JASA ARTICLE

The role of voiced consonant duration in sung vowel-consonant and consonant-vowel recognition[a)]

Allan Vurma,[1,b)] Einar Meister,[2] Lya Meister,[2] Jaan Ross,[1] Marju Raju,[1] Veeda Kala,[1] and Tuuri Dede[1]

[1]*Estonian Academy of Music and Theatre, Tallinn 10116, Estonia*

[2]*Tallinn University of Technology, Tallinn 19086, Estonia*

ABSTRACT:

Sung text intelligibility is often a problem, especially in reverberant acoustics, at high pitch, and in the presence of a loud accompaniment. This study aims to discover whether elongating the duration of voiced consonants /m/, /n/, /l/, and /v/ in sung vowel-consonant (VC) and consonant-vowel (CV) sequences improves their recognition. Perception tests were conducted with 42 participants, and the data were analyzed using generalized linear mixed models. Results showed that consonant durations of 20–35 ms are sufficient for a near-perfect recognition when singing at close-to-speaking pitch in non-reverberant acoustics and without accompaniment. However, in reverberant acoustics for VC sequences, longer consonant durations allow the reverberation from the preceding vowel to fade more fully during the consonant, reducing masking by the room reverberation and thereby improving recognition. Also, elongating consonant duration up to about 200 ms (or even longer) improved consonant recognition in the case of stimuli with added *Brown Noise*, imitating orchestral accompaniment, whereas only negligible impact on the order of succession (CV versus VC) was observed. Recognition tended to be poorer at higher pitch and with longer reverberation.

 https://doi.org/10.1121/10.0039581

(Received 20 May 2025; revised 12 September 2025; accepted 28 September 2025; published online 20 October 2025)

[Editor: Zhaoyan Zhang] Pages: 3120–3132

I. INTRODUCTION

A common issue in singing is the poor intelligibility of the sung text, especially at high pitches, in reverberant acoustics and when accompanied by instruments (Eberhart, 1962; Gregg, 1991; Nelson and Tiffany, 1968; Phillips, 2002; Titze, 1982; Meyer, 2009). Text intelligibility relies primarily on direct (bottom-up) auditory processing mechanisms, whereas top-down processes also contribute by assigning meaning to sound sequences and enhancing predictability through memory, prior experience, language skills, and contextual and cultural awareness (Behrman, 2023).

Although singers have limited means to enhance listeners' comprehension of sung text through top-down mechanisms, certain articulatory and prosodic strategies related to bottom-up perception may help. This study aims to investigate whether modifying consonant duration in vowel-consonant (VC) and consonant-vowel (CV) sequences can be an effective strategy to improve consonant identification in Western classical singing.

Phoneme duration is a core aspect of prosody and can influence lexical meaning, particularly in quantity languages, where word meaning is dependent on duration of the vowels and consonants (Lehiste, 1970). However, meaning often depends not only on the absolute duration of individual phonemes but also on their relative temporal relationships within words and phrases (Menn *et al.* 2023; Halle, 1967). In singing, musical structure predominantly governs phoneme duration, frequently overriding the temporal patterns typical of speech. Moreover, in reverberant concert halls, even identifying the language being sung can sometimes be difficult. In such cases, if a singer can enhance the intelligibility of individual phonemes—even at the expense of prosodic naturalness—this may still be justified.

In our study, we focus on the identification of the voiced consonants /m/, /n/, /l/, and /v/ as a function of their duration in sequences with the vowel /a/, independent of specific language or semantic content (i.e., relying primarily on bottom-up auditory processing). These phoneme sequences are common in Western classical vocal music across languages. Several other consonants were excluded for specific reasons. The consonant /r/ was excluded because for native Estonian speakers (the majority of participants), it is realized as an alveolar trill that is acoustically and perceptually distinct from other consonants. The semivowels /w/ and /j/ were excluded because the articulatory and perceptual boundaries between semivowels and their corresponding vowels (/i/–/j/, /u/–/w/) are not clearly defined. The sibilants /z/ and /ž/ were excluded as well as they occur only in rare loanwords in Estonian (Trost, 2025). Finally, voiced plosives were excluded because they cannot be sustained for long durations.

[a)]Portions of this work were presented in "The influence of the duration of voiced consonants on their recognition in sung vowel-consonant and consonant-vowel junctions," in 37th Finnish Phonetic Symposium, Turku, Finland, and "The influence of the duration of voiced consonants on their recognition in sung vowel-consonant and consonant-vowel junctions," in 54th Annual Symposium of the Voice Foundation, Philadelphia, PA.

[b)]Email: allan.vurma@eamt.ee

https://doi.org/10.1121/10.0039581

Voice pedagogy literature presents different views on how singers should shape consonant duration. Classical singing often favors a legato style, in which vowels flow seamlessly and consonants are produced quickly and precisely without prolongation (Ware, 1998; Davids and LaTour, 2012; Eberhart, 1962). In contrast, some argue that sufficient consonant duration is essential for text clarity (Sharnova, 1947; Nair, 2021; Appelman, 1986; LaBouff, 2008; Waring, 1945; Shaw and Blocker, 2004). However, empirical evidence is lacking regarding the minimum duration required for consonant identification and whether a threshold exists beyond which duration no longer improves intelligibility. Furthermore, it remains unclear if the required duration depends on acoustic factors such as room reverberation, background noise (e.g., from accompanying instruments or ensemble partners), pitch level, and consonant type.

Most research on phoneme identification has focused on speech, but singing, particularly at high pitches, differs acoustically from speech. Thus, findings from speech research may not fully apply (Sundberg, 1987). In vowel identification, vocal tract (VT) resonance frequencies (formants) shape the spectral envelope, where the first two or three formants play the most significant role (Sundberg, 1987). When singing at a high fundamental frequency (f_o), particularly when f_o exceeds the first formant frequency (F1), vowel intelligibility decreases because the sparsity of harmonics in the voice spectrum makes formant structure acoustically less distinct (Sundberg, 1987). The consonants under study—/m/, /n/, /l/, and /v/—can be sustained a way that is similar to vowels, and their spectral envelopes are likewise shaped by VT resonances.

For all four consonants, the sound source is the periodically opening and closing glottis, with /v/, additionally, involving turbulent airflow at the labiodental constriction (Kent and Read, 2002). Their spectral energy is concentrated in the low F1 range, typically around 250–300 Hz (referred to as murmur for /m/ and /n/). When singing high notes with f_o above F1, the low F1 cannot be present in the consonant's spectrum as a result of the absence of harmonics in that frequency range. Because of their acoustic similarities, nasals /m/ and /n/ as well as /l/ and /v/ may be confused with each other in perception (Kent and Read, 2002).

For /m/, /n/, and /l/, spectral shaping is also influenced by antiformants, which absorb energy in specific frequency regions (Johnson, 2012). Antiformants occurring at or near the frequencies of formants can effectively cancel them out, broadening formant bandwidths and suppressing harmonics in those frequency ranges, thereby lowering the overall intensity of these consonants compared to vowels (Johnson, 2012; Fuchs and Birkholz, 2019).

In the case of nasals, antiformants are generated because of the parallel acoustic path (the oral cavity), which functions as a shunt to the nasal tract (through which the sound exits). This shunting channel is longer for the bilabial /m/ and shorter for the alveolar /n/. Additional passages from the nasal cavity lead to the paranasal sinuses, which also create their own resonances functioning as antiformants. Due to considerable individual anatomical variation, the frequency placement of these antiformants and nasal tract formants varies greatly from person to person. However, for a given speaker, when articulating /m/ or /n/, they stay unchanged (Johnson, 2012). Thus, only the shape of the oral cavity and its influence on the resonator system's transfer function are key factors in distinguishing sustained /m/ and /n/ sounds as the invariant shapes of the nasal cavities and paranasal sinuses are not suitable for this distinction (Kent and Read, 2002). Also, the antiformants created by oral cavity in the cases of /m/ and /n/ may not serve as salient identification cues because of their low energy level, which can easily be masked by noise from various sources (Johnson, 2012). Instead, visual cues may play a more significant role in distinguishing between /m/ and /n/ (Fuchs and Birkholz, 2019).

For lateral /l/, antiformants arise from the cavity behind the tongue–palate contact (Johnson, 2012). As with nasals, interactions between formants and antiformants result in considerable individual variability in the spectral transfer function of /l/ compared to vowels (Fuchs and Birkholz, 2019). The voiced fricative /v/ has a more diffuse spectral shape and lower intensity, resulting from the labiodental constriction and energy loss caused by friction (Kent and Read, 2002; Koffi, 2020).

