pISSN 2005-8063
eISSN 2586-5854
2026. 3. 31.
Vol.18 No.1
pp. 65-73

말소리와 음성과학

Phonetics and Speech Sciences

한국음성학회지

https://doi.org/10.13064/KSSS.2026.18.1.065

Testing prosodic boundary-induced phonetic categorization in AI-based automatic speech recognition*

Jiyoung Jang · Richard Hatcher**

Hanyang Institute for Phonetics and Cognitive Sciences of Language, Hanyang University, Seoul, Korea

Abstract

Phonetic categorization is shaped by systematic relationships between segmental cues and prosodic structure. In human speech perception, stop consonant voicing varies according to prosodic boundary strength, reflecting boundary-conditioned patterns of phonetic realization. This study examines whether such prosodic effects are observable in AI-based automatic speech recognition (ASR). We manipulated the voice onset time (VOT) of word-initial stop consonants along an English voiced–voiceless continuum while varying the presence of a major prosodic boundary preceding the target. The stimuli were presented to a state-of-the-art ASR model (Whisper), and the transcription outputs were analyzed to determine how voicing categorization varied across prosodic boundary contexts. Results showed an effect of VOT, with longer VOTs yielding higher probabilities of voiceless responses. While prosodic boundary condition interacted with VOT, Whisper did not exhibit a human-like boundary-dependent shift in the voicing category boundary, and these effects were contingent on the voicing of the original token and place of articulation. Particularly, global acoustic properties associated with the source exerted stronger influence on categorization, often overriding VOT cues. These findings suggest that while Whisper encodes sufficient acoustic detail to support coarse phonetic categorization, it does not recalibrate segmental cue interpretation for prosodic boundary structure. This study highlights a fundamental divergence between human perceptual normalization and end-to-end ASR inference, with implications for prosody-sensitive modeling of speech perception and recognition.

Keywords: AI-based automatic speech recognition, stop voicing perception, prosodic boundary, English

1. Introduction

Speech perception involves the interpretation of an acoustically variable signal in relation to its surrounding structural context. Voice onset time (VOT) is one of the primary acoustic cues used by listeners to distinguish voiced and voiceless stop consonants (Lisker

* This work was supported by the research fund of Hanyang University (HY-202300000003653). This work was also supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2021S1A5C2A02086884). We are grateful to the editorial team and three anonymous reviewers for their thoughtful comments, which substantially strengthened this work.

** richard.j.hatcher.jr@gmail.com, Corresponding author

Received 9 February 2026; Revised 5 March 2026; Accepted 5 March 2026

& Abramson, 1964, 1970). Yet even robust acoustic cues do not map onto phonological categories in a one-to-one fashion; their interpretation depends on listeners' expectations about how speech sounds are typically realized in particular linguistic environments. A large body of phonetic research has shown that human listeners exploit regularities associated with prosodic structure when categorizing speech sounds, supporting perceptual stability despite systematic variation in the speech signal (Kim & Cho, 2013; McQueen & Dilley, 2020; Mitterer et al., 2016; Steffman et al., 2022). However, it remains an open question whether contemporary automatic speech recognition (ASR) systems exhibit comparable sensitivity to such prosodically conditioned variation.

One well-established source of structured phonetic variation arises at prosodic boundaries (Cho, 2016; Cho & Keating, 2009; Fletcher, 2010; Fougeron & Keating, 1997). In many languages, consonants produced at the beginnings of higher-level prosodic domains exhibit strengthened articulatory realizations, including longer durations, increased articulatory displacement, and enhanced acoustic cues. In English stop consonants, this domain-initial strengthening is often manifested as increased VOT, particularly at major prosodic boundaries such as intonational phrase onsets. Importantly, this variation is not random: it reflects a systematic relationship between prosodic structure and phonetic realization that listeners can learn and exploit. As a result, human listeners' interpretation of ambiguous segmental cues is shaped by expectations associated with prosodic position.

Kim & Cho (2013) provided direct experimental evidence that perceived prosodic boundary strength influences how listeners categorize an upcoming stop consonant along a /b-p/ VOT continuum. In their study, both native English listeners and non-native Korean listeners showed a systematic rightward shift in the category boundary following an intonational phrase (IP) boundary compared to a word (Wd) boundary. That is, a longer VOT was required for a voiceless /p/ percept after an IP boundary, consistent with listeners' expectations about domain-initial strengthening. Crucially, this boundary-induced perceptual shift was observed regardless of listeners' language background, suggesting that listeners exploit abstract prosodic structure—reflected in boundary-related phonetic cues—when interpreting segmental contrasts.

The findings of Kim & Cho (2013) have informed theories of speech perception by demonstrating that phonetic categorization is context-dependent and guided by higher-level prosodic structure. Rather than relying solely on absolute phonetic values, listeners integrate segmental cues with expectations about prosodic organization, calibrating category judgments to match boundary-conditioned phonetic distributions. Prosodic boundary effects are therefore perceptually relevant regularities, and not merely articulatory side effects.

