JASA ARTICLE

The role of phonetic overlap for speaker discrimination

Leah Bradshaw,[a] Eleanor Chodroff, and Volker Dellwo
Department of Computational Linguistics, University of Zürich, Zürich, Switzerland

ABSTRACT:
Linguistic information influences processing of speaker information in a multitude of ways, whether this arises from the listener's familiarity with the language or dialect spoken, or existing linguistic relationships between spoken words. Specifically, phonological and semantic relationships between spoken words have been observed to influence a listeners' ability to discriminate voices. This study aims to develop our understanding of how different kinds of linguistic relationships, namely, phonetic relationships, influence the processing of speaker information. We conducted two experiments, a voice discrimination task and a voice similarity rating task, in which listeners were presented with pairs of speakers producing two words with various degrees of phonetic overlap. On the whole, higher quantities of phonetic overlap corresponded to higher speaker discrimination performance and higher similarity scores; however, the type of the phonetic overlap also mattered. Overlapping vowel segments showed substantial utility, while overlap of the phonological rhyme alone substantially lower performance. Results from this condition suggest that a phonological relationship within the word pair can interfere with otherwise increased quantities of phonetic overlap. These findings highlight the salience of the phonological rhyme in voice processing, as well as the overall impact of phonetic overlap. © *2025 Acoustical Society of America*.
https://doi.org/10.1121/10.0036562
(Received 8 August 2024; revised 10 April 2025; accepted 12 April 2025; published online 12 May 2025)
[Editor: James F. Lynch] Pages: 3572–3589

I. INTRODUCTION

Speech signals encode not only linguistic cues necessary for interpreting speech content, but also indexical cues, which carry information relating to the personal characteristics of the speaker (Abercrombie, 1967). The processing of linguistic and indexical information in speech is shown to interact to a considerable degree, with each kind of information observed to impact the processing of the other. From one perspective, indexical information is shown to substantially impact how listeners process linguistic information. For instance, listeners display a greater capacity for word and sentence recognition when they are presented with speech from a familiar speaker (Nygaard and Pisoni, 1998; Nygaard *et al.*, 1994; Levi *et al.*, 2011; Souza *et al.*, 2013). In addition, being presented with multiple talkers slows processing in speeded phoneme classification tasks (Mullennix and Pisoni, 1990) and reduces the listener's capacity to recall lists of spoken words (Goldinger *et al.*, 1991). Further, when speakers are presented with different regional accents, word recognition is impaired (Imai *et al.*, 2005) and slower (Scott and Cutler, 1984), while cross-model priming from an acoustic prime to a visual word is either reduced or absent entirely (Sumner and Samuel, 2009).

Equally, the reverse effect has been demonstrated, with studies showing the influence of linguistic information on the processing of indexical features. Arguably, the most well-known phenomenon in this case is the language familiarity effect (LFE) (for detailed overviews, see Levi, 2019; Perrachione, 2019), which denotes that listeners show a greater capacity to process speaker or indexical information in their native language (e.g., Goldstein *et al.*, 1981; Hollien *et al.*, 1982; Thompson, 1987). Monolingual English speakers have been observed repeatedly to have impaired voice recognition when presented with voices speaking a foreign language [e.g., with Mandarin speakers (Perrachione and Wong, 2007); with French speakers (Phillipon *et al.*, 2007)]. The effect also persists when the speakers of both languages are kept consistent, and only the language is modulated; for instance, Goggin *et al.* (1991) observed that, despite using English–German bilingual speakers for the voice stimuli, monolingual English and German listeners remained significantly impaired when recognizing targets speaking in their non-native language.

Historically, a key argument regarding the LFE was that the issue for speaker recognition in non-native languages was challenging due to limited speech comprehension, in which listeners were unable to effectively process the lexical–semantic information in speech. Critically, however, the LFE has also been observed in cases where processing of lexical–semantic information is extremely limited. For example, when English and Mandarin listeners were given pairs of time-reversed sentences in each language, in which the intelligibility of the speech itself is completely disrupted, while some phonological information is preserved (i.e., the formant structure of long vowels; Binder *et al.*, 2000). With this time-reversed stimuli, all listeners,

[a]Email: leah.bradshaw@uzh.ch

https://doi.org/10.1121/10.0036562

regardless of native status, were unable to process the linguistic content of the utterances (Fleming *et al.*, 2014). Findings showed that both groups of listeners rated speakers as more dissimilar in their native language, suggesting that it is the processing of phonological information rather than lexical–semantic information that drives the LFE. However, Perrachione *et al.* (2015) were not able to replicate these findings in a speaker identification task, observing that neither English nor Mandarin listeners were better at identifying speakers of their own language when speech was time-reversed. Therefore, while phonological information may have been sufficient for making speaker similarity judgments, lexical–semantic information appeared to be more useful for speaker identification.

Further studies have sought to tease apart the role of lexical–semantic and phonological information for the LFE, finding both are likely to be influential, but to differing extents. For example, lexical–semantic information was shown to be significantly beneficial when listeners are presented with real words compared to non-words (Perrachione *et al.*, 2015; Xie and Myers, 2015; Zarate *et al.*, 2015; Abu El Adas and Levi, 2022). When presented with non-word structures that follow the phonological structure of a listener's native language, this meaningless speech corresponds to lower speaker discrimination accuracy. However, further observations suggest that phonological information does still play a role in the LFE, but with lesser importance compared to lexical–semantic information. Specifically, being able to access the lexical–semantic content of the speech shows an advantage over phonologically similar but meaningless speech (e.g., English vs "Jabberwocky English"; Xie and Myers, 2015). Equally, having access to the phonological information, but not the lexical–semantic information, is advantageous over speech contexts where neither information is accessible to the listener (e.g., "Jabberwocky" English vs Mandarin; Xie and Myers, 2015). In addition, speaker discrimination is higher for speech with a similar phonological structure to a listener's native language compared to one that is highly dissimilar (e.g., English listeners with German vs Mandarin; Perrachione *et al.*, 2015). Therefore, it seems that, combined lexical–semantic and phonological information has the edge for the LFE; however, phonological information remains advantageous.

Further evidence of the utility of phonological processing can be observed for dialectal familiarity, whereby lexical–semantic information is completely accessible, but specific pronunciations may be unfamiliar. In particular, the familiarity of listeners with a speaker's dialect can also influence voice recognition performance, a phenomenon coined the "other-accent" effect (Stevenage *et al.*, 2012). The other-accent effect has been observed in dialects of English with large [Australian vs British English (BE); Vanags *et al.*, 2005] and small (Glaswegian vs Southhampton English; Stevenage *et al.*, 2012) geographical proximities, as well as in some Dutch dialects (standard vs non-standard Dutch; Kerstholt *et al.*, 2006).

The influence of linguistic processing on the processing of indexical information is already highly complex, with confounds observed when presenting listeners with languages or dialects to which they are unfamiliar. However, these kinds of processing interactions do not exist solely in relation to familiarity with linguistic or phonological content. More recent studies have also considered the influence of different types of linguistic information on speaker perception, when both the lexical–semantic and phonological information are accessible to the listeners. In particular, Narayan *et al.* (2017), and later, Quinto *et al.* (2020) in a replication of the original study, explored the influence of *top-down information* on the perception of speaker differences. In both studies, participants were required to judge whether the voices in two audio samples were from the *same* or two *different* speakers. They explored differences in speaker discrimination performances in four experimental conditions of varying degrees of lexical–semantic coherence and cross-compared all conditions. Word pairs formed either a semantic relationship, namely, forming lexical compounds (e.g., "day–dream") or reversed lexical compounds (e.g., "dream–day"), a phonological relationship, namely, rhyming pairs (e.g., "day–bay"), or no lexical relationship in an unrelated word condition (e.g., "day–bee").

In the first study, Narayan *et al.* (2017) observed that participants were more likely to judge two words as being spoken by the same speaker when the presented word-pair was semantically related, specifically in the lexical compound condition, or phonologically related as opposed to being unrelated (no lexical relationship). However, an argument can be made for potentially inflated findings in this experiment, given the inclusion of cross-gender pairs for different-speaker comparisons. Further, these findings stem from an observation that discrimination accuracy is higher and RTs are lower for *same speaker* trials in these conditions, but no specific measure of response bias was employed.

To address the methodological issues in the original study, Quinto *et al.* (2020) conducted a replication of the study but with a gender-matched speaker sample and additional analysis procedures: specifically, the use of signal detection theory metrics for participant sensitivity (d') and bias (criterion, c). Indeed, they observed greater discriminability of speaker identity in the phonological rhyme condition using d', and the same *same speaker* response bias observed in the previous study, indicated by c being significantly greater than zero. Comparatively, the findings for the lexical compound condition did not hold, but rather, greater performance was observed for the reverse compound condition. However, a large *same speaker* response bias was observed for the lexical compound condition, likely accounting for the increased performance observed by Narayan *et al.* (2017) for this condition.