In all syllables examined, sound identification may be influenced not only by the consonant's stationary part but also by VT formant transitions, occurring as a result of changes in the VT shape as it moves from one sound to another (Behrman, 2023). Measuring these transitions through acoustic analysis can be challenging or even impossible because of variability in duration and the precise timing of their onset and offset (Kent and Read, 2002). Different studies have assessed the relative importance of formant transitions versus the stationary portion of a sound in its identification diversely. For example, Kurowski and Blumstein (1984) found that in nasals, nasal murmur and formant transitions contribute about equally as cues to the place of articulation, whereas Liberman *et al.* (1954) argued that formant transitions play the primary role. It can be assumed that for consonants sung with a high f_o, VT formant transitions become less informative as the alignment between formant frequencies and spectral partials becomes more incidental and acoustically less salient.

In the context of sound identification, masking can also play a significant role: a louder sound can reduce the audibility of a quieter sound occurring simultaneously (Howard and Angus, 2006). Masking can be either complete or partial. In partial masking, the masked sound remains audible, but its sound level approaches the increased hearing threshold. In addition to simultaneous masking, forward and backward masking are also possible, where the masker affects sounds that occur either shortly after or before it in time. Backward masking can occur within a time window of up to

about 20 ms, and forward masking can occur within a time window of up to about 200 ms (Meyer, 2009).

In reverberant spaces, the reverberation field itself can act as a masker, arising from multiple sound reflections from room boundaries and objects within the space. The formation of the reverberation field and its decay after the sound source stops require time, which is longer in large spaces with small sound absorption. For short-duration sounds, the reverberation field may not reach a steady-state level, meaning that these sounds may remain quieter and more susceptible to masking (Howard and Angus, 2006).

According to our assumptions, longer consonant duration may also enhance identification accuracy by providing more time for the cognitive processes involved in recognition (Heald and Nusbaum 2014; Koutsogiannaki, 2016). Additionally, sounds shorter than approximately 200 ms may be perceived as quieter (Howard and Angus, 2006), reducing the subjective signal-to-noise ratio. Here, noise can be understood as any simultaneous sounds with longer duration than the target consonant, including background noise, instrumental accompaniment, or sounds produced by ensemble partners. To summarize, our perception tests aim to investigate whether and to what extent the recognition of voiced consonants /l/, /m/, /n/, and /v/ in sung CV and VC sequences depends on consonant duration across various acoustic environments, pitch levels, and noise conditions.

II. METHOD

A. Stimuli

A mezzo-soprano (36 years old) and a baritone (38 years old) with Western classical training were asked to sing a series of phoneme sequences: /ma/, /na/, /la/, /va/ and /am/, /an/, /al/, /av/ at different pitches using a comfortable mezzo-forte/forte dynamic. The selected pitches were G3, A♭4, and F5 for the mezzo-soprano, and B2, B3, and E♭4 for the baritone. Although voice category of singers was not critical to the study, we limited the number of singers to one female and one male to keep the number of perception test stimuli manageable. Both singers perform nationally in principal operatic roles and oratorios (taxonomy categories 3.1a and 3.4). The mezzo-soprano also performs internationally in concerts and oratorios (category 2.4; see Bunch and Chapman, 2000). In general, the characteristics of singers' voices vary smoothly within wide limits, and the boundaries between voice categories are conditional. The selected singers represent the mid-range of this continuum rather than its extreme limits. Including a larger number of singers from all typical voice categories would have been beyond our capabilities.

The difference in pitches used for the mezzo-soprano and baritone reflects their respective vocal ranges; female voices generally have a wider *zona di passaggio* and shorter chest register than male voices (Miller, 1986). The pitch G3 corresponds to the mezzo-soprano's speaking voice range, whereas B2 is similarly characteristic of the baritone's speech, in which both use the chest register. A♭4 lies within the mezzo-soprano's comfortable middle register, whereas F5 falls into the *secondo passaggio* region, where phonation becomes technically more challenging and text intelligibility often decreases. However, this pitch remains roughly a fourth below the mezzo-soprano's upper pitch limit. For the baritone, B3 lies in the *primo passaggio* region, where phonation starts becoming vocally demanding. E♭4 is relatively high for the baritone, falling into the head voice range about a fourth below its upper boundary (Miller, 1986; Ware, 1998).

In sung sequences, consonant durations never exceeded approximately one vibrato period (where the maximum consonant length is 200 ms). As the perception of vibrato requires at least several periods of a given sound, no perceptible vibrato was present in the consonant part of the stimuli. For vowels, the mezzo-soprano's vibrato frequency was typically 6.15 Hz with an amplitude of up to 0.7 semitones. For the baritone, the vibrato frequency was 5.3 Hz with an amplitude of up to 0.9 semitones. Typically, the vibrato rate of opera singers falls within the range of 5–7 Hz, with an amplitude of up to two semitones (Sundberg, 1987).

In addition to the sung sequences, both singers also produced spoken versions of the same stimuli, with naturally varying pitch typical of speech (hereafter referred to as Sp_bar and Sp_mez for the baritone and mezzo-soprano, respectively).

The recordings were conducted in a low-reverberation studio (reverberation time T30 = 0.2 s) using a DPA SC4061-FM omnidirectional microphone (DPA Microphones, Kokkedal, Denmark) placed off-axis 3.5 cm from the corner of the singer's mouth, an Audient ID4 audio interface (Audient, Hampshire, UK), and a Dell Latitude 5400 laptop (with sampling rate 44 100 Hz and bit depth of 16 bits; Dell Technologies, Round Rock, TX). Recording levels were calibrated with a sound pressure level (SPL) meter Limit 7000 (Luna AB, Alingsas, Sweden) with linear (dBC) frequency response correction, positioned 30 cm from the singer's mouth as instructed by Svec and Granqvist (2018).

Using PRAAT software (Boersma and Weenink, 2024), first, we created a set of *clear* stimuli by modifying the phoneme sequences to have consonant durations of 0, 20, 35, 50, 75, 100, 150, and 200 ms. Consonants were either shortened or extended by duplicating original segments (for the method, see Charpentier and Stella, 1986). Vowel durations were standardized to 0.5 s across all stimuli to prevent vowel duration from serving as a cue for recognition under different acoustic conditions. A 20 ms onset (for VC stimuli) and offset (for CV stimuli) smoothing was applied to preserve the naturalness of the sounds.

To ensure better ecological validity, intensity differences across pitches and consonants were not artificially equalized. In singing, vowel intensity typically increases with pitch (Sundberg, 1987). Additionally, vocal intensity tends to be higher for the oral consonant /l/, where the VT remains relatively open, and lower for the fricative /v/, where the airflow is more obstructed (Koffi, 2020).

The prevocalic lateral /l/ differs somewhat from its postvocalic counterpart by exhibiting a lower F1 and higher

https://doi.org/10.1121/10.0039581

F2 (Kent and Read, 2002), resulting in some covariate differences in our perception test stimuli between VC and CV contexts. Nasal consonants generally have lower intensity than oral sounds (vowels) as the nasal cavity is less efficient at radiating sound energy because of the small size of the nasal openings (Johnson, 2012).

For both singers, consonants as well as vowels sung at mid-range pitch were more intense than those sung at lower pitches. At high pitches (F5 for the mezzo-soprano and E♭4 for the baritone), consonant intensity tended to decrease slightly, whereas the intensity of vowels increased further. This effect can be attributed to the fundamental frequency rising above the consonant's first formant (F1). As the strongest spectral component largely determines the overall intensity of the sound (Sundberg, 1987) and, for our consonants, this component lies near F1, its influence weakens or disappears when f_o surpasses F1.

At different pitches, vowel intensity measured at 30 cm from the singer's mouth reached up to 105 dBC for the mezzo-soprano at F5 and the baritone at E♭4. Spoken stimuli had an average intensity of approximately 70 dBC, whereas sung stimuli at speech-like pitch reached around 80 dBC. Consonant intensity was generally a few to 10 dB lower than the adjacent vowel, but this difference reached up to 30 dB in stimuli sung by the mezzo-soprano at F5. A similar effect—where vowels sung at high pitches become significantly more intense than adjacent consonants compared to speech—was also reported by Vurma *et al.* (2023).

For stimuli where pitch, consonant type, and phoneme order (VC or CV) were held constant and only consonant duration varied, phoneme intensity remained consistent across the entire series. As a result, within each series, consonant identification depended solely on consonant duration. However, because stimuli with different pitches exhibited some variation in phoneme intensity, which could also influence consonant identification, pitch as a factor affecting phoneme identification is somewhat conditional in this study. It encompasses not only pitch itself but also covarying properties such as phoneme intensity and, possibly, timbre. This study does not attempt to quantify the contribution of these covarying factors as they do not interfere with the primary goal of investigating the influence of phoneme duration. In practical singing, such covariation between pitch, intensity, and timbre is typical.