While such context-sensitive perception has been well-established for human listeners, it remains unclear whether comparable sensitivity emerges in computational systems trained on large-scale speech data. Despite rapid progress in ASR performance, relatively little is known about whether such systems internalize structured relationships between prosodic context and segmental realization, as opposed to relying on local acoustic cues alone. Modern ASR models are trained on massive amounts of naturalistic speech data in which prosodic boundaries and their phonetic correlates occur frequently and systematically. These systems must therefore learn to cope with boundary-conditioned variation in segmental realization, even though they are provided little to no annotation of prosodic structure during training. Whether such systems nevertheless learn structured relationships between prosodic context and phonetic realization remains largely unexplored.

Recent work has shown that neural ASR models encode substantial phonetic information and can capture fine-grained distinctions relevant to speech perception (Millet & Dunbar, 2022 ; Millet et al., 2019). These findings are consistent with the design of contemporary speech models, which rely on self-supervised representation learning and end-to-end sequence modeling to learn internal acoustic representations that reflect distinctions such as place and manner of articulation (Mohamed et al., 2022). However, most evaluations of ASR performance focus on recognition accuracy at the word or sentence level, leaving open how these systems behave when confronted with systematically manipulated phonetic ambiguity. In particular, it is unclear whether ASR models adjust their interpretation of segmental cues depending on prosodic boundary context, as human listeners do.

In examining prosodically conditioned phonetic categorization in automatic speech recognition, it is important to clarify the nature of the system being evaluated. The present study focuses on Whisper, a sequence-to-sequence ASR model that predicts (sub)word tokens rather than phonemes (Radford et al., 2023). As a result, distinctions such as stop voicing arise in this model indirectly through lexical hypotheses based on the entire acoustic signal (e.g., *ban* vs. *pan*), rather than through explicit segmental decisions.

Given this model architecture, the present study takes Kim & Cho (2013) as a point of departure and asks whether their core finding—prosodic boundary-induced modulation of phonetic categorization—can be replicated in ASR systems. Focusing specifically on the effect of prosodic boundary strength, we examine whether contemporary ASR models adjust their categorization of word-initial stop consonants along a voiced-voiceless VOT continuum depending on whether the target sound follows an intonational phrase boundary or a word boundary.

If the ASR system encodes and utilizes prosodically conditioned phonetic regularities in a manner comparable to human listeners, we expect categorization along the VOT continuum to shift as a function of preceding prosodic boundary strength, with longer VOTs required for voiceless classifications following stronger boundaries. Alternatively, if ASR categorization is driven primarily by local acoustic cues without prosodic normalization, categorization should remain stable across boundary conditions, or reflect only global acoustic differences unrelated to prosodic structure.

Investigating this issue is theoretically informative for both phonetics and speech technology. This comparison allows us to assess whether prosodically conditioned phonetic interpretation emerges naturally from large-scale distributional learning, or instead depends on specialized perceptual or representational mechanisms not captured by current end-to-end ASR models. From a phonetic perspective, ASR systems provide a test case for assessing which aspects of context-sensitive phonetic interpretation can emerge from data-driven learning alone, without explicit representations of prosodic structure or perceptual normalization mechanisms. Evidence of

boundary-conditioned categorization would indicate that structured prosodic–phonetic relationships are recoverable from the acoustic signal alone. In contrast, their absence would highlight a divergence between human perceptual strategies and current end-to-end approaches to speech recognition.

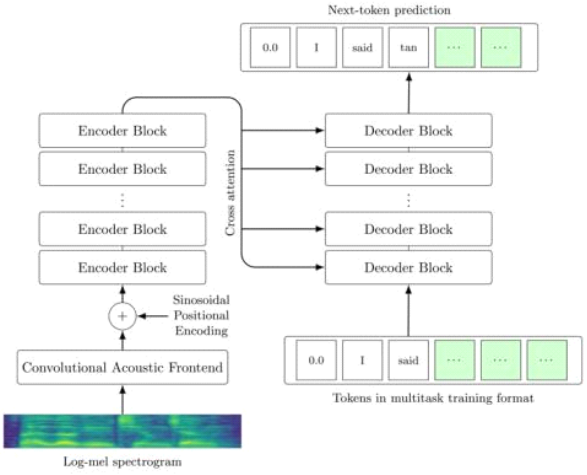
Using controlled speech stimuli modeled on the design of Kim & Cho (2013), we manipulate the VOT of target stops while varying only the presence or absence of a major prosodic boundary immediately preceding the target word. These stimuli are presented to a state-of-the-art ASR model (Whisper), and the model's outputs are analyzed to determine whether voicing classification shifts systematically across boundary conditions. Evidence for such shifts would indicate that ASR systems, like human listeners, incorporate expectations about boundary-conditioned phonetic realization when interpreting ambiguous segmental cues.

The present study does not attempt to examine the full range of prosodic factors known to influence phonetic categorization, but instead provides an initial test of whether prosodic boundary information alone is sufficient to induce context-dependent categorization effects in an ASR system. By directly paralleling a well-established human perception study, this work provides a focused test of whether prosodically conditioned phonetic expectations emerge in computational systems trained solely on large-scale speech data. In doing so, it lays the groundwork for follow-up research that more broadly explores how prosodic structure is represented and utilized in automatic speech recognition.

2. Methods

2.1. Automatic speech recognition (ASR) model

This study evaluates whether a state-of-the-art ASR system exhibits prosodic boundary-dependent modulation in phonetic categorization. We focus on a single ASR model, Whisper (small-v2; Radford et al., 2023). The model was used as released, without task-specific fine-tuning or adaptation to the experimental materials. All analyses therefore reflect patterns acquired during general training rather than optimization for the present task.