The role of lexical–semantic coherence in these studies seemed to impact the likelihood of a *same speaker* response bias more so than the accuracy with which participants performed in the task. We can hypothesize here that this impact of the semantic relationship might stem from the higher likelihood of hearing a lexical compound as spoken by a

single speaker in everyday situations, rather than by two speakers in succession. In comparison, the findings for the phonological rhyme condition suggest two plausible explanations. The increased performance that is observed for words containing a phonological rhyme may be a function of top-down processing, i.e., advantages stemming from the phonological relationship between the word pair, or a function of bottom-up processing, i.e., advantages stemming from the increased phonetic similarity between the word pair. Specifically, the increased phonetic similarity in the word pair allows a more direct comparison of speakers' voices. This is further highlighted by the fact that other kinds of linguistic relationships did not facilitate speaker discrimination. Specifically, being presented with a lexical compound did not improve discrimination accuracy, despite it having a clear lexical–semantic relationship. As these studies did not vary the degree of phonetic similarity across conditions, it is difficult to fully gauge the role of bottom-up processing.

To summarize, previous findings have shown that the influence of linguistic information on the processing of speaker information is incredibly complex and influenced by numerous factors. There are between-language differences with the LFE, within-language differences with the other-accent effect, and even differences within the same language and dialect depending on the linguistic relationships between the words. Interestingly, there is also uncertainty between the role of top-down (phonological) and bottom-up (phonetic) processing for word pairs that have a phonological relationship, i.e., a phonological rhyme. It is likely that increased phonetic similarity[1] is highly beneficial for bottom-up processing, as it allows listeners to make more direct comparisons between talkers, resulting in more accurate discrimination judgments. Equally, as a salient phonological unit, the phonological rhyme may also be driving this improvement.

While increased phonetic overlap plausibly facilitates speaker discrimination, it is nevertheless unlikely that all kinds of phonetic overlap are equally useful. We could reasonably assume that the utility of the phonetic overlap in previous studies is highly driven by the overlapping vowel portions in the two words, given what has previously been shown about the speaker discriminatory capacities of vowels. From an acoustic perspective, speaker specificity has frequently been stated in vowel productions (Amino and Arai, 2007; Andics *et al.*, 2007; Eatock and Mason, 1994; Loakes, 2004). More specifically, regarding speaker discrimination, vowels are shown to be the most effective speech segments for discriminating speakers across several languages (for a review, see Amino *et al.*, 2006), and listeners have previously been shown to be capable of discriminating speakers from vowel portions alone (Dellwo *et al.*, 2018). Therefore, it is possible that the nature of the phonetic overlap will play a large role in its utility for speaker discrimination.

The present study aims to expand on previous studies that imply that bottom-up (phonetic) processing drives performance for speaker discrimination, and specifically, the idea that being presented with words with increased phonetic similarity renders voice discrimination easier. We consider how varying degrees of phonetic overlap influences a listener's ability to discriminate speakers and how this corresponds to judgments of overall speaker similarity. The study consists of two experiments: a speaker discrimination task and a voice similarity judgment task.

First, we conducted a speaker discrimination experiment, in line with previous studies (Narayan *et al.*, 2017; Quinto *et al.*, 2020), in which participants were presented with a pair of voices and asked to judge whether the voices in the samples came from the *same* or two *different* speakers. Listeners were presented with word pairs corresponding to five experimental conditions that manipulated the amount of overlap in a consonant-vowel-consonant (CVC) word, where each C was a stop consonant. These were a *same word* condition, where listeners were presented with the exact same CVC word (e.g., *tap–tap*); a *phonological rhyme* condition, where listeners were presented with rhyming words, i.e., _VC overlap (e.g., *tap–gap*); a *vowel overlap* condition, where listeners were presented with words that contained the same vowel portion, i.e., _V_ overlap (e.g., *tap–bat*); a *frame overlap* condition, where listeners were presented with words with C_C overlap (e.g., *tap–top*); and a *different word* condition, where listeners were presented with words containing no overlapping segments in the same phonotactic position (e.g., *tap–got*).[2]

These conditions were chosen to explore in more detail how the degree of phonetic overlap influences speaker discrimination. Namely, it allows us to tease apart how the quantity of overlap, i.e., how many overlapping segments the words contain, interacts with the quality of the overlap, i.e., how useful the overlapping segment is for speaker discrimination. Indeed, we predict that increasing the amount of phonetic overlap will show some performance improvements, in line with previous findings regarding phonological rhyme overlap (Narayan *et al.*, 2017; Quinto *et al.*, 2020). However, given that vowels are inherently more useful than consonants for speaker discrimination (Amino *et al.*, 2006), especially in comparison to the stop consonants that we employ, we expect performance improvements in conditions that contain an overlapping vowel portion. Specifically, we anticipate that performance will be better in the *same word*, *phonological rhyme* and *vowel overlap* conditions compared to the *frame overlap* and *different word* conditions due to differing quality of overlap, and graded improvement from the *vowel overlap* to the *same word* conditions from increased quantity of overlap.

Further, the use of these conditions offers a comparison to the phonological rhyme condition presented in previous studies (Narayan *et al.*, 2017; Quinto *et al.*, 2020). Specifically, if the observed performance improvements from the phonological rhyme condition are a result of bottom-up processing, i.e., an advantage of the increased phonetic similarity between the words, we should observe greater performance with more phonetic overlap (e.g., *same word* > *phonological rhyme* > *vowel overlap*). Alternatively, if the performance improvements stem from top-down processing,

https://doi.org/10.1121/10.0036562

i.e., an advantage of the salient phonological relationship in the word pair, then performance in the *phonological rhyme* condition should be uniquely higher than in other conditions. Since previous studies observed greater performance improvements for the phonological rhyme condition compared to other conditions with a linguistic relationship, e.g., word pairs forming a lexical compound, we expect that it is the increased phonetic similarity, and as such, the utility of the overlap for bottom-up processing, that corresponds to performance improvements.

In addition to the speaker discrimination task, we completed a second experiment, which consisted of a voice similarity judgment task. This was implemented to further investigate the response bias observed in Quinto *et al.* (2020), in which speakers were more likely to respond that voices came from the *same speaker* in their phonological rhyme condition compared to their unrelated condition, in which there was no linguistic relationship between the word pairs. We asked participants to listen to a pair of voices and judge how similar they sounded on a scale from 1 to 6. This allowed us to validate the findings from Experiment 1 using a somewhat related task and also assess the source of any response bias. If present, we can ascertain whether it is related to a perceptual difference caused by the conditions, or if it was merely an artefact of the task, i.e., listener uncertainty. If it is the case that response biases are linked to perceptual differences, we expect to observe higher similarity scores for all conditions which are observed to have a *same speaker* response bias in Experiment 1, particularly for different speaker pairings.

II. EXPERIMENT I: SPEAKER DISCRIMINATION

A. Stimuli recordings

Fifty-five monosyllabic CVC English words were identified in which the initial and final *C* elements were a stop consonant, /p t k b d ɡ/, and the *V* was one of /æ ɪ ɛ ɔ/ (e.g., "tap" /tæp/, "beg" /bɛɡ/). Given our interest in the influence of vowel overlap on speaker identification, we avoided the use of any other segments shown to be useful speaker discriminants, such as nasals (Amino and Arai, 2007; Eatock and Mason, 1994) and the fricative /s/ (Andics *et al.*, 2007; Eatock and Mason, 1994). In addition, some vowels, such as the close-back and open-mid vowels, /ʊ/ and /ʌ/, were excluded given dialect variation which may be present among the recorded speakers. All possible BE words that could be constructed using these segments were recorded. Word pairs for experimental stimuli were then created from these recordings (Appendix).

Stimuli were recorded from eight female speakers of BE—seven of whom had been living in Switzerland for at least 1 year (Table I). All speakers were assessed by a native BE speaker prior to recording to confirm they had a sufficiently similar sounding dialect and ensure any influence of living in Switzerland on the speakers' dialects was minimal. Recordings were made in a sound-attenuated booth at the Linguistic Research Infrastructure (LiRI) laboratory at the University of Zürich using a Røde NT1 microphone (Røde Microphones, Silverwater, Sydney, Australia). Recordings were made directly to disk in WAV format at 44.1 kHz sampling rate, 16 kbit/s bitrate using ProTools software (Avid Technology Inc, Burlington, MA; Brooks and Gotcher, 2023). Each speaker participated in one recording session, lasting approximately 15 min. In the session, each word was presented individually to the speakers. Speakers were instructed to read the 55-item word list aloud at a comfortable volume and to maintain a consistent pitch and prosodic pattern throughout the session to minimize strong effects of list intonation. Two or three repetitions were recorded for each speaker, depending on the success of the first production, i.e., whether the speakers made any errors (e.g., incorrect pronunciation) or the recordings were of poor quality.