To simulate different room acoustics (a concert hall and a church) and the presence of accompanying instruments, additional stimulus series were created from the *clear* stimuli by adding artificial reverberation using the Praat Vocal Toolkit plugin (Corretge, 2012–2025). The presets *church* (*Ch*) 70% and *big room* (*BR*) 70% were used, where the proportion of reverberant sound energy was 70%. For the *Ch* condition, reverberation time was estimated at approximately 5 s, representing the acoustics of a medium-sized church, whereas for the *BR* condition, it was estimated at 1.3 s, which is typical of a medium-sized concert hall (Meyer, 2009).

To simulate masking by orchestral accompaniment, brown noise (BN) was added using the *BN* 80% preset. The long-term average spectrum (LTAS) of BN decreases at −6 dB per octave, which is sufficiently similar for the purpose of our study, to the typical LTAS slope of symphony orchestras, −9 dB per octave (Lindblom and Sundberg, 2007).

In summary, the complete stimulus paradigm comprised 2 singers (mezzo-soprano and baritone) × 2 sequence types (CV and VC) × 8 consonant durations (0, 20, 35, 50, 75, 100, 150, and 200 ms) × 4 consonants (/m/, /n/, /l/, and /v/) × 4 pitches per singer (Sp_mez, G3, A♭4, and F5 for the mezzo-soprano; Sp_bar, B2, B3, and E♭4 for the baritone) × 4 acoustic conditions (*clear*, *BN*, *BR*, and *Ch*), resulting in a total of 2048 stimuli.

B. Participants and procedure

A total of 42 participants (13 men and 29 women; aged 16–69 years old) were recruited via personal contacts within the research group and through social media. Most participants were native Estonian speakers, where one was a native Russian speaker; only some of them had a musical background.

Each participant completed the perception tests individually using their personal computers and headphones. They were instructed to adjust their computer volumes to a comfortable level before starting, ensuring that the volume remained unchanged throughout the session.

The tests were conducted using PRAAT and divided into four sections: CV stimuli by the baritone, VC stimuli by the baritone, CV stimuli by the mezzo-soprano, and VC stimuli by the mezzo-soprano. Stimuli were presented in random order, where each one was presented once. Participants responded by selecting one of four response buttons labeled with the corresponding CV or VC sequences; a question mark option was provided for cases where the consonant was completely unrecognizable.

Breaks of self-determined duration were automatically prompted after every 32 stimuli, corresponding to approximately 2–3 min depending on participant's response times. Longer breaks were provided between test sections. The full session, including breaks, lasted approximately 1.5 h, although test series could be completed over one or multiple days. Allowing participants to take breaks at their own pace helped mitigate potential fatigue effects. On completion, participants saved their results and submitted them to the researchers. They received a €20 bookshop gift card as compensation for their time.

C. Data analysis

The results were analyzed using several generalized linear mixed models (GLMMs), fitted with the *lme4* package (Bates *et al.*, 2015) in *R* version 4.3.2 (R Core Team, 2021) to investigate the effects of different factors. Tukey *post hoc* pairwise comparisons were conducted using the *emmeans* package (Lenth, 2024) to assess differences between all

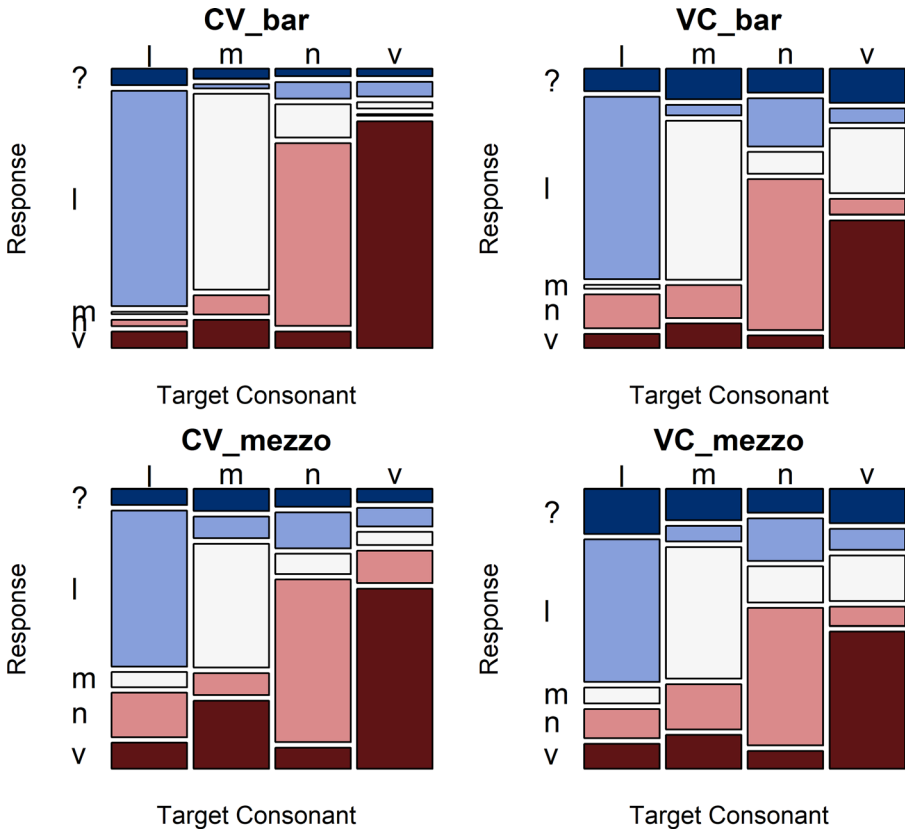

FIG. 1. Mosaic plots illustrating the distribution of responses /l/, /m/, /n/, /v/, and "?" to the CV (left) and VC (right) stimuli sung by baritone (top) and mezzo-soprano (bottom).

levels of the fixed factors. All models included a binary response variable (1 = correct, 0 = incorrect) and *duration* (eight levels) as a fixed factor, whereas additional fixed and random factors varied across models. Each model is described in Sec. III (Results).

III. RESULTS

A. Distribution of responses

Figure 1 illustrates the overall distribution of responses. Only 8.6% of all responses were categorized as "?,"indicating that participants were generally successful in identifying the consonants. However, the proportion of "?" responses was higher for VC stimuli (11.9%) than for CV stimuli (5.3%) and higher for stimuli sung by a mezzo-soprano (9.9%) compared to those sung by a baritone (7.4%).

The recognition rate was higher for CV stimuli (70.2%) than for VC stimuli (56.8%) and higher for baritone stimuli (70.1%) compared to mezzo-soprano stimuli (56.9%). Among consonants, overall recognition rates ranged as follows: /l/, 55.4%–83.8%; /m/, 48.1%–76.2%; /n/, 53.4%–71.1%; and /v/, 49.8%–88.3%.

B. The impact of consonant duration on its recognition—A general trend

A GLMM was fitted to the pooled data with consonant duration as the only fixed factor for durations of 0, 20, 35, 50, 75, 100, 150, and 200 ms, using 200 ms as the baseline. Pitch, consonant, acoustic condition, singer, and sequence type were included as random effects. As depicted in Fig. 2, the proportion of correct responses predicted by the model increased consistently with longer consonant durations. Analysis of variance (ANOVA) for the GLMM indicates that the effect of duration was statistically significant ($\chi^2 = 3568.31$, degree of freedom, df = 7, $p < 0.001$). When consonant duration increased from 0 to 200 ms, predicted average recognition improved by 35percentage points (from 41% to 76%).

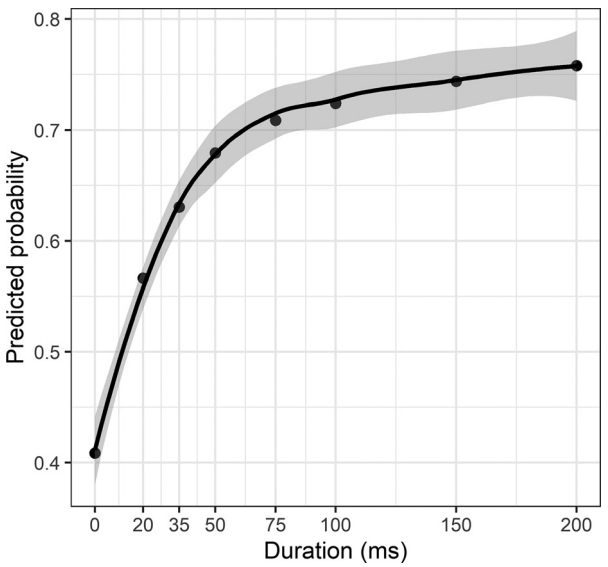

FIG. 2. The proportion of correct responses predicted by the GLMM as the function of consonant duration, where all data are pooled (with 95% confidence intervals).

https://doi.org/10.1121/10.0039581

TABLE I. The results of the ANOVA for the GLMM. In the model, the duration, singer, acoustic condition, and their interactions are included as main factors.