Log-Mel spectrograms are encoded using stacked transformer encoder blocks, and textual output is generated autoregressively by a transformer decoder conditioned on the encoded acoustic representation. Adapted from Radford et al. (2023) with CC-BY-4.0.

Figure 1. Schematic overview of the Whisper encoder–decoder architecture used in this study.

Whisper is an encoder–decoder transformer model that operates directly on log-Mel spectrogram inputs and generates text autoregressively as a sequence of (sub)word tokens (Figure 1). The encoder transforms the acoustic input into a learned latent representation, which is then used by the decoder to predict the most probable token (i.e., text) sequence given the preceding context. Because the model is trained to optimize transcription accuracy rather than to make explicit phonetic decisions, distinctions such as stop voicing are not represented as independent outputs but instead influence lexical predictions indirectly through the acoustic evidence available to the model.

2.2. Speech materials

The speech materials were modeled closely on those used in Kim & Cho (2013), with adaptations appropriate for ASR evaluation. Recordings were produced by a native speaker of American English (male) in a sound-attenuated booth with a Shure SM10 head-mounted microphone, digitized at 48 kHz/24-bit.

Target stimuli consisted of a word-initial stop consonant followed by the low vowel /æ/, forming a voiced-voiceless contrast at three places of articulation (bilabial, alveolar, and velar). Example minimal pairs include *ban-pan*, *Dan-tan*, and *gam-cam*. These targets were embedded in carrier sentences of the form *[Pronoun] said # [target word] today*, where the subject pronoun was *I, you, he, she,* or *they*. This carrier sentence frame was selected to minimize lexical or semantic bias toward any particular target word, ensuring that differences in model responses would primarily reflect phonetic and prosodic factors rather than contextual predictability.

In the carrier phrase above, the symbol "#" marks the location of a prosodic boundary immediately preceding the target-bearing syllable. Two boundary conditions were created: (1) IP boundary condition, in which the target word followed a major prosodic boundary; (2) Wd boundary condition, in which the target word occurred phrase-medially without a major boundary.

The prosodic boundary contrast was realized using acoustic cues characteristic of English prosodic structure. In the IP condition, the preceding material exhibited phrase-final lengthening and was followed by a brief pause, while the Wd condition lacked both lengthening and pause. The two boundary contexts differed systematically in prosodic structure while remaining comparable in overall segmental content. The recording design included three places of articulation, two voicing categories, two prosodic boundary conditions, five carrier sentences, and ten repetitions. A total of 600 sentences were recorded, of which six were excluded due to disfluencies.

2.3. Stimulus manipulation

To create controlled phonetic ambiguity, VOT continua were generated for English plosives (bilabial, alveolar, velar) using acoustic manipulation in Praat (Boersma, & Weenink, 2024). Starting from naturally produced tokens, the VOT of the target stop consonant was manipulated by adjusting the duration of aspiration noise following stop release while keeping the

closure, burst, and following vowel unchanged.

Fifteen VOT steps were created, spanning a range from short-lag to longer positive VOT values (e.g., 0–70 ms in 5 ms increments). This range was chosen to encompass values that are typically perceived as ambiguous between voiced and voiceless stops in English, following Kim & Cho (2013). Crucially, the VOT manipulation was identical across boundary conditions, ensuring that any differences in categorization could be attributed to the preceding prosodic context rather than to acoustic differences in the target segment itself.

Each stimulus was created by concatenating three components: (1) the sentence-initial portion (*e.g., they said*), produced with either an IP or Wd boundary; (2) the manipulated target syllable containing the VOT continuum; (3) the sentence-final portion (*today*). The full stimulus set consisted of all combinations of: prosodic boundary (IP vs. Wd); and VOT (15 levels along the voicing continuum). After exclusion of six disfluent recordings, each of the remaining 594 base tokens was expanded into a 15-step VOT continuum, yielding 8,910 stimuli.

2.4. AI-based automatic speech recognition (ASR) input and response extraction

Because Whisper outputs (sub)word tokens rather than phoneme-level representations, the model does not provide frame-synchronous phoneme posterior probabilities (Radford et al., 2023). Decoding used the default settings in the open-source Python implementation (temperature=0.0; beam_size=None), resulting in deterministic greedy decoding without sampling. We therefore operationalized the model's phonetic categorization behavior in terms of discrete lexical outcomes, treating the recognized token's form as the model's categorical response, while acknowledging that lexical hypotheses may exert strong constraints on phonetic categorization. This approach follows directly from the model's inference mechanism and parallels the forced-choice responses used in human phonetic categorization.

Each stimulus was presented to the Whisper model as an independent audio file. Whisper's transcription output was used to determine the model's categorization of the target stop consonant. Because Whisper sometimes merged words from the carrier phrase and target into a single token (e.g., *Disappend today* for *They said pan today*), categorization was based on the identity of the stop consonant corresponding to the manipulated segment in the model's output, regardless of token boundaries. Model responses were classified as *voiced* or *voiceless* depending on whether this consonant was realized as a voiced or voiceless plosive. Outputs that did not clearly map onto either category (e.g., deletions or substitutions unrelated to the target contrast such as *I said and today* for *I said pan today*) were excluded from analysis (approx. 3% of Whisper's output). Example output-to-category mappings for each minimal pair are provided in Table A1 in the Appendix.