TABLE I. Demographic and acoustic information for each of the eight speakers. Standard deviations for mean *f*0 and mean word duration are noted in parentheses. SSBE refers to a Standard Southern British English dialect.

Speaker	Age	Dialect	Mean *f*0 (Hz)	Mean word duration (ms)
Speaker 1	23	Midlands	197 (45)	535 (76)
Speaker 2	28	SSBE	179 (56)	510 (60)
Speaker 3	27	SSBE	210 (40)	566 (74)
Speaker 4	27	SSBE	189 (53)	487 (50)
Speaker 5	26	Midlands	192 (43)	530 (55)
Speaker 6	24	SSBE	184 (53)	603 (56)
Speaker 7	23	SSBE	200 (61)	611 (83)
Speaker 8	23	SSBE	177 (49)	497 (81)

Recordings were all assessed auditorily to select stimuli for the perception experiments. Typically, a speaker's second and third repetitions were used in the stimulus set, with the second repetition being the default production in the stimulus set, and the third repetition being used as the comparison material for *same speaker–same word* trials. For speakers who only recorded two repetitions, the first production was the default production in the stimulus set and the second repetition was used for *same speaker–same word* trials. Following the default selections, coherence of stimuli pairs was further assessed auditorily to ensure no large dialectal differences were present in a single trial. For instance, our recordings contained one speaker who tended to hyperarticulate /æ/ vowels in a way that was inconsistent with the other speakers. In instances where this occurred, all repetitions of the word in question were assessed and the production that was most in line with the other speaker in the pair was selected for use in the trial.

B. Methods

1. Participants

Ninety-nine native speakers of American English with no reported history of speech or hearing disorders (34 female, 57 male, seven non-binary, one undisclosed gender) were recruited using the online-recruitment platform, *Prolific* (Prolific Academic Ltd., London, UK). American English participants were selected based on the assumption of lesser familiarity with BE dialects. Therefore, if any of

the speakers' dialects had been influenced by their time living in Switzerland, these small differences would not influence task performance. We acknowledge that the use of American English listeners may influence the overall performance in the task as a result of the other-accent effect (Sec. I); however, all listeners should be approximately equally influenced by this difference, and we expect that listeners will still be able to complete the task successfully. Moreover, the lesser familiarity with the dialect will further remove the possibility of observing ceiling effects.

All participants were monolingual English speakers between ages 18 and 35 yrs who had not lived outside of the United States for a period of longer than six months. Participants were pre-screened using the *Prolific* built-in filters to ensure they met these criteria and participants also later confirmed their status using a demographic questionnaire. As the experiment was administered online, participants were required to complete a debriefing questionnaire after completion of the experiment in which they could disclose any technical issues they faced during the experiment. In total, two participants were excluded from the data analysis after reporting connection issues resulting in unsuccessful completion of some trials. In addition, participants were asked to briefly explain the task they had just completed, to ensure the task was understood correctly. One additional participant was excluded based on their answer to this question which indicated they had misunderstood the task. In total, 96 participants were included in the final analyses.

2. Materials

The 55 monosyllabic English words were combined into pairs according to five experimental conditions with varying degrees of phonetic similarity. The conditions were (1) *same word*, in which the exact same CVC word was presented (e.g., tap–tap); (2) *phonological rhyme*, in which the words rhyme, i.e., _VC overlap (e.g., tap–gap); (3) *vowel overlap*, in which words had contained the same vowel portion, i.e., _V_ overlap (e.g., tap–bat); (4) *frame overlap*, in which words had C_C overlap (e.g., tap–top); and (5) *different word*, in which words had no overlapping segments in the same phonotactic position (e.g., tap–-got). These conditions were created such that the quantity and quality of phonetic overlap was adjusted slightly in each condition; a visualization of how these differ is presented in Table II.

Pairs were created to ensure that, in the *frame overlap* and *phonological rhyme* conditions, the non-overlapping consonants always differed in place of articulation, i.e., no non-overlapping segments had the same place of articulation but only differed in voicing. Additionally, pairs were created to avoid any common phases or compound words, e.g., *tip–top* and *kit–cat*, that might create a semantic relationship between the two words. To minimize the opportunity for learning, unique word pairings were created for all possible *same* and *different speaker* combinations for each condition, such that no voice clip was heard twice throughout the experiment. Speakers were not heard to repeat the same word twice within the stimulus set, aside from in the *same word* condition, where the samples presented were always different productions. In total, there were 180 possible trials in the final experiment.

TABLE II. Segment overlap in the five experimental conditions.

Condition	C	V	C
Same word e.g., *tap–tap* /tæp/–/tæp/	✓	✓	✓
Phonological rhyme e.g., *tap–gap* /tæp/–/gæp/	-	✓	✓
Vowel overlap e.g., *tap–bat* /tæp/–/bæt/	-	✓	-
Frame overlap e.g., *tap–top* /tæp/–/tɔp/	✓	-	✓
Different word e.g., *tap–got* /tæp/–/gɔt/	-	-	-

3. Procedure

Participants completed a timed AX-discrimination task, in which they were asked to determine whether the words in the two audio samples were spoken by the same or two different speakers by pressing a button on the screen. These buttons were inactive, but visible, to participants until the end of the audio stimulus presented, i.e., the offset of the second word. In each experiment trial, the first stimulus was presented to the listener 2000 ms after the screen appearance or the button response from the previous trial, resulting in an intertrial interval of 2000 ms. The second stimulus was presented 1700 ms after the onset of the first stimulus, resulting in an interstimulus interval (ISI) of approximately 1000 ms based on a maximum stimulus duration of 700 ms. Due to differing articulation rates of the speakers, the ISI varied slightly depending on the stimulus pair (IQR: 132 ms). ISI differences were balanced across the conditions and further accounted for by a random effect of trial in the relevant statistical models.

Listeners were instructed to answer as quickly and accurately as possible and were given 3000 ms after the offset of the second word to input their response. Participants completed 80 experimental trials in total (40 same speaker, 40 different speaker). Eight same and different speaker trials were presented in each of the five experimental conditions, corresponding to the eight voices present within the experiment. All participants were given a same speaker comparison for each of the voices plus a random selection of different speaker comparisons to balance the number of same and different speaker trials in the experiment, given the higher possible number of different speaker comparisons. No participant received an identical stimulus set for different speaker trials. The task took around 15 min to complete (median time: 12.22 min).

The experiment was designed and administered using the online experiment builder platform, Gorilla.sc (Cauldron Science Ltd., Cambridge, UK; Anwyl-Irvine *et al.*, 2020).

https://doi.org/10.1121/10.0036562

Participants received a remuneration of £2.25 for their participation, corresponding to an average hourly rate of £9 as recommended by *Prolific*. Prior to the main experiment, listeners completed a headphone screening task to ensure they were wearing headphones (Woods, *et al.*, 2017). Participants were required to listen to three pure tones at 200 Hz and were required to select the softest tone. The task was designed such that one of the tones was in antiphase across stereo channels, meaning it was attenuated when heard over loudspeakers, but not headphones. Therefore, the "softest" tone in each trial would differ depending on how the audio was heard, resulting in very low task performance when completed without headphones. This task was completed three times, and participation in the full experiment was contingent on 100% accuracy across the trials. Participants were allowed three attempts to pass this headphone check. If they failed all three attempts, it was assumed the participant had not followed the instructions to wear headphones and they were rejected from the experiment. However, no participant within the experiment failed the headphone task.

4. Analyses

The analysis procedure was conducted using the same three sets of analyses as Quinto *et al.* (2020); mean accuracy, RT (ms), and the signal detection statistics of performance sensitivity (d') and bias (c). The first two measures allowed for direct comparisons of our findings with two of the previous studies (Narayan *et al.*, 2017; Quinto *et al.*, 2020), while the signal detection statistics offer direct comparison with the findings in Quinto *et al.* (2020).

First, mean accuracy was calculated by dividing the total number of correctly answered trials by the total number of trials and multiplying by 100. Following previous studies, mean accuracy was assessed with separate consideration to same and different speaker trials. However, as this measure is substantially influenced by listener response bias and therefore does not provide an accurate overview of performance accuracy, we opted only to include this measure for congruence with the previous studies and to assess it using descriptive but not inferential statistics.

To complement this first analysis, we also assessed RT with the same separate consideration of same and different speaker trials. RT (ms) was calculated as the time taken to respond after audio offset. Following Narayan *et al.* (2017) and Quinto *et al.* (2020), RT measures were log-transformed and RTs greater than 2000 ms were excluded ($n = 78$). In addition, RTs less than 100 ms were excluded to remove any possible accidental responses ($n = 103$). In total, 5558 responses were included in the final analysis.