Response: Correct			
	χ^2	df	p
(Intercept)	27.19	1	<0.001
Duration	1072.56	7	<0.001
Singer	2.68	1	0.1
Acoustics	181.54	3	<0.001
Duration:singer	39.27	7	<0.001
Duration:acoustics	220.28	21	<0.001
Singer:acoustics	107.07	3	<0.001
Duration:singer:acoustics	75.04	21	<0.001

Considering only durations between 20 and 200 ms, the increase was 20 percentage points (from 56% to 76%).

The improvement was steepest during the first ∼50 ms and became progressively slower at longer durations (0.8% per 1 ms between 0 and 20 ms; 0.43% per 1 ms between 20 and 35 ms; but only 0.02% per 1 ms between 100 and 200 ms). Notably, even when the stationary part of the consonants was removed—leaving only information from VT formant transitions—the predicted average recognition (41%) was still above chance.

C. The impact of consonant duration by acoustic conditions

In the next model, consonant duration, acoustic condition (*BR, Ch, BN*, with *clear* as baseline), singer (mezzo-soprano, with baritone as baseline), and their interactions were included as fixed effects, whereas pitch, consonant, and sequence type were included as random effects. The effects of the duration, acoustic condition, and their interactions were statistically significant at $p < 0.001$, whereas the effect of singer was significant only in interactions (see Table I for the GLMM's ANOVA results, and Fig. 3). On average, recognition was about 15 percentage points higher for stimuli sung by the baritone compared to that for the mezzo-soprano. Recognition declined with increasing reverberation: *clear* > *BR* > *Ch*. The lowest predicted recognition was observed for mezzo-soprano stimuli with *BN* simulating orchestral accompaniment. In this condition, recognition improved by about 10 percentage points when consonant duration increased from 75 to 200 ms. In contrast, there was no significant improvement for mezzo-soprano stimuli in the *clear* acoustic condition between 75 and 200 ms.

D. The impact of consonant duration at different pitch levels

In the next GLMM, we examined pitch effects alongside consonant duration. Consonant duration and pitch (B2, B3, E♭4, Sp_mez, G3, A♭4, F5, with Sp_bar as the baseline) were fixed effects, with acoustic condition, sequence type, and consonant as random effects. Main effects and their interactions were significant ($p < 0.001$; see Table II and Fig. 4). Recognition improved with longer consonant duration across all pitch levels, including the spoken stimuli. Recognition was poorer at higher pitches, where performance for the highest pitch F5 approaches or falls below chance level (25%), depending on consonant duration. For

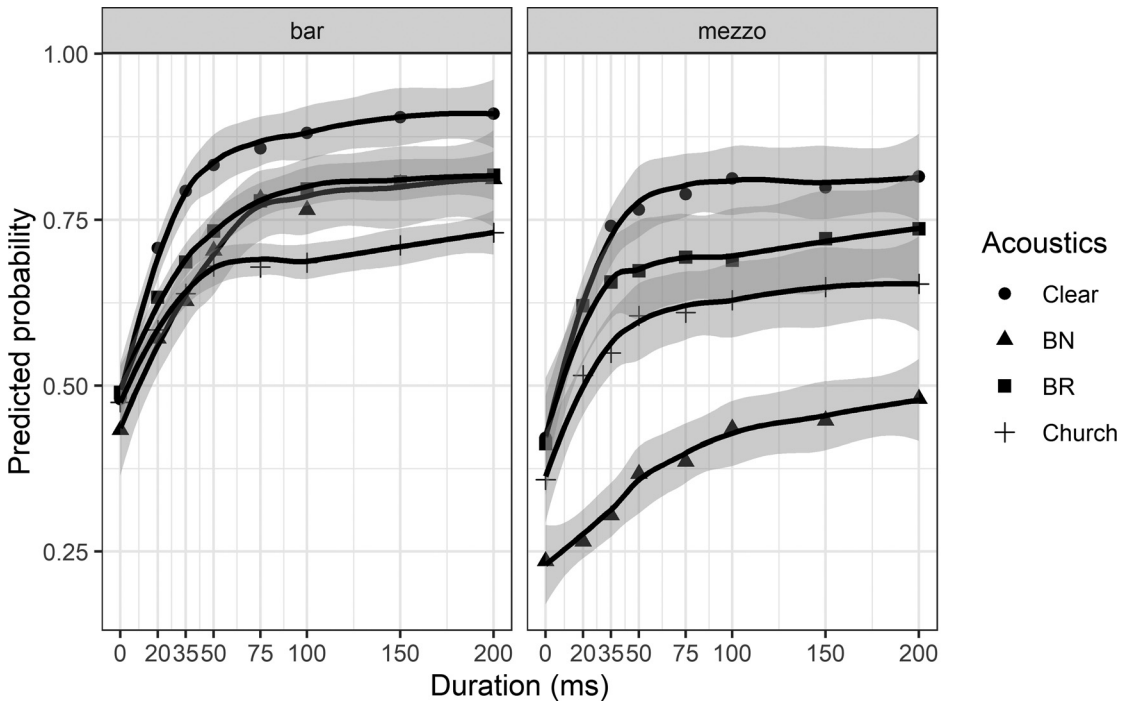

FIG. 3. The probability of correct responses by acoustic conditions predicted by the GLMM as the function of consonant duration. Left panel is for the baritone and right panel is for the mezzo-soprano stimuli (with 95% confidence intervals).

TABLE II. The results of ANOVA for the GLMM. The effects of duration, pitch, and their interactions were applied as main factors.

Response: Correct			
	χ^2	df	p
(Intercept)	30.53	1	<0.001
Duration	408.22	7	<0.001
Pitch	1541.66	7	<0.001
Duration:pitch	310.59	49	<0.001

spoken stimuli, recognition was similar to that in sung stimuli at pitches close to the speaking pitch (B2 for the baritone and G3 for the mezzo-soprano).

As consonant duration increased from 20 to 200 ms, recognition improved by 30 percentage points (from 46% to 76%) for E♭4 stimuli and 26 percentage points (from 53% to 79%) for B3 stimuli. However, when duration increased from 50 to 200 ms, recognition improved by less than 5 percentage points for spoken stimuli and sung stimuli at pitches close to speaking pitch. In contrast, at the highest pitch levels F5 and A♭4, recognition improved more slowly between 0 and 50 ms compared to that for the stimuli at lower pitch levels.

E. The influence of CV versus VC succession order (type) in different acoustic conditions

To examine whether consonant recognition is influenced by the order of vowels and consonants (VC versus CV), we fitted a GLMM with fixed effects for duration, acoustic condition, sequence type (VC, with CV as the baseline), and their interactions. Singer and consonant were included as random effects. ANOVA results revealed that all main effects and their interactions were statistically significant ($p < 0.001$; see Table III and Fig. 5).

In the left panel of Fig. 5, we observe that for CV sequences, the predicted consonant recognition for the stimuli with consonant duration shorter than 100 ms remained almost the same across the *clear*, *BR*, and *Ch* acoustic conditions. At longer durations (100 and 200 ms), differences in recognition probabilities across these acoustic conditions were minimal (up to about 5 percentage points), where recognition of *clear* stimuli tends to be slightly higher. According to *post hoc* test results, this difference was statistically significant only at 200 ms consonant durations between *clear* and *BR* acoustic conditions ($z = 5.6$, $p < 0.001$). An exception was found for stimuli with added *BN*, where recognition was about 20 percentage points lower than in the other acoustic conditions.

In contrast, recognition of VC stimuli was strongly affected by the acoustic condition, declining with increasing reverberation (see Fig. 5, right panel). For example, at 100 ms consonant duration, recognition was about 15 percentage points lower in *BR* acoustics and 30 percentage points worse in *Ch* acoustics compared to *clear* stimuli; both differences are statistically significant ($z = 9.4$, $p < 0.001$ and $z = 16.39$, $p < 0.001$, respectively). Recognition probabilities for VC stimuli with *BN* were similar to those of CV stimuli under the same conditions.