2.5. Statistical analysis

We fit a logistic regression model predicting whether Whisper produced a voiceless (1) or voiced (0) stop response. The model was implemented in the *lme4* package (Bates et al., 2015) in R (R Core Team, 2025). Fixed effects included VOT (continuous), prosodic boundary condition (IP vs. Wd), source-token voicing (voiced vs. voiceless), and place of articulation, along with their interactions. Binary predictors were contrast-coded (IP=0.5, Wd=−0.5; voiced=0.5, voiceless=−0.5). This analysis tested whether prosodic boundary context shifted voicing categorization along the VOT continuum in Whisper's responses.

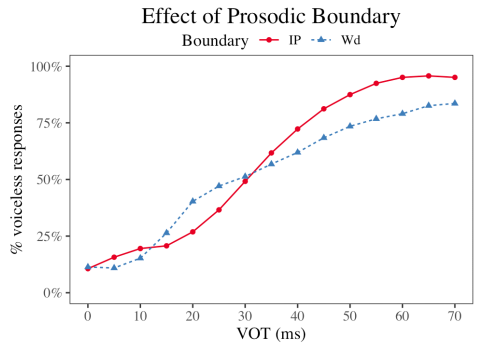
3. Results

The statistical model revealed a robust main effect of VOT (β=0.20, *SE*=0.01, z=19.90, p<.001), indicating that increases in VOT were associated with higher log-odds of a voiceless response in the reference condition. Significant two-way interactions showed that the effect of VOT was modulated by both boundary (β=0.05, *SE*=0.02, z=2.62, p=.009) and source-token voicing (β=−0.12, *SE*=0.02, z=−6.04, p<.001), as well as by place of articulation (bilabial: β=−0.05, *SE*=0.01, z=−4.20, p<.001; velar: β=−0.09, *SE*=0.01, z=−8.19, p<.001).

Crucially, these effects were further conditioned by higher-order interactions. A marginal three-way interaction between VOT, boundary, and source-token voicing was observed (β=−0.07, *SE*=0.04, z=−1.78, p=.075), suggesting that boundary-related modulation of the VOT effect differed across source-token voicing categories. This pattern was sharpened by significant four-way interactions involving place of articulation (bilabial: β=0.20, *SE*=0.05, z=3.99, p<.001; velar: β=0.18, *SE*=0.04, z=4.28, p<.001), indicating that the joint influence of VOT, boundary, and source-token voicing on response probability varied systematically across places of articulation. In the following subsections, we unpack these main and interaction effects across boundary conditions, source-token voicing categories, and places of articulation by examining both the empirical response curves (Figures 2–4) and model-derived crossover locations (VOT_{50}) below.

3.1. Overall stop categorization across prosodic boundaries

Figure 2 presents the proportion of voiceless responses as a function of VOT for the IP and Wd boundary conditions in the ASR model. Across both boundary conditions, Whisper exhibited a clear sensitivity to VOT: the proportion of voiceless responses increased as VOT lengthened, yielding a monotonic categorization function.



VOT, voice onset time; IP, intonational phrase; Wd, word.

Figure 2. Percentage of voiceless responses across the VOT continuum for IP and Wd boundary conditions.

Despite this sensitivity to VOT, the shape of the categorization function differed markedly between the ASR model and what is typically reported for human listeners. Human listeners exhibit a relatively steep categorization slope, with voicing judgments transitioning rapidly from voiced to voiceless over a narrow VOT range (Kim & Cho, 2013). In contrast, the ASR model displayed a substantially shallower slope, with a gradual increase in voiceless responses extending over a wider portion of the VOT continuum. This broader range of intermediate response proportions suggests that the ASR system maintains a larger region of phonetic ambiguity than is typically observed in human categorization behavior.

Descriptively, the empirical curves show boundary-related differences at certain VOT ranges. Between approximately 15–30 ms, fewer voiceless responses were observed in the IP condition than in Wd. At higher VOT values (approx. ≥35 ms), this pattern reversed. While these local differences are visible in the curves, it is necessary to examine the cross-over location directly to determine whether a systematic boundary-induced shift is present.

To that end, we estimated the VOT value yielding a 50% probability of a voiceless response (VOT_{50}) for each boundary condition. When averaged across all contexts, the estimated crossover was 28.2 ms for IP and 28.2 ms for Wd, yielding an overall ΔVOT_{50} of 0.04 ms [95% CI (−1.40, 1.61)]. The confidence interval is tightly centered around zero, indicating no reliable overall boundary-induced crossover shift. Importantly, the 95% confidence interval excludes shifts of the magnitude typically reported in human studies (approx. 5–10 ms), suggesting that any aggregate boundary effect in Whisper is substantially smaller than the human benchmark (Kim & Cho, 2013). Thus, although the response curves show localized boundary-related differences, these do not translate into a consistent overall shift in category crossover.

3.2. Effects of source-token voicing and place of articulation on stop categorization

Figure 3 illustrates the proportion of voiceless responses as a function of VOT, separated by the source-token's original voicing category. Across both conditions, the ASR model continued to show an increase in voiceless responses with increasing VOT, indicating sensitivity to VOT. However, the overall level and shape of the categorization curves differed substantially depending on source-token voicing.