Variation in RT was modelled with a linear-mixed effects regression using the *lme4* R package (Bates *et al.*, 2003), with fixed effects of *Word condition* (*same word*, *phonological rhyme*, *vowel overlap*, *frame overlap*, *different word*) and *Trial type* (*same speaker*, *different speaker*), their interactions, plus by-*Participant* and by-*Trial* random intercepts. Factors were sum-coded, such that the interpretation of comparisons are relative to the average RT across all trials; a baseline condition was not assumed. For better interpretations of the interactions, pairwise comparisons of the *Word condition* levels within each *Trial type* were obtained using the *emmeans* R package (Lenth, 2017).

In addition, for a more accurate assessment of performance, we calculated listener sensitivity and bias using the corresponding signal detection statistics of d' and c. The measures were calculated using the four possible combinations of trial and response types present in the experiment. Namely, responding *different* in a different speaker trial was coded a *hits*, while responding *same* in these trials was coded as a *miss*. Responding *same* in a same speaker trial was coded as a *correct rejection*, while responding *different* in these trials was coded as a *false alarm*. Using this, we calculated each participant's hit rate (H) and false alarm rate (FA) to later calculate participant sensitivity (d') (Grier, 1971) and bias (criterion location, c) (Macmillan, 2002; Macmillan and Creelman, 1990). For reference, H is the proportion of different speaker (signal trials) where a participant responded *different*: $H = P(\text{response } \textit{different} \mid \text{different speaker trials})$. While FA is the proportion of same speaker trials (noise trials) where a participant responded *different*: $\text{FA} = P(\text{response } \textit{different} \mid \text{same speaker trials})$. d' and c were then calculated using the following formulas:

$$d' = z(H) - z(FA) \tag{1}$$

and

$$c = -0.5\big(z(H) + z(FA)\big), \tag{2}$$

where H is the hit rate, FA is the false alarm rate, and z() indicates a z-score. To avoid generating z- scores of positive or negative infinity, H and FA of 0 and 1 were replaced, respectively, with $1/(2*N)$ and with $1 - (1/(2*N))$, where N is the maximum possible number of false alarms (Stanislaw and Todorov, 1999). A d' and c measure was calculated for each participant in each of the five experimental conditions. A higher d' indicates greater participant sensitivity in distinguishing between *same speaker* and *different speaker* trials. Criterion location (c) at a neutral point, $c = 0$, is indicative of a lack of bias, while negative values of c indicate a bias towards responding *different speaker* and positive values of c indicate a bias towards responding *same speaker*.

For each of d' and c, we fit linear mixed-effects models with one fixed effect (Word condition with five levels: *same word*, *phonological rhyme*, *vowel overlap*, *frame overlap*, *different word*) and a by-*Participant* random intercept on d' and c separately. As the calculation of the signal detection theory measures was aggregated across trials, we could not include a by-*Trial* random intercept in these models. Factors were again sum-coded, such that the interpretation of the comparisons are relative to the average performance across all participants and a baseline condition is not assumed. In addition, we conducted a series of *post hoc* pairwise comparisons on the factor *Word condition* using the *emmeans* package

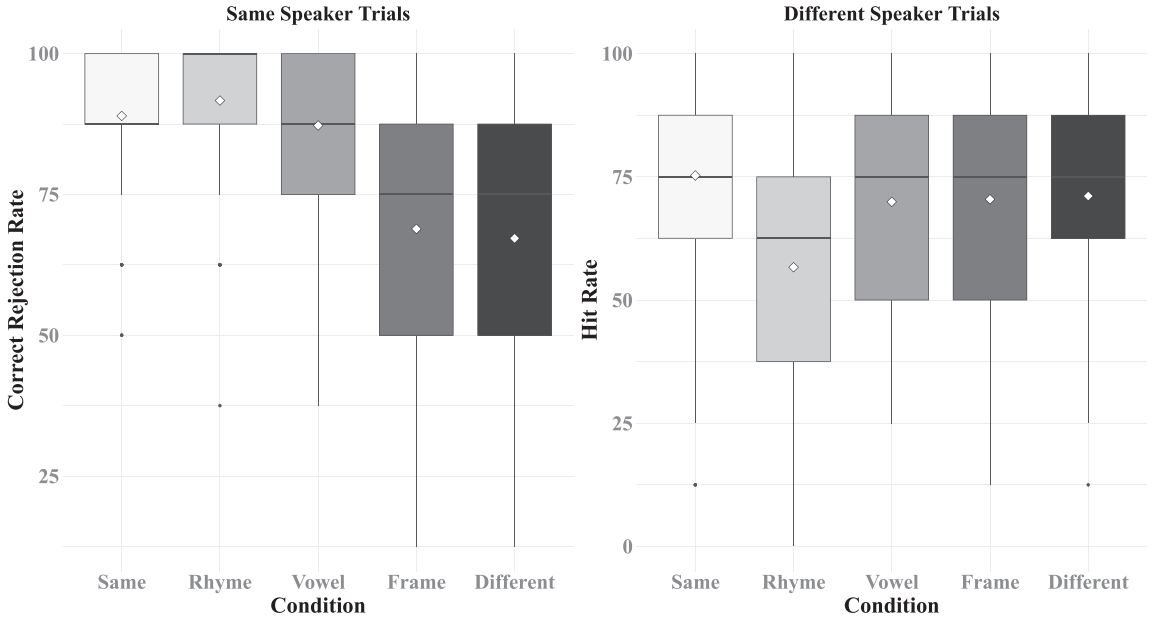

FIG. 1. Mean participant accuracy by *Trial type* (same or different speaker) across the five experimental conditions. For same speaker trials, accuracy corresponds to the correct rejection rate, and for different speaker trials, accuracy corresponds to the hit rate. Grand means are represented by the white points.

(Lenth, 2017). For the analysis of c, we also first conducted a one-sample t-test on c measures for each condition to test whether the bias differed significantly from zero. All statistical analyses were conducted in *R* version 4.3.2 (R Core Team, Vienna, Austria). Unless stated otherwise, the α level for assessing significance was set to 0.05 for all statistical analyses.

C. Results

1. Mean accuracy

To align with previous studies, analysis of mean accuracy across conditions was assessed separately for the two trial types (*same* vs *different speaker*) (Fig. 1). For *same speaker* trials, we observed higher mean accuracy in the *same word*, *phonological rhyme*, and *vowel overlap* conditions, but lower mean accuracy for the *frame overlap* and *different word* conditions. For *different speaker* trials, we observed lower mean accuracy in the *phonological rhyme* condition, but few obvious differences in any of the other conditions.

2. Reaction time

The linear mixed-effects model of RT revealed that, for same speaker trials, participants exhibited significantly faster RTs in same speaker trials in the same word

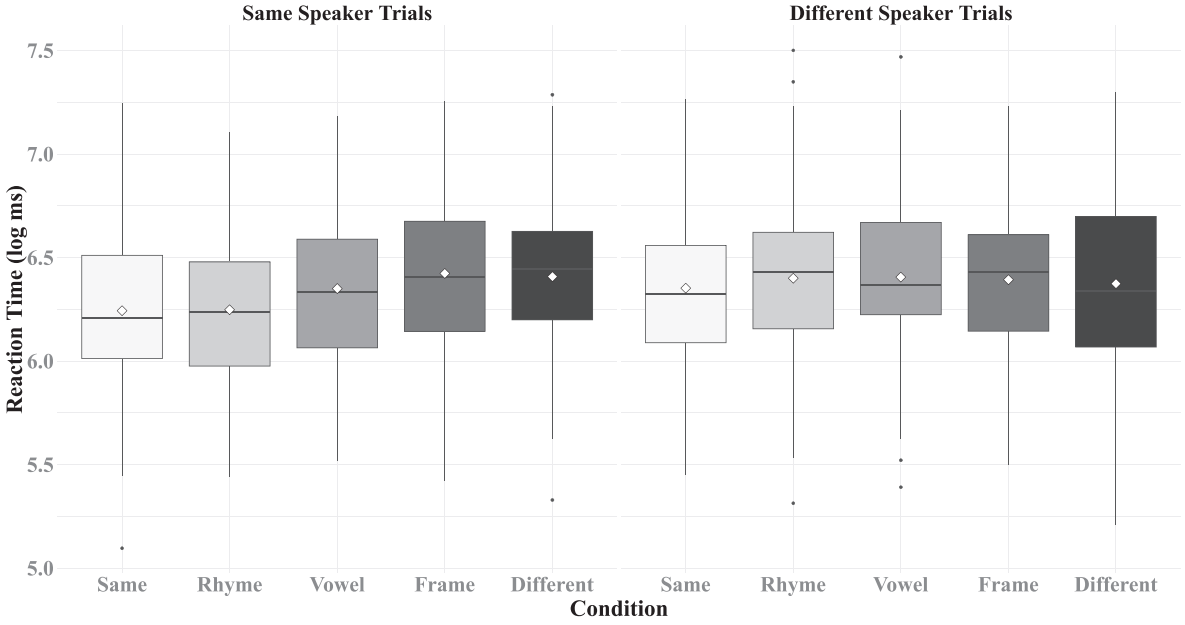

FIG. 2. Log-transformed reaction times in milliseconds (ms) for correct trials, averaged by participant and faceted by *Trial type* (same vs different speaker) across the five experimental conditions. Grand means are indicated by white points.

https://doi.org/10.1121/10.0036562

TABLE III. Pairwise comparisons (*emmeans*) of *Trial type* and *Word condition* levels on predicted estimates of the linear mixed-effects regression model for log-transformed reaction time (ms). Comparisons that reach the significance threshold (<0.05) are highlighted in bold.

