels more like a song"*. Another added, *"The criteria were whether it doesn't sound like regular speech and whether it has a rhythm you can follow"*.

Comparison with Previous Studies and Broader Implications

This study's findings align with Deutsch's (2011) research, which highlights repetition as a key driver of the STS illusion. Repetition amplifies melodic and rhythmic patterns within speech, facilitating the perceptual transformation into song-like qualities. By including tonal stimuli such as Chinese and the pitch-accented Gyeongsang dialect, this study extends Deutsch's conclusions, demonstrating how tonal characteristics interact with repetition to enhance STS effects. This suggests that tonal variations can serve as an important acoustic feature, provided that linguistic meaning does not dominate auditory perception.

However, the results of this study contrast with Jaisin's (2016) findings, which suggested that tonal language speakers exhibit reduced STS effects due to their reliance on pitch for linguistic processing. Unlike Jaisin's observations, this study demonstrates that Korean speakers retain strong STS effects across various stimuli, regardless of their exposure to tonal or non-tonal linguistic environments. This discrepancy may be attributed to the unique linguistic structure of Korean.

Korean is a language that incorporates both tonal elements (e.g., dialects) and non-tonal elements (e.g., Standard Korean), providing speakers with extensive experience in navigating between these two pitch systems. This dual experience may enhance perceptual flexibility, allowing speakers to interpret pitch not only as a linguistic feature but also as a musical one. In particular, the pitch-accented Gyeongsang dialect may further reinforce this flexibility. Consequently, Korean speakers may develop the ability to process pitch effectively in both linguistic and musical contexts, which could explain the robust STS effects observed in this study.

These findings offer a compelling example of how the universal mechanisms underlying the STS illusion can be modulated or strengthened by specific linguistic experiences. Furthermore, this aligns with Kachlicka (2024), who emphasized the universality of the STS illusion. The dual linguistic characteristics of Korean provide a unique opportunity to explore the interaction between universal cognitive mechanisms and language-specific factors, offering novel insights into the study of the STS illusion.

Cognitive and Cultural Implications

The findings of this study underscore the shared cognitive and neural resources involved in processing speech and music. Neuroimaging studies have shown that the Speech-to-Song Illusion engages a broader network, including the inferior frontal gyrus (IFG), superior temporal gyrus (STG), and supplementary motor area (SMA), which facilitate auditory-motor integration during the speech-to-music transition (Tierney, Dick, Deutsch, & Sereno, 2013; Tsai & Li, 2019; Hymers et al., 2015). This convergence aligns with our results, emphasizing the critical role of repetition and tonal features in transforming speech into a song-like perception.

From a cultural perspective, the inclusion of diverse dialects and languages underscores the role of considering linguistic and cultural diversity in auditory perception research. The differential effects observed in language stimuli emphasize the role of cultural exposure in shaping auditory processing. The dual exposure of Korean speakers to tonal and non-tonal speech offers insight into cross-cultural differences in the perception of STS illusion. Standard Korean and regional dialects coexist, exposing individuals to varied accents through media from a young age. This environment familiarizes Standard Korean speakers with the acoustic features of dialects, which may influence their perception of dialectal accents as song-like. As one participant noted, *"Korean sentences are clear in meaning, but dialect accents might sound like songs to foreigners"*, reflecting how cultural and linguistic contexts influence auditory experiences.

Conclusion

This study investigated the Speech-to-Song Illusion (STS) among native Korean speakers using six language stimuli: Standard Korean, Gyeongsang dialect, Jeju dialect, English, German, and Chinese. The results confirmed that STS is strongly influenced by repetition, rhythm, and pitch. Among the Korean stimuli, the Jeju dialect produced the strongest STS effect, whereas among the foreign languages, English showed the largest Δ, with Mandarin's smaller gain reflecting its already elevated baseline—underscoring how baseline differences can modulate the apparent magnitude of the illusion. However, no significant difference was found between Standard Korean and dialect speakers.

Participant feedback highlighted repetition as the key factor in STS, with rhythm and melody playing a crucial role in transforming speech into song. In contrast, strong semantic content tended to weaken the STS effect, indicating an interaction between meaning and sound.

This study has limitations, including a restricted participant age range (19–28) and reliance on specific language stimuli. Future research should incorporate diverse linguistic backgrounds and use neuroimaging techniques such as fMRI or EEG to explore the neural mechanisms of STS. Future research should explore linguistic diversity and applications in music and language learning. This study advances understanding of language-music boundaries, laying groundwork for future STS research.

References

Deutsch, D. (1974). An auditory illusion. *Nature*, *251*(5473), 307–309.

Deutsch, D. (1975). Musical illusions. *Scientific American*, *233*(4), 92–105.

Deutsch, D., Henthorn, T., & Lapidis, R. (2011). Illusory transformation from speech to song. *The Journal of the Acoustical Society of America*, *129*(4), 2245–2252.

Hymers, M., Prendergast, G., Liu, C., Schulze, A., Young, M. L., Wastling, S. J., … Millman, R. E. (2015). Neural mechanisms underlying song and speech perception can be differentiated using an illusory percept. *NeuroImage*, *108*, 225–233.

Jaisin, K., Suphanchaimat, R., Figueroa Candia, M. A., & Warren, J. D. (2016). The speech-to-song illusion is reduced in speakers of tonal (vs. non-tonal) languages. *Frontiers in Psychology*, *7*, 662.

Jo, L., & Lee, W. (2018). *Korean open-source speech corpus for speech recognition by zeroth project.* Dataset available at http://www.openslr.org/40/.

Kachlicka, M., Patel, A. D., Liu, F., & Tierney, A. (2024). Weighting of cues to categorization of song versus speech in tone-language and non-tone-language speakers. *Cognition*, *246*, 105757. doi: https://doi.org/10.1016/j.cognition.2024.105757

LT and Telecooperation Group. (2015). *Open speech data corpus for german.* Dataset available at https://www.voxforge.org/home/forums/other-languages/german/open-speech-data-corpus-for-german.

MAGIC DATA Technology. (2019). *Mandarin chinese read speech corpus was developed by magic data technology co., ltd.* Dataset available at http://www.openslr.org/68/.

Park, K. (2017). *The world english bible speech dataset.* Dataset available at https://www.kaggle.com/bryanpark/the-world-english-bible-speech-dataset.

Park, K. (2018). *Chinese single speaker speech dataset.* Dataset available at https://www.kaggle.com/datasets/bryanpark/chinese-single-speaker-speech-dataset.

Park, K. (2019). *Jejueo single speaker speech dataset.* Dataset available at https://www.kaggle.com/bryanpark/jejueo-single-speaker-speech-dataset.

Selvas AI. (2020). *Emotional and conversational speech synthesis dataset.* Dataset available at https://github.com/emotiontts/emotiontts_open_db.

Shepard, R. N. (1964). Circularity in judgments of relative pitch. *The journal of the acoustical society of America*, *36*(12), 2346–2353.

Tierney, A., Dick, F., Deutsch, D., & Sereno, M. (2013). Speech versus song: multiple pitch-sensitive areas revealed by a naturally occurring musical illusion. *Cerebral Cortex*, *23*(2), 249–254.

Tierney, A., Patel, A. D., & Breen, M. (2018). Acoustic foundations of the speech-to-song illusion. *Journal of Experimental Psychology: General*, *147*(6), 888.

Tsai, C.-G., & Li, C.-W. (2019). Is it speech or song? effect of melody priming on pitch perception of modified mandarin speech. *Brain Sciences*, *9*(10), 286.