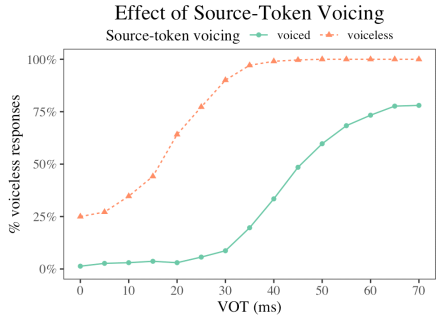
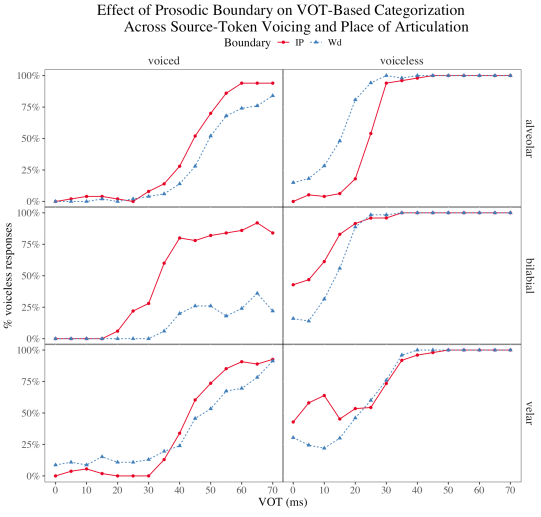


Figure 3. Percentage of voiceless responses across the voice onset time (VOT) continuum as a function of source-token voicing (voiced vs. voiceless).

When the target stop was embedded in stimuli derived from a source-token bearing a voiced stop, voiceless responses remained relatively low across much of the VOT continuum, increasing gradually and reaching ceiling only at higher VOT values. In contrast, stimuli derived from a source-token bearing a voiceless stop elicited substantially higher rates of voiceless responses across the continuum, even at shorter VOT values. This shift resulted in an earlier rise in the categorization curve in the voiceless source-token condition. These results indicate that global acoustic properties associated with the original source sentence exert a strong influence on Whisper's stop categorization behavior, beyond the manipulated VOT values.

We next examine how stop categorization patterns by prosodic boundary condition, source-token voicing, and place of articulation (Figure 4). Across alveolar, bilabial, and velar stops, the model remained sensitive to VOT (i.e., increased voiceless response with increase in VOT), but the steepness and alignment of the categorization curves varied by prosodic boundary context, source-token voicing, and place of articulation. Several descriptive patterns in the empirical curves are worth highlighting.



VOT, voice onset time; IP, intonational phrase; Wd, word.

Figure 4. Percentage of voiceless responses across the VOT continuum for IP and Wd boundary conditions by source-token voicing and place of articulation.

First, although most panels show a boundary effect in the opposite direction from what is typically observed in human listeners, one condition—alveolar stops in a voiceless frame sentence—exhibits a reversed pattern, with IP conditions showing an expected rightward shift. Second, for bilabial stops in voiced frame sentences under the Wd condition, the proportion of voiceless responses never reaches 50%, suggesting a strong bias from voicing cues in the broader acoustic context that overrides VOT-based modulation of categorization in the ASR model. Finally, for velar stops in voiceless frame sentences, shorter VOT values are associated with higher voiceless response rates (up to approx. 10 ms), followed by a decrease in the 15–25 ms range, yielding a non-monotonic pattern that diverges from expectations based on human perception.

Table 1. Prosodic boundary-induced shifts in category crossover (ΔVOT_{50}=IP−Wd) across source-token voicing and POA

source-token voicing	POA	IP	Wd	ΔVOT_{50} (ms)
voiced	alveolar	45.0	52.6	−7.55 [−9.92, −5.25]
voiceless	alveolar	23.9	13.5	10.4 [8.45, 12.2]
voiced	bilabial	37.4	NA	(no in-range crossover)
voiceless	bilabial	5.3	12.6	−7.25 [−10.7, −4.46]
voiced	velar	45.2	48.2	−2.97 [−6.33, 0.145]
voiceless	velar	11.6	18.5	−6.86 [−11.3, −2.61]

VOT, voice onset time; IP, intonational phrase; Wd, word; POA, place of articulation.

To quantify the crossover differences suggested by the curves in Figure 4, we estimated VOT_{50} for each Boundary×source-token voicing × place of articulation condition and computed the boundary-induced shift Equation (1). Table 1 presents these crossover shifts with bootstrapped 95% confidence intervals.

$$\Delta VOT50 = VOT\frac{IP}{50} - VOT\frac{Wd}{50} \quad (1)$$

The resulting pattern is highly context-dependent. For alveolar stops, the sign of the shift depends on source-token voicing. In the voiceless condition, a robust rightward shift is observed [ΔVOT_{50}=10.4 ms, 95% CI (8.45, 12.2)], corresponding to a later crossover in the IP condition and aligning with the direction typically reported in human listeners. In contrast, in the voiced alveolar condition the shift is re-versed and leftward [ΔVOT_{50}=−7.55 ms, 95% CI (−9.92, −5.25)], indicating earlier crossover in IP relative to Wd.