Trial type	*Condition comparison*	Estimate	SE	*z*	*p*
Same speaker	*Same vs rhyme*	0.014	0.025	0.58	1.0000
	Same vs vowel	−0.094	0.025	−3.806	**0.0064**
	Same vs frame	−0.175	0.027	−6.586	**<0.001**
	Same vs different	−0.149	0.027	−5.617	**<0.001**
	Rhyme vs vowel	−0.109	0.025	−4.397	**0.0005**
	Rhyme vs frame	−0.189	0.027	−7.149	**<0.001**
	Rhyme vs different	−0.164	0.027	−6.174	**<0.001**
	Vowel vs frame	−0.081	0.027	−3.037	0.1079
	Vowel vs different	−0.055	0.027	−2.073	1.0000
	Frame vs different	0.025	0.028	0.905	1.0000
Different speaker	*Same vs rhyme*	−0.036	0.029	−1.251	1.0000
	Same vs vowel	−0.041	0.027	−1.501	1.0000
	Same vs frame	−0.054	0.027	−1.994	1.0000
	Same vs different	−0.024	0.027	−0.897	1.0000
	Same vs vowel	−0.005	0.029	−0.158	1.0000
	Rhyme vs frame	−0.018	0.029	−0.617	1.0000
	Rhyme vs different	0.012	0.029	0.401	1.0000
	Vowel vs frame	−0.013	0.028	−0.488	1.0000
	Vowel vs different	0.016	0.028	0.593	1.0000
	Frame vs different	0.030	0.028	1.081	1.0000

[$\beta = -0.002.59$, standard error (SE) $= 0.0012$, $p < 0.05$] and phonological rhyme conditions ($\beta = -0.0057$, SE $= 0.0012$, $p < 0.001$), and significantly slower RTs in same speaker trials in the frame overlap condition ($\beta = 0.0033$, SE $= 0.0012$, $p < 0.01$), each compared to average (see Fig. 2). The emmeans analysis of the interactions assessed the significance of all within-Trial type marginal contrasts (Table III). The results showed significant differences in same speaker trials between the same word condition and the vowel overlap, frame overlap, and different word conditions, as well as significant differences between the phonological rhyme condition and the vowel overlap, frame overlap, and different word conditions, also in same speaker trials. All other comparisons for same speaker trials and all comparisons for different speaker trials were not significant.

3. Sensitivity (d′)

Overall performance in both same and different speaker trials was analyzed using the parametric sensitivity measure d′ (Fig. 3). We found that participants showed significantly better performance in the same word ($\beta = 0.447$, SE $= 0.057$, $p < 0.001$) and vowel overlap ($\beta = 0.238$, SE $= 0.057$, $p < 0.001$) conditions compared to average. We also observed significantly worse performance in the frame overlap ($\beta = -0.315$, SE $= 0.057$, $p < 0.001$) condition compared to average.

As shown in Table IV, the analyses of the pairwise comparisons using emmeans indicated significant differences between the phonological rhyme, frame overlap, and different word conditions compared to the same word condition. Further, we observed significant differences between the phonological rhyme condition and the vowel overlap, frame overlap, and different word conditions, as well as between the frame overlap and different word conditions compared to the vowel overlap condition.

4. Bias (Criterion location, c)

The results of the one-sample *t*-tests for assessing the overall presence of bias revealed that *c* was significantly greater than zero in the *same word* [$t(95) = 5.25$, $p < 0.001$[, *vowel overlap* [$t(95) = 6.34$, $p < 0.001$], and the *phonological rhyme* [$t(95) = 13.53$, $p < 0.001$] conditions, indicating a bias for responding *same speaker*. In contrast, *c* measures in the *frame overlap* [$t(95) = -0.52$, $p = 0.602$] and the *different word* [$t(95) = -1.35$, $p = 0.18$] conditions did not significantly differ from zero, indicating no response bias (see Fig. 4).

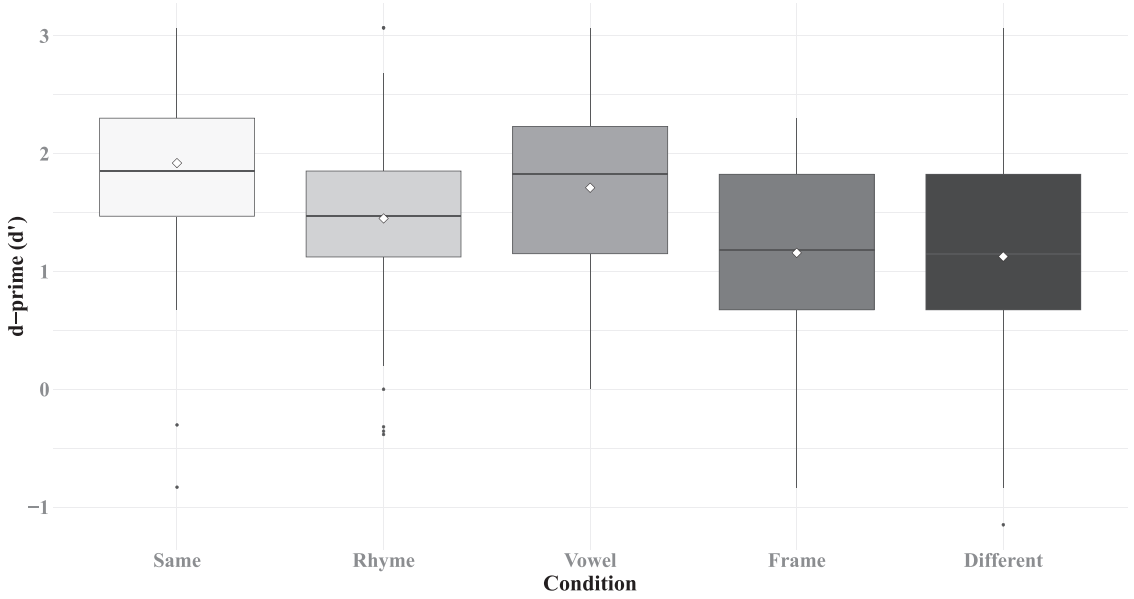

FIG. 3. Boxplot of participant performance (d') for the five experimental conditions. Grand means are represented by white points.

TABLE IV. Pairwise comparisons (*emmeans*) of *Word condition* levels on predicted estimates of the linear mixed-effects regression model for d'. Comparisons that reach the significance threshold (<0.05) are highlighted in bold.

Condition comparison	**Estimate**	**SE**	*z*	*p*
Same vs rhyme	0.469	0.090	5.240	**<0.001**
Same vs vowel	0.209	0.090	2.336	0.136
Same vs frame	0.762	0.090	8.502	**<0.001**
Same vs different	0.794	0.090	8.857	**<0.001**
Rhyme vs vowel	−0.260	0.090	−2.904	**0.032**
Rhyme vs frame	0.292	0.090	3.262	**0.011**
Rhyme vs different	0.324	0.090	3.616	**0.0031**
Vowel vs frame	0.552	0.090	6.166	**<0.001**
Vowel vs different	0.584	0.090	6.520	**<0.001**
Frame vs different	0.032	0.090	0.355	0.997

TABLE V. Pairwise comparisons (*emmeans*) of *Word condition* levels on predicted estimates of listener response bias (*c*) from the linear mixed-effects regression model. Comparisons that reach the significance threshold (<0.05) are highlighted in bold.

Condition comparison	**Estimate**	**SE**	*z*	*p*
Same vs rhyme	−0.331	0.054	−6.13	**<0.001**
Same vs vowel	−0.042	0.054	−0.770	0.939
Same vs frame	0.232	0.054	4.306	**<0.001**
Same vs different	0.271	0.054	5.020	**<0.001**
Rhyme vs frame	0.290	0.054	5.360	**<0.001**
Rhyme vs frame	0.564	0.054	10.436	**<0.001**
Rhyme vs different	0.602	0.054	11.150	**<0.001**
Vowel vs frame	0.274	0.054	5.076	**<0.001**
Vowel vs different	0.313	0.054	5.790	**<0.001**
Frame vs different	0.039	0.054	0.714	0.9532

The linear mixed-effects analysis indicated that *c* was significantly higher than average in the *phonological rhyme* ($\beta = 0.357$, $SE = 0.034$, $p < 0.001$) and *vowel overlap* ($\beta = 0.068$, $SE = 0.034$, $p = 0.048$) conditions, suggesting a bias towards responding *same speaker* in these conditions. In addition, *c* was observed to be lower than average in the *frame overlap* condition ($\beta = -0.206$, $SE = 0.034$, $p < 0.001$), suggesting a lower likelihood to respond *same speaker* in this condition. The *post hoc* analysis of the pairwise comparisons show differences in *c* between all conditions except the *same word* and *vowel overlap* conditions, and the *frame overlap* and *different word* conditions (Table V).