F. Interactions with the consonant duration, acoustic condition, succession type, and pitch

Figures 6 (baritone stimuli) and 7 (mezzo-soprano stimuli) illustrate interactions between consonant duration, pitch level, acoustic condition, and sequence type. Separate GLMMs were applied for each of the four acoustic conditions, with consonant duration, sequence type, pitch, and their interactions as fixed effects and consonant and participant as random effects. Tables IV and V show the statistical significance of the corresponding factors and their interactions for baritone and mezzo-soprano stimuli, respectively.

For spoken CV sequence, predicted recognition was nearly ceiling (~100%) at the shortest consonant duration (20 ms) and remained almost stable across longer durations, except in the *BN* condition. Similarly, near-ceiling recognition

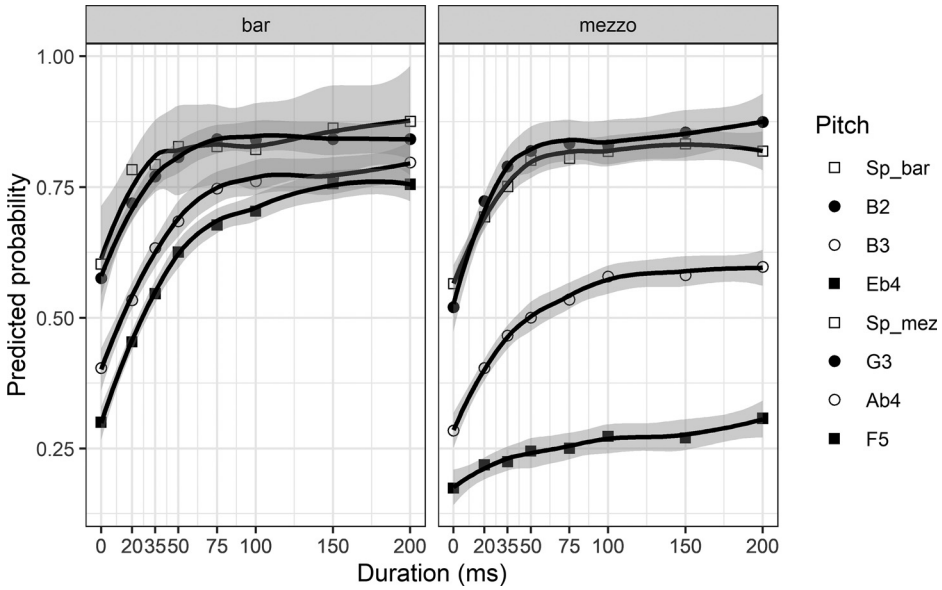

FIG. 4. The probability of correct responses predicted by the GLMM as a function of the duration of consonants by pitch levels. Left panel is for the baritone and right panel is for the mezzo-soprano stimuli (with 95% confidence intervals).

https://doi.org/10.1121/10.0039581

TABLE III. The results of ANOVA for the GLMM. Duration, acoustic condition, type, and their interactions are applied as main factors in the model.

Response: Correct			
	χ^2	df	p
Intercept	52.02	1	<0.001
Duration	531.62	7	<0.001
Acoustics	267.71	3	<0.001
Type	13.61	1	<0.001
Duration:acoustics	99.19	21	<0.001
Duration:type	52.92	7	<0.001
Acoustics:type	225.35	3	<0.001
Duration:acoustics:type	192.32	21	<0.001

was observed for sung stimuli at pitches close to the speaking pitch, except for baritone stimuli in reverberant acoustics *BR* and *Ch*, where recognition was up to about 10 percentage points below maximum. For VC sequence sung by the mezzo-soprano in the *clear* condition, recognition for spoken and sung stimuli near the speaking pitch approached ceiling within the first 50 ms of consonant duration. In contrast, recognition remained notably lower at all higher pitch levels across all acoustic conditions. In particular, for the highest pitch (F5), it remained close to chance regardless of consonant duration, and was especially low in the *BR* and *Ch* conditions (see F5 curves in Fig. 7). Additionally, under reverberant acoustics, recognition of VC sequences became more sensitive to consonant duration at longer durations. Minor fluctuations observed in some curves across Figs. 6 and 7 were not statistically significant.

G. Predicted recognition by consonants

Figure 8 shows that the GLMM-predicted probability of correct responses increased with longer consonant duration for all four consonants. In this model, consonant duration and consonant identity were included as fixed effects, whereas sequence type and acoustic condition were treated as random effects. ANOVA results indicated significant effects of duration ($\chi^2 = 1358.66$, df = 7, $p < 0.001$), consonant ($\chi^2 = 156.49$, df = 3, $p < 0.001$), and their interactions ($\chi^2 = 570.71$, df = 21. $p < 0.001$). The largest improvement in recognition occurred at consonant durations of up to 75–100 ms. An exception was /v/, for which recognition plateaued within the first 50 ms with no further improvement at longer durations.

IV. DISCUSSION

The results of our study showed that elongating voiced consonants in sung VC and CV sequences tend to improve their recognition for durations up to at least 200 ms. However, additional factors, such as room acoustics, consonant position before or after the vowel in the sequence, and pitch, may influence the effect. Moreover, recognition can also exceed chance even when the stationary part of the consonant is absent; in such cases, VT formant transitions provide the primary cue.

In favorable conditions—CV sequence order, speaking or singing at pitches close to speaking pitch, and the absence of accompaniment or noise— only 20–35 ms of a stationary consonant segment was sufficient for near 100% recognition and further elongation of the consonant yielded little or no additional improvement.

The greatest benefit of longer consonant durations arises in less favorable acoustic conditions involving masking and high pitch (the exception is the highest female pitch range). A singer's voice may be masked either by sounds from accompanying instruments, ensemble partners or surrounding noise, and in reverberant rooms by reflections of the singer's own voice. Our results support the hypothesis that in VC stimuli, longer consonant durations allow more time for the reverberation field of the preceding vowel to decay, thereby reducing masking during the latter portion of the consonant and improving recognition. In contrast, in CV

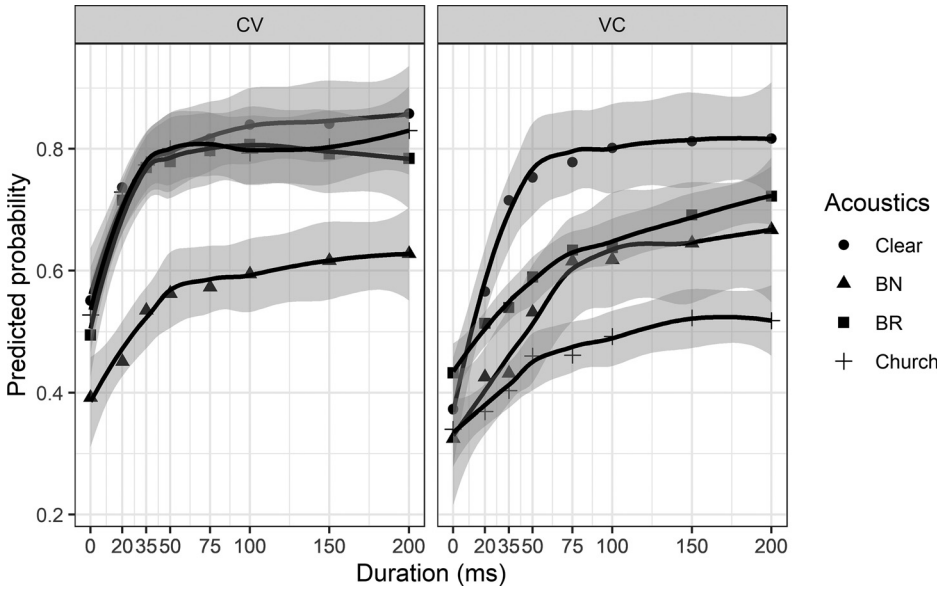

FIG. 5. The proportion of correct responses predicted by the GLMM as a function of the consonant duration split by four acoustic conditions. Left panel is for the CV and right panel is for VC stimuli (with 95% confidence intervals).