Across bilabial and velar stops, most estimable shifts are leftward. The voiceless bilabial [ΔVOT_{50}=−7.25 ms, 95% CI (−10.7, −4.46)] and voiceless velar [ΔVOT_{50}=−6.86 ms, 95% CI (−11.3, −2.61)] conditions both show significant negative shifts. The voiced velar condition exhibits a small negative shift (ΔVOT_{50}=−2.97 ms) whose confidence interval includes zero, indicating no reliable boundary-induced cross-over difference in that context.

In the voiced bilabial condition, the Wd curve does not reach the 50% criterion within the tested VOT range; accordingly, VOT_{50} and ΔVOT_{50} are undefined in-range. This absence of a crossover reflects a strong anchoring of categorization to the original voicing category in that context, overriding the manipulated VOT continuum.

These crossover estimates show that the direction and magnitude of boundary-induced shifts vary systematically with source-token voicing and place of articulation. Although a rightward shift consistent with human findings is observed in one context (voiceless alveolar), most other contexts exhibit leftward shifts or no reliable effect. Thus, when expressed in canonical categorical-perception terms, the boundary effect in Whisper is not stable across phonetic environments.

4. Discussion

4.1. Evidence for prosodic boundary effects in AI-based automatic speech recognition

The results indicate that Whisper exhibits systematic sensitivity to variation in VOT along the English voiced-voiceless continuum. As VOT increased, the probability of voiceless classification increased in a roughly monotonic fashion. However, the resulting categorization functions were substantially shallower than those typically observed for human listeners (Lisker & Abramson, 1964, 1970), suggesting that the model maintains a wide region of probabilistic ambiguity rather than forming sharply defined phonetic categories.

Prosodic boundary context did influence categorization, but not in a consistent or human-like manner. While some conditions showed differences between IP and prosodic Wd boundary contexts, the direction and magnitude of these effects varied depending upon the source-token's voicing and place of articulation. In several cases, the boundary effect was opposite in direction to that reported by Kim & Cho (2013) for human listeners, and in others it was absent or highly attenuated. These findings indicate that prosodic boundary information does not function as a stable conditioning factor for segmental cue interpretation in Whisper, in the current study. Rather than shifting the VOT boundary in a systematic way, boundary-related differences seem to act as just one more acoustic factor in the signal.

A much stronger influence on categorization came from the voicing of the stop in the original source token. Stops derived from originally voiced tokens tended to be recognized as voiced even at longer VOT values, while stops derived from originally voiceless tokens frequently remained voiceless despite substantial VOT shortening. This asymmetry indicates that manipulation of VOT alone was often insufficient to override other acoustic properties preserved during resynthesis. Rather than reflecting noise or inconsistency, this pattern suggests that Whisper integrates VOT with residual global acoustic properties of the original recording, leading to a strong anchoring effect toward the source category.

Overall, these results suggest that Whisper encodes sufficient acoustic detail to support coarse phonetic categorization of stop voicing, but does not implement prosodically conditioned normalization of segmental cues. Instead, voicing decisions emerge from the interaction of multiple acoustic factors, with global

acoustic properties often exerting stronger influence than local VOT manipulations. Prosodic boundary effects, when present, appear to arise as secondary consequences of these broader acoustic interactions rather than as evidence of abstract prosodic conditioning of phonetic interpretation.

4.2. Comparison with patterns in human speech perception

Compared with human listeners, Whisper's categorization pattern differs in several important ways. Kim & Cho (2013) found that both native English listeners and non-native Korean listeners exhibited a robust prosodic boundary effect in stop voicing perception: a longer VOT was required to identify a stop as voiceless following an intonational phrase boundary than following a word boundary. This shift was interpreted as reflecting listeners' expectations about domain-initial strengthening and the systematic relationship between prosodic structure and phonetic realization.

Whisper, in contrast, did not show a stable or uniform boundary-conditioned shift in its categorization patterns. Although differences between IP and Wd contexts were observed, their direction and magnitude depended strongly on the voicing of the source token and the stop's place of articulation. In many cases, the boundary effect was reversed relative to the pattern reported for human listeners, and in others it was small or absent. Rather than producing a consistent rightward shift in the VOT boundary following stronger prosodic boundaries, Whisper's responses suggest that boundary-related variation interacts with other acoustic properties in ways that are highly context-dependent.

Another clear difference lies in the sharpness of categorization. This contrasts with the steep category transitions typically reported in human perception (Lisker & Abramson, 1964, 1970). Whisper's broader and more gradual response curves indicate weaker commitment to discrete phonetic categories. This suggests that, unlike human listeners, the model does not form strongly delimited phonetic categories that can be predictively shifted by prosodic structure.

These differences point to a fundamental divergence in how prosodic information is used. Human listeners appear to incorporate prosodic structure as part of a predictive framework for interpreting segmental cues, adjusting category boundaries based on expectations about how sounds are realized in different prosodic positions. Whisper's behavior, in contrast, is better characterized as weighted acoustic pattern matching, in which prosodic cues are treated as part of the acoustic signal rather than as a structural factor that recalibrates segmental interpretation.