D. Discussion

In the present experiment, we aimed to explore how varying degrees of phonetic overlap influenced speaker discrimination performance, with a specific focus on the type of overlapping segments. First, we hypothesized that the quantity of phonetic overlap would play a role in how well participants were able to discriminate speakers. Specifically, we predicted that the more phonetic overlap in the word pair, the better the discrimination performance would be. Relatedly, if the utility of the *phonological rhyme* condition for speaker discrimination observed in Narayan *et al.* (2017) and Quinto *et al.* (2020) was driven by increased phonetic overlap, participants would perform equally as well when presented with the *same word*, as when presented with rhyming words. Second, we also predicted that the quality of the phonetic overlap would play a large role; namely, discrimination performance would be better for vowels, given that these segments have been previously shown to be highly speaker-specific and useful for speaker discrimination (see Sec. I). To examine these

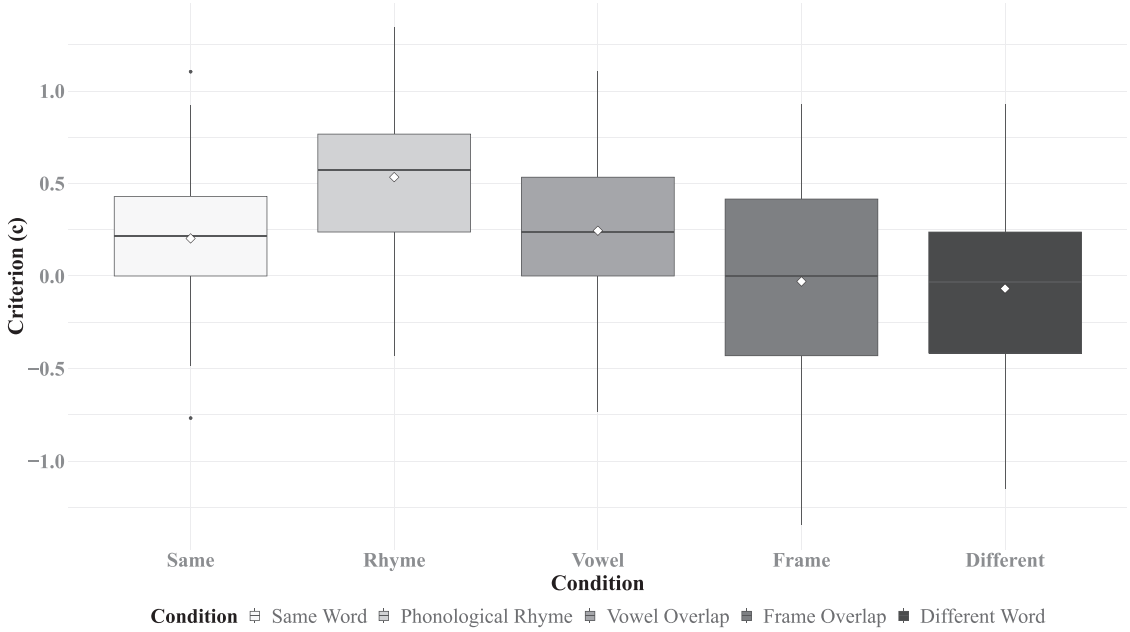

FIG. 4. Boxplot of participant criterion location (*c*) for the five experimental conditions. Grand means are represented by white points. A criterion location above zero indicates a bias towards responding *same speaker*.

https://doi.org/10.1121/10.0036562

hypotheses, we tested speaker discrimination performance using monosyllabic CVC English words across five experimental conditions; two corresponding to the original studies, the *phonological rhyme* and *different word* conditions, and three new conditions, the *same word*, *vowel overlap*, and *frame overlap* conditions. In addition to performance, we also examined patterns in RT and bias.

Overall, increased phonetic overlap generally improved performance, but with some notable exceptions. As revealed by d', higher performance was observed in the *same word*, *phonological rhyme*, and *vowel overlap* conditions in comparison to the *different word* condition, suggesting that having higher quantities of phonetic overlap corresponds to higher performance, and highlighting the importance of bottom-up processing for speaker discrimination. No difference, however, was observed between the *frame overlap* and *different word* conditions. Notably, performance in the *phonological rhyme* condition was significantly lower compared to both the *same word* and *vowel overlap* condition, suggesting that bottom-up processing advantages may not be the only thing at play in this condition.

Indeed, our findings suggest that the *phonological rhyme* condition was somewhat unique. Nevertheless, when considering the *phonological rhyme* condition compared with the linguistically unrelated *different word* condition, our findings were mostly complementary with the findings in the analogous conditions of the previous studies (i.e., *phonological rhyme* and *unrelated*) (Narayan *et al.*, 2017; Quinto *et al.*, 2020). Similar to Quinto *et al.* (2020), the improved accuracy in the *phonological rhyme* condition was only found for *same speaker* trials. For *different speaker* trials, accuracy in the phonological rhyme condition was noticeably worse than in all other conditions. It may be that the substantially higher performance in Narayan *et al.* (2017) for the *different speaker* trials may have been inflated by the inclusion of cross-gender trials. Further, our d' analysis further corroborates the findings from Quinto *et al.* (2020), in that significant performance improvements were nevertheless found in the *phonological rhyme* condition compared to the *different word* condition.

As predicted, the nature of the phonetic overlap was seemingly more crucial for its utility in speaker discrimination compared to mere quantity. In particular, we saw a tendency for listeners to perform better in conditions that contain an overlapping vowel portion, i.e., the *same word, phonological rhyme*, and *vowel overlap* conditions, compared to those that do not, i.e., the *frame overlap* and *different word* conditions. If it were the case that simply having a higher proportion of overlapping segments improved speaker discrimination performance, we would expect to see better performance in conditions with more overlap. Therefore, in the *frame overlap* condition, which has two overlapping segments, we would predict better performance than the *vowel overlap* condition, where there is only one overlapping segment. However, we observed the opposite in our findings. Nonetheless, given what we know already about the speaker discriminatory capacities of various segments, this finding was anticipated. Vowels have previously been observed to be highly speaker-specific and informative for speaker discrimination (Amino and Arai, 2007; Andics *et al.*, 2007; Dellwo *et al.*, 2018; Eatock and Mason, 1994). Further, we employed a comparison between vowels and stop consonants, which are not shown to be useful speaker discriminants, within these experimental conditions. As such, it is unsurprising that their performance is poor in comparison to the vowel segments.

Following the previous studies (Narayan *et al.*, 2017; Quinto *et al.*, 2020), we also examined the effect of each word condition on RT. For *same speaker* trials, we observed significantly faster RTs in the *same word* and *phonological rhyme* conditions compared to the *frame overlap* and *different word* conditions, with no difference in RT between the *same word* and *phonological rhyme* conditions. However, we did not observe any major facilitation of the *vowel overlap* on RT at least compared to the *different word* condition, despite it showing some utility for discriminating speakers. This suggests that increased quantities of phonetic overlap may be useful for the speed of decision making, while the quality of phonetic overlap may be useful for the overall performance. For *different speaker* trials, we observed no significant differences in the length of the decision-making period. Regardless of the amount of phonetic overlap, speakers took equally as long to decide that the voices came from two different speakers.

Finally, a significantly greater bias towards *same speaker* responses was observed in all conditions containing a vowel overlap, namely *same word*, *phonological rhyme*, and *vowel overlap*. As with the performance results, we once again observed that the response pattern in the *phonological rhyme* condition was relatively unique. Specifically, the bias towards *same speaker* responses was strongest in the *phonological rhyme* condition, and this bias was significantly greater than in the *same word* or *vowel overlap* conditions. The bias towards *same speaker* responses in the *phonological rhyme* condition was also observed in Quinto *et al.* (2020). Importantly, this might also account for the overall lower performance accuracy observed in this condition.

Most interestingly here is the differing behavior in the *same word* and *phonological rhyme* conditions. Despite both technically containing a phonological rhyme overlap, the two conditions were not treated equally by listeners. Our findings provide additional support for the idea first stated in Quinto *et al.* (2020) that the *phonological rhyme* condition offers a situation where bottom-up (*phonetic*) and top-down (*phonological*) processing skills are in fact competing. While it is true that the *phonological rhyme* condition contains a level of similarity which is akin to the *same word* condition, it is also the case that this condition contains a highly salient phonological unit. Perhaps the influence of bottom-up processing for this condition corresponds to the observed performance improvements compared to conditions with less useful phonetic similarity, i.e., the *frame overlap* and *different word* conditions. In comparison, the

influence of top-down processing may correspond to the strong *same speaker* response, which interferes with the utility of the phonetic overlap. This can be further supported by previous findings that equally observed a *same speaker* response bias in other linguistically related conditions, e.g., the semantically related lexical compound condition (Quinto *et al.*, 2020). Therefore, it could be inferred that the top-down relationship (the salience of a phonological unit) between the words may be overriding the bottom-up relationship (phonetic similarity).