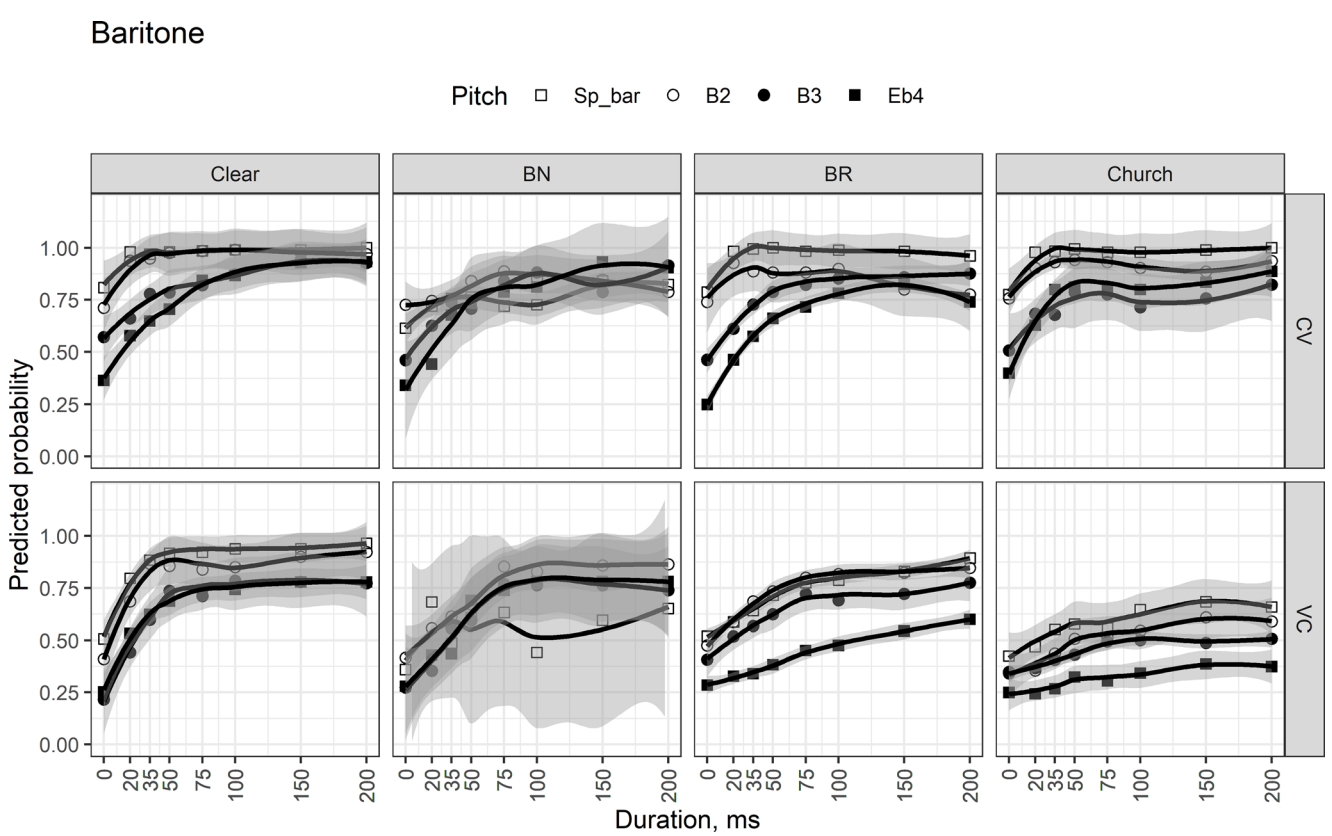

FIG. 6. The proportion of correct responses predicted by the GLMM in the case of stimuli with baritone voice split by acoustic conditions, pitch levels, and succession type (with 95% confidence intervals).

sequences, where the consonant precedes the vowel, simultaneous masking of the consonant by the vowel's reverberation tail is not possible. Consequently, consonant recognition in CV stimuli tended to be higher than that in VC stimuli. Although backward masking remains theoretically possible in CV sequences, we estimate its influence to be minor, likely limited to the shortest consonant durations. For example, we typically observed saturation in recognition already at 20–35 ms in CV sequences with low and medium pitches in *BR* and *Ch* acoustics, indicating no significant backward masking. In contrast, for VC sequences, where masking of consonants by the reverberation tail of the vowel is possible, recognition improved steadily with increasing duration and did not reach saturation even at 200 ms (the maximum duration tested; e.g., see the *BR* and *Ch* panels in Fig. 6).

In VC sequences, forward masking of the consonant by the preceding vowel was also possible. In *clear* acoustics, this may explain why recognition of consonants in VC sequences at 20 ms and 35 ms was somewhat poorer than for that in CV sequences as the influence of forward masking typically lasts considerably longer than that of backward masking (Meyer, 2009). *Post hoc* test confirmed that the difference between VC and CV sequences was statistically significant at 20 ms ($z = 9.42$, $p < 0.001$).

TABLE IV. The results of four ANOVAs for GLMMs by acoustic conditions in the case of baritone stimuli.

Response: Correct									
	Clear		BN		BR		Ch		
	χ^2	p	χ^2	p	χ^2	p	χ^2	p	df
(Intercept)	0.0032	0.96	31.06	<0.001	60.94	<0.001	3.83	0.05	1
Duration	75	<0.001	29	<0.001	78.81	<0.001	95.08	<0.001	7
Type	0.002	0.96	12.29	<0.001	5.61	0.018	3.52	0.06	1
Pitch	3.57	0.31	15.77	0.0013	31.87	<0.001	13.93	0.003	3
Duration:type	3.85	0.8	18.92	0.008	33.08	<0.001	47.23	<0.001	7
Duration:pitch	48	<0.001	128.71	<0.001	114.21	<0.001	55.14	<0.001	21
Type:pitch	0.42	0.94	23.22	<0.001	13.9	0.003	11.34	0.01	3
Duration:type:pitch	36.2	0.02	73.56	<0.001	90.23	<0.001	40.7	0.006	21

https://doi.org/10.1121/10.0039581

TABLE V. The results of four ANOVAs for GLMMs by acoustic conditions in the case of mezzo-soprano stimuli.

Response: Correct									
	Clear		BN		BR		Ch		
	χ^2	p	χ^2	p	χ^2	p	χ^2	p	df
(Intercept)	55.86	<0.001	12.86	<0.001	58.18	<0.001	42.47	<0.001	1
Duration	73.57	<0.001	24.56	<0.001	87.21	<0.001	111.53	<0.001	7
Type	1.12	0.29	12	<0.001	3.46	0.063	23	<0.001	1
Pitch	128.55	<0.001	74.2	<0.001	153.87	<0.001	115.07	<0.001	3
Duration:type	30.78	<0.001	21.92	0.0026	14.23	0.047	35.59	<0.001	7
Duration:pitch	74.93	<0.001	44.42	0.0021	111.83	<0.001	132.82	<0.001	21
Type:pitch	7.48	0.058	7.83	0.05	9.46	0.024	28.73	<0.001	3
Duration:type:pitch	32.84	0.048	46.23	0.0012	44.34	0.002	78.46	<0.001	21

The benefit from elongating consonant duration was most pronounced at medium pitches. For spoken stimuli and stimuli sung at low pitches close to speaking range, recognition typically reached saturation near 100% accuracy already at 20 ms. On the other hand, at high pitches (especially the mezzo-soprano's F5) in reverberant acoustics, recognition remained low and close to chance level even for long consonant durations. In such cases, recognition could not decline further at shorter durations as performance cannot fall below chance.

Typically, when recognition reached high levels only at longer consonant durations, it usually improved quickly by about 25 percentage points or even more at short consonant durations from 0 to 20, 35, or 50 ms. Further elongation of consonants up to 200 ms continued to improve recognition, primarily under reverberant conditions when the consonants were partially or fully masked by the reverberation tail of the preceding vowel in VC sequences. Under favorable acoustic conditions, recognition saturated near 100% at 20 or 35 ms, whereas under less favorable conditions, saturation was not achieved even at 200 ms. Although improvements at longer durations were generally smaller, gains could still reach 30 percentage points between 20 and 200 ms without approaching saturation (e.g., see E♭4 curve in the BR/CV panel of Fig. 6). This pattern was typical for CV stimuli in reverberant acoustics (*BR* and *Ch*). To determine the