4.3. Implications for prosody-sensitive speech recognition

Because Whisper predicts lexical tokens from a global encoding of the acoustic signal, its categorization behavior reflects integrated acoustic–lexical hypotheses rather than local adjustments to individual segmental cues. As shown above, source-token voicing exerted a strong influence on categorization. Even when VOT was substantially lengthened or shortened through resynthesis, tokens often continued to be recognized in line with their original category. This resistance to local cue manipulation shows that broader acoustic properties of the sentence can outweigh changes to a single phonetic dimension, contributing to the shallow categorization slopes and to the inconsistent boundary effects observed here.

Within this framework, the present findings offer insight into how prosodic information is represented and used in contemporary end-to-end ASR systems. Although Whisper responded to VOT variation, it did not reliably adjust how that cue was interpreted as a function of prosodic boundary context. This suggests that prosodic information may be present in the acoustic representation without being used to condition segmental interpretation. Boundary-related differences influenced the signal, but they did not consistently shift the model's decision pattern in the way observed for human listeners. Instead, prosodic cues appear to be absorbed into a wider set of acoustic properties that jointly shape lexical predictions.

From a modeling perspective, this pattern aligns with Whisper's training objective. Because the system is optimized to predict text tokens from entire utterances, it can achieve good performance by learning stable mappings between global acoustic patterns and lexical outputs. Under this strategy, boundary-conditioned variation in phonetic realization can be accommodated without requiring dynamic, context-dependent adjustment of phonetic category boundaries. Prosodic structure may still contribute to recognition—for example, by aiding segmentation or lexical expectation—but it does not appear to function as a dedicated mechanism for recalibrating segmental cue interpretation.

These findings highlight the value of controlled phonetic probing for understanding ASR behavior. Standard metrics such as word error rate reveal little about how models handle prosodically structured phonetic variability. By adapting methods from human speech perception research, the present study exposes a key difference between human perceptual normalization and end-to-end ASR inference. The comparison points both to the strengths of current data-driven systems and to the limits of their sensitivity to prosodically conditioned phonetic structure.

4.4. Limitations and directions for future research

The present study provides an initial test of how prosodically conditioned phonetic variation is handled by an end-to-end ASR system, but several limitations constrain the scope of the findings. First, the stimuli were based on recordings from a single speaker, which allowed tight acoustic control but limited the range of phonetic and prosodic variability. Future work should examine whether similar patterns emerge with multiple speakers and more diverse speech materials.

Second, the models were evaluated without task-specific fine-tuning. Training or fine-tuning on prosodically structured data may change how segmental cues are weighted and how contextual information is integrated. Comparing model behavior before and after such exposure will help clarify which aspects of prosody-sensitive phonetic interpretation can emerge from data-driven learning alone.

A further limitation concerns the source of the observed anchoring effect of the original token. Although VOT was systematically manipulated, other acoustic properties of the carrier frame—such as overall speech rate, mean F0, intensity profile, spectral tilt, duration patterns, or cues outside the immediate stop release—were not independently controlled. It therefore remains possible that the model's responses were influenced by global frame-level or distributed acoustic information beyond the manipulated segmental cue. Although the present dataset was not designed for independent manipulation of these frame-level cues, future work will incorporate explicit measurement and control of speech rate, mean and range of F0, intensity, duration, and spectral properties in order to determine whether systematic covariation

aligns with the observed categorization patterns. More generally, future research should implement fuller normalization or resynthesis designs, including complete frame-level control or parametric manipulation of prosodic and spectral properties. Such approaches would allow a more precise diagnosis of whether the anchoring effect reflects reliance on distributed acoustic correlates of voicing, broader prosodic structure, or token-specific idiosyncrasies.

Finally, the analysis focused on a single model, Whisper. Other ASR architectures may represent and use prosodic and phonetic information differently. In particular, self-supervised speech models such as wav2vec 2.0 (Baevski et al., 2020) and HuBERT (Hsu et al., 2021) provide access to intermediate acoustic representations that are not tied directly to lexical prediction. Probing these models could help determine whether prosodically conditioned phonetic patterns are present in learned representations even when they do not appear in token-level ASR outputs. In future work, this paradigm will be extended to a broader range of speech models, including self-supervised representation models and alternative ASR architectures, with systematic comparisons between pre-trained and fine-tuned versions of each model to assess how exposure to controlled speech data influences sensitivity to prosodically conditioned phonetic variation.

5. Conclusion

This study asked whether a modern end-to-end ASR system uses prosodic boundary information to condition phonetic categorization in a manner comparable to human listeners. By systematically manipulating VOT and prosodic boundary context, we tested whether boundary strength would shift voicing categorization in Whisper in the way previously reported for human perception (Kim & Cho, 2013).

The results show that Whisper is sensitive to VOT, but that its categorization behavior differs fundamentally from human listeners. Prosodic boundary context influenced responses in some conditions, yet often reversed in direction, and strongly mediated by the voicing and place of articulation of the source token. In contrast, the most reliable determinant of the model's responses was the global acoustic profile of the original recording, which frequently anchored categorization outcomes even when VOT was substantially manipulated.

This pattern indicates that, under the present experimental setting, Whisper did not show consistent evidence of using prosodic structure as a stable, predictive framework for interpreting segmental cues. Instead, voicing decisions reflect weighted integration of multiple acoustic properties, with global sentence characteristics often outweighing local cue manipulations. Human listeners, by contrast, appear to use prosodic structure to form expectations about how phonetic cues should be interpreted in context, leading to systematic boundary-conditioned shifts in categorization.