Albeit present, it remains unclear whether the response bias observed in our experiment and in Quinto *et al.* (2020) is a result of a perceptual belief that different speakers indeed sound more similar when producing a phonological rhyme or if it is simply an artefact of the task itself, i.e., greater listener uncertainty, corresponding to a *same speaker* response. Therefore, we conducted a second experiment to explore listeners' perception of voice similarity in each of these conditions. It is likely that a listeners' ability to discriminate voices is highly influenced by their perception of how similar the voices sound; therefore, our results will be highly complementary to those of the previous experiment. Running a second experiment with a similar but more fine-grained experimental method allows us to explore further the response biases observed in Experiment 1 while also testing whether the above results replicate for a different participant group.

III. EXPERIMENT TWO: VOICE SIMILARITY JUDGMENTS

A. Methods

1. Participants

One hundred native speakers of American English (36 female, 61 male, three non-binary) with no reported history of speech or hearing disorders were recruited using *Prolific*. Participants were selected with the same procedure as in Experiment 1: participants were first pre-screened using *Prolific* filters and later confirmed their suitability in a demographic questionnaire. Participants who completed Experiment 1 were not eligible to complete Experiment 2. The same post-experiment questionnaire to report technical issues was included in this experiment. Two participants were excluded after reporting technical issues meaning they did not complete every trial within the experiment and a further participant was excluded based on an assessment of their English language background, resulting in 97 participants in the final analysis.

2. Materials

The same stimulus set was used for Experiment 2 as for Experiment 1. Therefore, the experimental conditions were identical to Experiment 1: *same word, phonological rhyme, vowel overlap, frame overlap*, and *different word*. Participants were again presented with a unique word pairing in each trial and no speaker was heard saying the same word twice in the experiment, except from within the *same word* condition for *same speaker* pairings.

3. Procedure

Participants completed a Likert rating scale task designed and administered using the online experiment builder platform, Gorilla.sc (Anwyl-Irvine *et al.*, 2020). As in Experiment 1, the task was completed remotely, so participants completed the same headphone screening task (Woods *et al.*, 2017). The trial setup was also similar, with audio played at the exact same intervals as in Experiment 1; the first word was played 2000 ms after trial onset and the second 3700 ms after trial onset. Listeners were asked to rate the similarity of the voices on a scale from 1 to 6, with 1 being *not very similar* and 6 being *very similar*. Participants were instructed to ignore the content of the words spoken and concentrate solely on the voice itself to ensure the phonetic similarity between the words was not the focus of the judgments. Participants judged all possible speaker pairings once, including *same speaker* comparisons, in each of the five experimental conditions, resulting in 180 trials. The task took around 25 min to complete, and participants received a remuneration of £3.50 for their participation, corresponding to an average hourly rate of £9 as recommended by *Prolific*, the same payment policy as Experiment 1.

4. Analyses

Statistical analyses were conducted on z-scored Likert scale ratings to account for individual differences in scale usage. A β-logistic regression model was fit using the *glmmTMB* package (Brooks *et al.*, 2024), with main effects of *Word condition* (*same word*, *phonological rhyme*, *vowel overlap*, *frame overlap*, *different word*) and *Trial type* (*same speaker*, *different speaker*), their interactions, plus by-*Participant* and by-*Trial* random intercepts. Factors were again sum-coded, such that comparisons are interpreted in relation to the average ratings in all trials. In addition, we conducted pairwise comparisons of all conditions within *same* and *different speaker* trials using the *emmeans* packed (Lenth, 2017) to assess the significance of all within-*Trial type* marginal contrasts.

B. Results

First, we evaluated the mean ratings for same speaker pairs to assess the validity of the data and ensure that listeners did in fact judge the same speakers as sounding similar. We observed an average rating of 4.9 in same speaker pairs, and a median score of 5. As a value of 6 on the Likert scale represented a *very similar* judgment, we can confirm that same speaker pairings were indeed judged as sounding similar. As shown in Fig. 5, *different speakers* were overall judged to sound less similar to one another than the *same speakers*, resulting in lower ratings. In addition, the word condition played a role in the similarity ratings for both

https://doi.org/10.1121/10.0036562

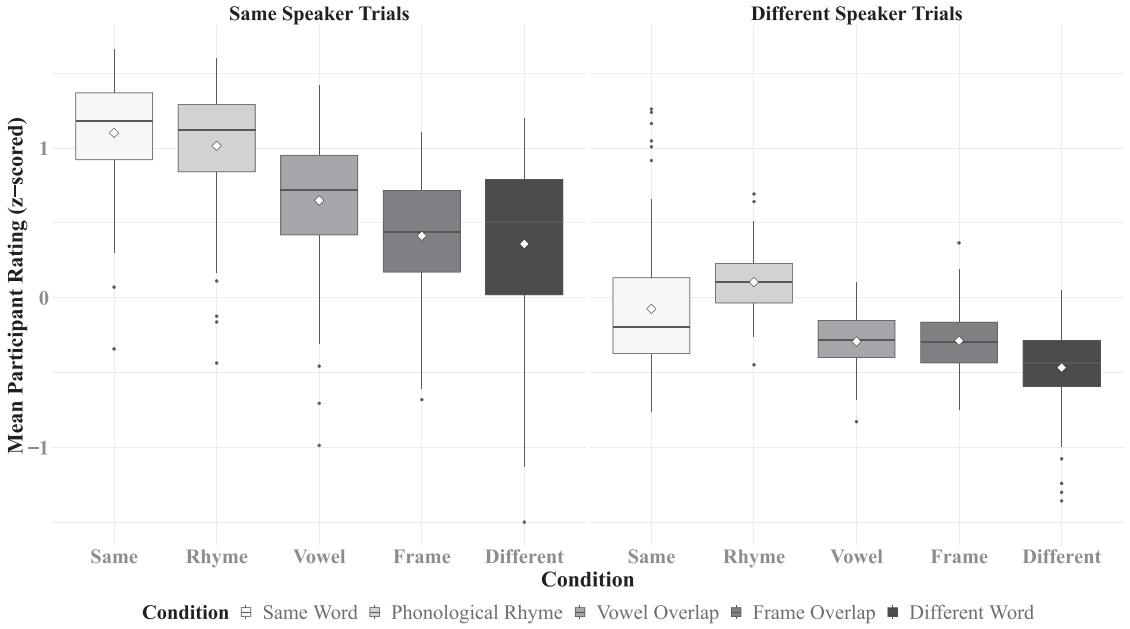

FIG. 5. *z*-scored Likert ratings of speaker similarity faceted by *Trial type* (e.g., same or different speaker) in each of the five experimental conditions. Grand means are represented by white points. Higher scores are indicative of higher perceived speaker similarity, while lower scores are indicative of lower perceived speaker similarity.

same and *different speaker* pairings, with differences observed in judgments as a result of phonetic overlap.

More specifically, the model output showed a significant two-way interaction between the effect of *Word condition* (i.e., *same word*, *phonological rhyme*, *vowel overlap*, *frame overlap*, *different word*) and *Trial type* (i.e., *same* or *different speaker* comparison). For *same speaker* trials in the *same word* and *vowel overlap* conditions, ratings were significantly higher than average (*same word*: $\beta = 0.16$, $SE = 0.022$, $p < 0.001$; *vowel overlap*: $\beta = 0.05$, $SE = 0.022$, $p = 0.024$), suggesting the same speakers were judged as sounding more similar when speaking the same word or same vowel. In comparison, for *same speaker* trials in the *different word* condition, ratings were observed to be significantly lower than average ($\beta = -0.146$, $SE = 0.022$, $p < 0.001$), indicating that same speakers were judged as sounding less similar in this condition.

As shown in Table VI, the analysis of the pairwise comparisons revealed statistically significant differences in all conditions except the *same word* and *phonological rhyme* condition in *same speaker* trials, the *frame overlap* and *different word* conditions in *same speaker* trials, and the *vowel overlap* and *frame overlap* conditions in *different speaker* trials. This suggests that the degree of phonetic similarity had a significant influence on similarity judgments for both same and different speaker pairings.