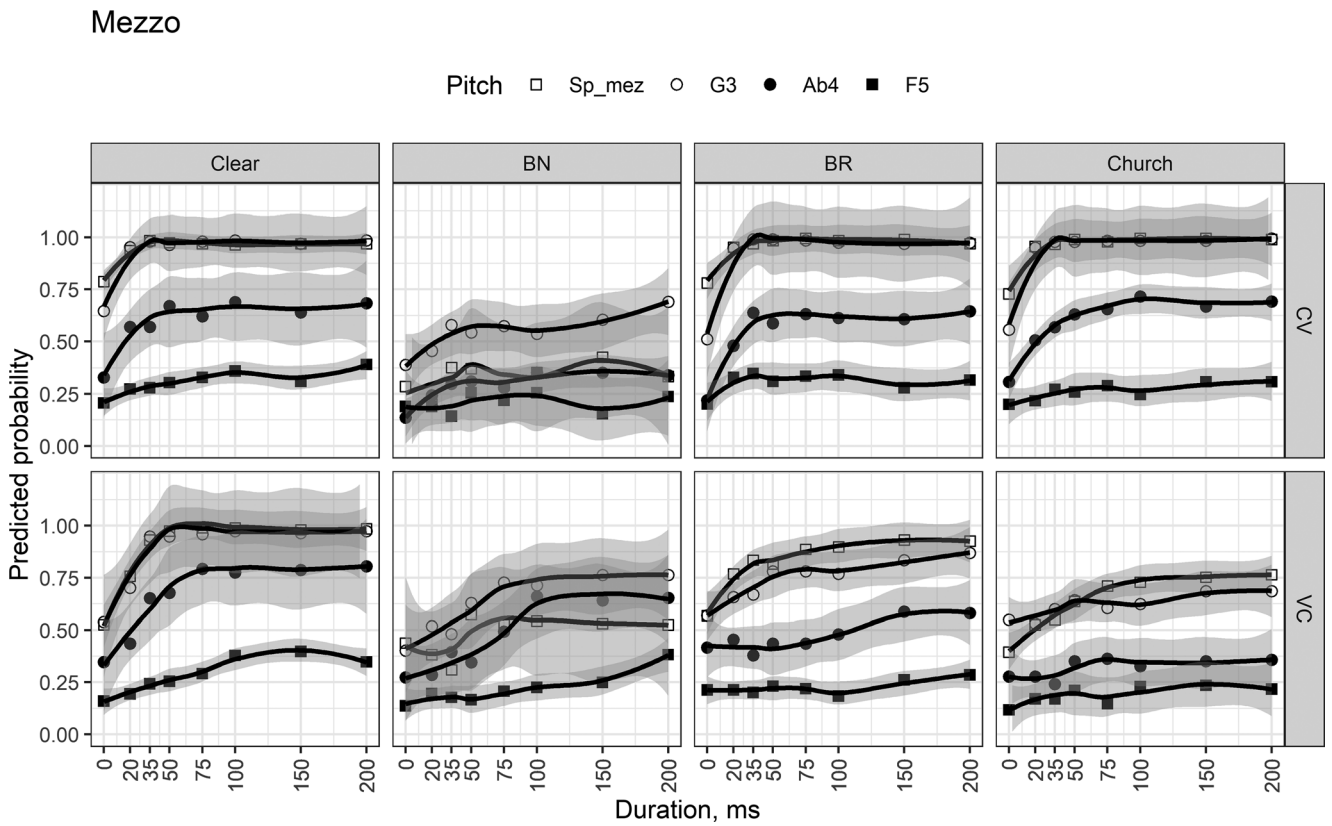

FIG. 7. The proportion of correct responses predicted by the GLMM in the case of stimuli with mezzo-soprano voice split by acoustic conditions, pitch levels, and succession type (with 95% confidence intervals).

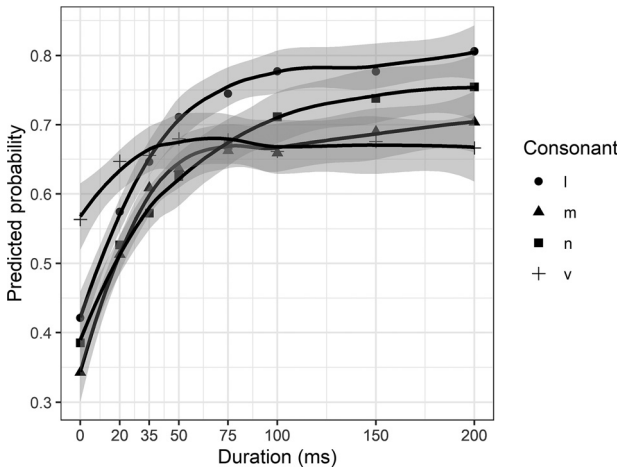

FIG. 8. GLMM-predicted probability of correct responses as a function of consonant duration for /l/, /m/, /n/, and /v/. Data are pooled across all acoustic conditions and pitch levels (with 95% confidence intervals).

consonant duration required for saturation in such cases, further experiments with consonant durations longer than 200 ms would be necessary.

In our study, each listener used their own headphones at a subjectively comfortable sound level, which likely varied across participants and contributed to variability in perceptual test responses (as reflected in the 95% confidence intervals in Figs. 2–8). Ideally, standardized calibrated equipment would have been used, but time and resource constraints prevented this. Variability may also have arisen from individual hearing-system differences as audiograms typically show intersubject standard deviations of up to about 12 dB even in young adults (Corso and Cohen, 1958), and greater variability can be expected in older listeners. The most relevant headphone parameter in this context is frequency response between about 100 Hz (approximately the lowest oscillation frequency of the stimuli) and 5.5 kHz (the upper range essential for vowel perception in female singers; Sundberg, 1987). Typical headphones vary by about 5–15 dB in this range (see frequency response graphs[1]), which is comparable to audiogram variability. Thus, differences in headphone types and individual hearing characteristics likely widened confidence intervals but are unlikely to have altered our conclusions. Moreover, as each participant used the same headphones across test series, headphone-related effects were consistent within listeners.

In our study, headphone variability can be compared to the situation in concert or theatre halls, where the aural information available to listeners depends on their seating position. Although participants used different headphones, each device remained consistent across the session, just as each listener's experience in a hall remains fixed for a given seat.

A positive effect of elongating the consonant was also observed in stimuli with added BN, although this effect was less consistent and showed greater interindividual variability in spoken stimuli compared to those in other acoustic conditions (see, e.g., Figs. 6 and 7). Recognition of spoken *BN* stimuli tended to be poorer than that of sung stimuli at pitches close to typical spoken pitch. Several factors may have contributed to this. First, the masking effect of BN on syllables can vary throughout the syllable as the pitch in spoken stimuli tends to glide slightly—a characteristic feature of natural speech. Lower-pitched sounds are more prone to masking because the spectral energy of BN decreases by approximately 6 dB per octave and is most concentrated in the low frequencies.

Second, individual headphone characteristics may have influenced masking by BN. Headphones with enhanced low-frequency response (sometimes boosting up to 10 dB below 100 Hz) could amplify BN energy in this region, increasing its ability to mask nearby spectral components of low-pitched voices (e.g., baritone stimuli at speaking pitch). By contrast, headphones without such a boost do not produce this effect. Consequently, variability in headphone frequency response likely contributed to the higher variability of perceptual responses in the *BN* condition (see Fig. 6, BN VC panel, spoken stimuli). This effect was largely limited to *BN* stimuli as in all other conditions, masker and maskee shared the same f_o, making headphone effects comparable across both.

Among the four consonants, /v/ showed the smallest effect of increased consonant duration on recognition—it was the highest at 0 and 20 ms but the poorest at longer durations of 150 and 200 ms. This may be related to singers' deviations from proper articulation, especially at high pitches. The narrow labiodental constriction required for /v/ makes producing this fricative with sufficient intensity difficult. At high f_o, an additional drop in the intensity of voiced consonant can occur because of the absence of spectral partials near a low F1 when f_o exceeds F1. At the same time, the intensity of high-pitch vowels adjacent to the consonants tends to increase naturally, e.g., as a result of the formant tuning technique that female singers use at high pitches. In such a situation, to increase the loudness of /v/, it is possible that singers tend to modify /v/ by slightly opening the labiodental constriction toward a semivowel /w/, especially when the target consonant is to be produced with a longer duration. It is possible that this hypothetical articulatory modification was reflected in the acoustic characteristics of the stimuli and could explain why the recognition of /v/ was the poorest among the consonants, as well as why the expected benefit of elongation at longer durations was absent. Additional spectral analysis supports this hypothesis. The /v/ spectrum exhibited clear harmonic partials, similar to vowels, and only a weak, irregular noise component unlike spoken /v/, which typically has a stronger noise component. In some cases, it is likely that attempts to increase the intensity of /v/ may lead singers to produce a firmer labiodental constriction (toward /f/) which, however, may reduce consonant audibility.

In producing the stimuli for the perception tests, we aimed to maintain ecological validity by preserving the natural intensity differences across different pitches, as well as the intensity relationships between vowels and consonants

depending on the consonant and sequence type. Consequently, our results more adequately reflect typical situations in real singing across various conditions. At the same time, though, the numerical results inevitably reflect some idiosyncrasies of the specific performances by the singers used. For example, the A♭4 curves in *clear* acoustics (see Fig. 7) show slightly higher predicted consonant recognition in VC sequences than in CV sequences. This may be attributed to our mezzo-soprano's tendency to produce slightly more intense consonants in VC sequences compared to that in CV sequences. However, in reverberant acoustics, masking of consonants by the vowel's reverberation tail in VC sequences superseded the probable influence of the intensity difference, resulting in significantly lower recognition than in CV sequences (see A♭4 curves in the *BR* and *Ch* panels of Fig. 7).