These findings highlight a central difference between human perceptual normalization and contemporary ASR inference. Prosodic information may be encoded in the acoustic signal and internal model representations, but it is not necessarily deployed to guide context-dependent phonetic interpretation. More broadly, the study demonstrates the value of controlled phonetic experiments for probing how ASR systems process structured variability in speech, offering insights that complement standard evaluation metrics such as word error rate. Future work will extend this paradigm to a broader range of speech models, including self-supervised representation models, and will involve fine-tuning these systems to compare how different architectures encode and utilize prosodically conditioned phonetic variation.

References

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, *33*, 12449-12460.

Bates, D., Maechler, M., Bolker, B., Walker, S. (2015). Package 'lme4'. *convergence*, *12*(1), 2.Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1-48.

Boersma, P., & Weenink, D. (2024). Praat: Doing phonetics by computer (Version 6.4.13) [Computer software]. Retrieved from https://www.praat.org/

Cho, T. (2016). Prosodic boundary strengthening in the phonetics–prosody interface. *Language and Linguistics Compass*, *10*(3), 120-141.

Cho, T., & Keating, P. (2009). Effects of initial position versus prominence in English. *Journal of Phonetics*, *37*(4), 466-485.

Fletcher, J. (2010). The prosody of speech: Timing and rhythm. In W. J. Hardcastle, J. Laver, & F. E. Gibbon (Eds.), *The Handbook of Phonetic Sciences* (2nd ed., pp. 523-602). Hoboken, NJ: Wiley-Blackwell.

Fougeron, C., & Keating, P. A. (1997). Articulatory strengthening at edges of prosodic domains. *The Journal of the Acoustical Society of America*, *101*(6), 3728-3740.

Hsu, W. N., Bolte, B., Tsai, Y. H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 3451-3460.

Kim, S., & Cho, T. (2013). Prosodic boundary information modulates phonetic categorization. *The Journal of the Acoustical Society of America, 134*(1), EL19-EL25.

Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, *20*(3), 384-422.

Lisker, L., & Abramson, A. S. (1970). The voicing dimension: Some experiments in comparative phonetics. In B. Malmberg (Ed.), *Proceedings of the 6th International Congress of Phonetic Sciences* (pp. 563-567).

McQueen, J. M., & Dilley, L. C. (2020). Prosody and spoken-word recognition. In C. Gussenhhoven, & A. Chen (Eds.), *The Oxford Handbook of Language Prosody* (pp. 509-521). Oxford, UK: Oxford University Press.

Millet, J., & Dunbar, E. (2022). Do self-supervised speech models develop human-like perception biases? arXiv. *https://doi.org/* 10.48550/arXiv.2205.15819

Millet, J., Jurov, N., & Dunbar, E. (2019). Comparing unsupervised speech learning directly to human performance in speech perception. *Proceedings of the Annual Meeting of the Cognitive Science Society, 41*, 2358-2364.

Mitterer, H., Cho, T., & Kim, S. (2016). How does prosody influence speech categorization? *Journal of Phonetics, 54*, 68-79.

Mohamed, A., Lee, H. Y., Borgholt, L., Havtorn, J. D., Edin, J., Igel, C., Kirchhoff, K., ... Watanabe, S. (2022). Self-supervised speech representation learning: A review. *IEEE Journal of Selected Topics in Signal Processing*, *16*(6), 1179-1210.

R Core Team. (2025). *R: A language and environment for statistical computing* (Version 4.5.2) [Computer software]. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org/

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Proceedings of the 40th International Conference on Machine Learning* (Vol. 202, pp. 28492-28518). PMLR.

Steffman, J., Kim, S., Cho, T., & Jun, S. A. (2022). Prosodic phrasing mediates listeners' perception of temporal cues: Evidence from the Korean accentual phrase. *Journal of Phonetics*, *94*, 101156.

- **Jiyoung Jang**
  Post-doctoral Researcher, Hanyang Institute for Phonetics and Cognitive Sciences of Language, Hanyang University
  222 Wangsimni-ro, Seongdong-gu
  Seoul 04763, Korea
  Tel: +82-2-2220-2507
  Email: jiyoungljang@hanyang.ac.kr
  Areas of interest: Phonetics, Laboratory Phonology, Prosody

- **Richard Hatcher,** Corresponding author
  Research Assistant Professor, Hanyang Institute for Phonetics and Cognitive Sciences of Language, Hanyang University
  222 Wangsimni-ro, Seongdong-gu
  Seoul 04763, Korea
  Tel: +82-2-2220-2507
  Email: richard.j.hatcher.jr@gmail.com
  Areas of interest: Phonetics, Phonology, Prosody, Korean

Appendix

Table A1. Examples of decoded outputs mapped to voicing categories for selected minimal pairs

Target pair	Mapped as voiceless	Mapped as voiced
pan/ban	pan, pat, pain	ban, band, better, bender
tan/Dan	tan, ten, 10, tender, time	Dan, down, dandelion
cam/GAM	cam, KM, camp, can	GAM, game, gambler

The table illustrates representative Whisper outputs that were assigned to voiceless and voiced categories based on initial consonant identity after normalization.