Finally, one may ask whether elongating consonants affects legato and sung text prosody. Because all four consonants were sonorants with a definite pitch that can be sustained similarly to vowels, elongation is unlikely to disrupt legato. Moreover, perceived prosody is likely more sensitive to relative durations between adjacent speech sounds than to absolute durations, therefore, extending consonant duration may have minimal impact on prosodic perception.

V. CONCLUSIONS

Based on the GLMMs fitted to the perception tests results from 42 participants, we conclude that elongating the duration of the voiced consonants /m/, /n/, /l/, and /v/ can, in certain cases, improve their recognition in sung VC and CV sequences. Elongating appears to reduce masking and allows listeners more time for cognitive processes. In reverberant acoustics, consonants in VC sequences may be masked by the reverberant field of the preceding vowel, as well as by forward masking, additional sounds, and noise. Recognition can also decline at high pitches, where the wide distance between spectral partials reduces access to information about precise locations of formant frequencies.

Vocalists may benefit from knowing that at pitches close to the spoken voice range, CV sequences beginning with a voiced consonant require only a short stationary part—about 20–35 ms—for near-perfect recognition, even in a reverberant acoustic environment. In such cases, recognition remains high (sometimes approaching 80% in our study) even if the stationary part is absent. In non-reverberant acoustics, recognition of consonants in VC sequences also does not deteriorate significantly.

The greatest benefit from elongating consonants beyond about 35 ms occurs in condition with moderate reverberation (*BR* in our study) at medium pitches in VC sequences. Longer consonants also tend to enhance their recognition when singing with accompaniment.

ACKNOWLEDGMENTS

This study was supported by the Estonian Science Foundation (PRG1552). We thank the vocalists and all participants who took part in the perception tests.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Ethics Approval

This research was approved by the Research Ethics Committee of the University of Tartu. Informed consent was obtained from all participants.

DATA AVAILABILITY

The data that support the findings of this study are available on request from the corresponding author.

[1]See https://kuulokenurkka.squig.link/ (Last viewed September 3, 2025).

Appelman, D. R. (**1986**). *The Science of Vocal Pedagogy: Theory and Application* (Indiana University Press, Bloomington, IN), pp. 171–172.
Bates, D., Mächler, M., Bolker, B. M., and Walker, S. C. (**2015**). "Fitting linear mixed-effect models using lme4," J. Stat. Software **67**, 1–48.
Behrman, A. (**2023**). *Speech and Voice Science*, 4th ed. (Plural, San Diego, CA).
Boersma, P., and Weenink, D. (**2024**). "Praat: Doing phonetics by computer (version 6.3.02) [computer program]," available at http://www.praat.org (Last viewed November 30, 2024).
Bunch, M., and Chapman, J. (**2000**). "Taxonomy of singers used as subjects in scientific research," J. Voice **14**, 363–369.
Charpentier, F., and Stella, M. (**1986**). "Diphone synthesis using an overlap-add technique for speech waveforms concatenation," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, April 7–11, Tokyo, Japan (IEEE, New York), pp. 2015–2018.
Corretge, R. (**2012–2025**). "Praat vocal toolkit [computer program]," https://www.praatvocaltoolkit.com/index.html (Last viewed October 9, 2025).
Corso, J. F., and Cohen, A. (**1958**). "Methodological aspects of auditory threshold measurements," J. Exp. Psychol. **55**, 8–12.
Davids, J., and LaTour, S. (**2012**). *Vocal Technique: A Guide for Conductors, Teachers, and Singers* (Waveland, Long Grove, IL).
Eberhart, C. (**1962**). "Diction," J. Sing. NATS bulletin **18**, 8–9.
Fuchs, S., and Birkholz, P. (**2019**). "Phonetics of consonants," in *Oxford Research Encyclopedia of Linguistics*, available at https://oxfordre.com/linguistics/view/10.1093/acrefore/9780199384655.001.0001/acrefore-9780199384655-e-410 (Last viewed May 13, 2025).
Gregg, J. W. (**1991**). "On articulation— Part I," J. Sing. **47**, 30–32.
Halle, M. (**1967**). "On the modern study of speech sounds," Int. Soc. Sci. J. **19**, 17–27.
Heald, S. L. M., and Nusbaum, H. C. (**2014**). "Speech perception as an active cognitive process," Front. Syst. Neurosci. **8**, 35.
Howard, D., and Angus, J. (**2006**). *Acoustics and Psychoacoustics*, 3rd ed. (Routledge, New York).
Johnson, K. (**2012**). *Acoustic and Auditory Phonetics* (Wiley-Blackwell, Malden, MA).
Kent, R. D., and Read, C. (**2002**). *The Acoustic Analysis of Speech* (Delmar, Clifton Park, NY).
Koffi, E. (**2020**). "A comprehensive review of intensity and its linguistic applications," Linguist. Portfolios **9**, 2–27.
Koutsogiannaki, M. C. (**2016**). "Intelligibility enhancement of casual speech based on clear speech properties," Ph.D. dissertation, University of Crete, Heraklion, Greece.
Kurowski, K., and Blumstein, S. E. (**1984**). "Perceptual integration of the murmur and formant transitions for place of articulation in nasal consonants," J. Acoust. Soc. Am. **76**, 383–390.
LaBouff, K. (**2008**). *Singing and Communicating in English—A Singer's Guide to English Diction* (Oxford University Press, New York).
Lehiste, I. (**1970**). *Suprasegmentals* (MIT Press, Cambridge, MA).
Lenth, L. R. (**2024**). "emmeans: Estimated marginal means, aka least-qquares means *R* package (version 1.10.3-090003) [computer program],"

available at https://rvlenth.github.io/emmeans/ (Last viewed July 13, 2024).
Liberman, A. M., Delattre, P. C., Cooper, F. S., and Gerstman, L. J. (**1954**). "The role of consonant-vowel transitions in the perception of the stop and nasal consonants," Psychol. Monographs: Gen. Appl. **68**, 1–13.
Lindblom, B., and Sundberg, J. (**2007**). "The human voice in speech and singing," in *Springer Handbook of Acoustics*, edited by D. T. Rossing (Springer, New York).
Menn, K. H., Männel, D., and Meyer, L. (**2023**). "Phonological acquisition depends on the timing of speech sounds: Deconvolution EEG modeling across the first five years," Sci. Adv. **9**, eadh2560.
Meyer, J. (**2009**). *Acoustics and the Performance of Music. Manual for Acousticians, Audio Engineers, Musicians, Architects and Musical Instrument Makers*, 5th ed. (Springer, New York).
Miller, R. (**1986**). *The Structure of Singing: System and Art in Voice Technique* (Schirmer, New York).
Nair, A. (**2021**). *The Tongue as a Gateway to Voice, Resonance, Style, and Intelligibility* (Plural, San Diego, CA).
Nelson, H. D., and Tiffany, W. R. (**1968**). "The intelligibility of song: Research results with a new intelligibility test," NATS Bull. **25**, 22–33.
Phillips, G. L. (**2002**). "Diction: A rhapsody," J. Sing. **58**, 405–409.
R Core Team (**2021**). "*R*: A language and environment for statistical computing" (R Foundation for Statistical Computing, Vienna, Austria), available at https://www.R-project.org/ (Last viewed October 9, 2025).
Sharnova, S. (**1947**). "Diction," NATS Bull. **3**, 4.
Shaw, R., and Blocker, R. (**2004**). *The Robert Shaw Reader* (Yale University Press, New Haven, CT).
Sundberg, J. (**1987**). *The Science of the Singing Voice* (Northern Illinois University Press, DeKalb, IL).
Svec, J., and Granqvist, S. (**2018**). "Tutorial and guidelines on measurement of sound pressure level in voice and speech," J. Speech. Lang. Hear. Res. **61**, 441–461.
Titze, I. (**1982**). "Why is the verbal message less intelligible in singing than in speech," NATS Bull. **38**, 37.
Trost, S. (**2025**). "Alphabet and character frequency: Estonian (Eesti)," available at https://www.sttmedia.com/characterfrequency-estonian (Last viewed May 13, 2025).
Vurma, A., Meister, E., Meister, L., Ross, J., Raju, M., Kala, V., and Dede, T. (**2023**). "The intensities of vowels and plosive bursts and their impact on text intelligibility in singing," J. Acoust. Soc. Am. **154**, 2653–2664.
Ware, C. (**1998**). *Basics of Vocal Pedagogy: The Foundations and Process of Singing* (McGraw-Hill, Boston, MA).
Waring, F. (**1945**). *Tone Syllables* (Shawnee, Delaware Water Gap, PA).