C. Discussion

Experiment 2 consisted of a 6-point Likert rating scale task in which participants were required to judge how similar two voices sounded. This experiment aimed to further validate the findings from the first experiment and to explore whether speakers were perceived as sounding more similar in specific word conditions. Overall, the findings for the rating task were highly complementary to the results from Experiment 1. If we consider the *same speaker* trials in comparison with the performance (d') results (see Sec. II C 3), we can observe the same graded effect of similarity according to the quantity of phonetic overlap in the word pairs. In

TABLE VI. Pairwise comparisons (*emmeans*) of *Trial type* and *Word condition* levels on predicted estimates of the β-logistic mixed-effects regression model for the Likert scale similarity ratings. Comparisons that reach the significance threshold (<0.05) are highlighted in bold.

Trial type	*Condition comparison*	Estimate	SE	*z*	*p*
Same speaker	*Same vs rhyme*	0.133	0.060	2.215	1.0000
	Same vs vowel	0.517	0.061	8.503	**<0.001**
	Same vs frame	0.880	0.061	14.369	**<0.001**
	Same vs different	0.928	0.612	15.070	**<0.001**
	Rhyme vs vowel	0.385	0.061	6.298	**<0.001**
	Rhyme vs frame	0.748	0.062	12.146	**<0.001**
	Rhyme vs different	0.795	0.062	12.872	**<0.001**
	Vowel vs frame	0.363	0.062	5.826	**<0.001**
	Vowel vs different	0.411	0.063	6.564	**<0.001**
	Frame vs different	0.047	0.063	0.750	1.0000
Different speaker	*Same vs rhyme*	−0.227	0.034	−6.620	**<0.001**
	Same vs vowel	0.298	0.035	8.624	**<0.001**
	Same vs frame	0.268	0.034	7.817	**<0.001**
	Same vs different	0.517	0.034	15.019	**<0.001**
	Rhyme vs vowel	0.525	0.035	15.221	**<0.001**
	Rhyme vs frame	0.495	0.034	14.447	**<0.001**
	Rhyme vs different	0.744	0.034	21.627	**<0.001**
	Vowel vs frame	−0.030	0.034	−0.859	1.0000
	Vowel vs different	0.219	0.034	6.378	**<0.001**
	Frame vs different	0.249	0.034	7.274	**<0.001**

particular, we observed that the same speakers were judged as sounding more similar in the *same word* and *phonological rhyme* conditions, while lowest similarity scores were found in the *frame overlap* and *different word* conditions, with the *vowel overlap* condition falling in-between. Given that the speakers were in fact the same in these trials, higher similarity ratings seemed to correspond with more accurate judgments, further validating the performance (d') results in Experiment 1. Specifically, these findings further emphasize the heightened perception of similarity of an overlapping vowel spoken by the same speaker, which as shown in Experiment 1, resulted in significantly higher performance.

For *different speaker* trials, the similarity ratings fall almost perfectly in line with our response bias (c) findings from Experiment 1 (see Sec. II C 4). In particular, the higher similarity ratings were found in the *phonological rhyme* condition, and these were significantly higher than the *same word* and *vowel overlap* conditions. The findings from the similarity ratings suggest that the *same speaker* response bias observed in the *phonological rhyme* condition in Experiment 1 (see also Quinto *et al*., 2020) may be accounted for by a heightened impression of speaker similarity. In addition to the exceptionality of the rhyme, the same graded effect of phonetic overlap on similarity ratings was mostly observed for *different speaker* trials. However, we did not observe a difference in the similarity ratings between the *vowel overlap* and *frame overlap* conditions for *different speaker* trials in Experiment 2, despite observing a significant difference in the *same speaker* response bias in the *vowel overlap* compared to the *frame overlap* condition in Experiment 1. Therefore, the response bias was unlikely to have arisen from a perceptual belief that speakers sounded more similar to one another when producing a word with an overlapping vowel.

In these cases, it could be that listeners were making their judgments based on the linguistic similarity of the word, rather than speaker similarity, whereby words with overlapping vowels were perceived to be more *linguistically* similar, resulting in the bias for *same speaker* responses for the *same word*, *phonological rhyme*, and *vowel overlap* condition. It is possible that there was some confusion between linguistic similarity and voice similarity, which affected the judgments of speakers, despite an explicit instruction to ignore the linguistic content of the words when making judgments. While a plausible explanation, listeners appeared to be more capable of disambiguating this linguistic similarity for the *vowel overlap* condition, and potentially offered judgments related solely to speaker similarity, at least in this ratings task.

In both the *same word* and *phonological rhyme* conditions, this increased linguistic similarity corresponded to higher similarity ratings for *different speakers* (Experiment 2) and a higher *same speaker* response bias (Experiment 1). Equally, in both the *frame overlap* and *different word* conditions, the similarity ratings were lower for *different speakers*, and the response bias was not significant, suggesting that linguistic similarity did not interfere in the task. However, the vowel overlap condition offered an intermediate overlap condition where there was enough overlap to succumb to bias in the discrimination task, but not to increase the perceptual judgment of speaker similarity in the ratings task. One possible reason for this could stem from the time pressure employed in the discrimination task, but not the ratings task. Since listeners were instructed that they would be timed out of a trial after 3 s in the discrimination tasks, snap judgments were required; however, no such time constraint was employed in the ratings task. As such, it may have given listeners a greater opportunity to disentangle linguistic similarity from speaker similarity. In comparison, the additional overlap presented in the *same word* and *phonological rhyme* conditions, may have been sufficiently substantial that it continued to influence judgments, even when listeners had slightly longer to consider them.

Despite this, it is interesting to further note that the similarity ratings did not differ significantly between the *vowel overlap* and *frame overlap* conditions. Given that different speakers were not deemed to sound more similar simply from vowel overlap alone, we can infer that vowel overlap serves as a useful speaker discriminant. It seems that vowel overlap did not perceptually influence the overall similarity judgments of different speakers, compared to having no overlapping vowel segment, but it did improve performance when discriminating speakers.

Finally, we frequently observed that when presented with words with no phonetic overlap (the *different word* condition) speakers were judged as sounding significantly less similar compared to other word conditions, in both *same* and *different* speaker trials. This suggests that any phonetic overlap in the word pair will lead to listeners perceiving different voices as more similar, potentially a result of being unable to disentangle linguistic similarity judgments from speaker similarity judgments. As such, we can infer that any degree of phonetic similarity interferes with a listener's perception of voice similarity and will lead to speakers being judged as more similar, either to themselves or another speaker.

IV. GENERAL DISCUSSION

Listeners' ability to discriminate voices from serially presented single words has previously been shown to be influenced by the linguistic content of the presented words (Narayan *et al*., 2017; Quinto *et al*., 2020). In those studies, the existence of a linguistic relationship between the words improved listener capacities for speaker discrimination, and of most relevance for this study, being presented with a word pair containing a *phonological rhyme* evoked higher speaker discrimination performance. The increased performance observed previously was argued to arise from either the phonological relationship within the word pair or the increased phonetic overlap, which rendered an easier speaker comparison. However, the extent to which either relationship played a role here had not been directly tested; equally, the exact influence of phonetic overlap for speaker

discrimination remained unknown. The experiments presented in this study aimed to bridge this gap by exploring how varying degrees of phonetic overlap influence speaker discrimination capacities, and the impact of this on overall judgments of speaker similarity.

Experiment 1 consisted of a speeded AX-discrimination task, similar to those presented in previous studies (Narayan *et al.*, 2017; Quinto *et al.*, 2020), in which speakers were asked to judge whether voices came from the *same* or two *different* speakers. Speakers were presented saying single words that varied in their degree of phonetic overlap using five different conditions: *same word, phonological rhyme*, *vowel overlap*, *frame overlap*, and *different word*. To summarize, we observed better discriminability of voice identity in conditions with overlapping vowels, i.e., the *same word*, *phonological rhyme*, and *vowel overlap* conditions. Further, we observed an almost graded effect of phonetic overlap on RT, with higher quantities of phonetic overlap resulting in, on the whole, faster RTs. However, this was not consistent in our *frame overlap* condition, in which the lack of an overlapping vowel portion slowed RTs. Finally, we observed a substantial bias for responding *same speaker* in the *phonological rhyme* condition, mirroring findings from previous studies, as well as in *same word* and *vowel overlap* conditions.

Experiment 2 aimed to validate findings from Experiment 1 using a different experimental method and to explore further perceptual biases in the different word conditions using a voice similarity rating task. For the most part, increased phonetic similarity between the words led to an increase in perceived speaker similarity. Further, our findings echoed our response bias (c) findings from Experiment 1, with regard to the *same word* and *phonological rhyme* conditions, in which *different speakers* were judged to sound more similar to one another. Finally, our findings for *same speaker* trials showed higher similarity ratings for conditions with an overlapping vowel segment, i.e., the *same word*, *phonological rhyme*, and *vowel overlap* conditions. Given that the speakers were in fact indeed the same in these trials, higher similarity scores should reflect more accurate judgments, meaning these findings complement the performance (d') findings in Experiment 1. Overall, we observed cohesive findings across the two experiments in this study that have important implications for understanding the relationship between language and speech in spoken language processing.

A. Quantity vs quality

Previous studies have suggested that increased phonetic overlap, or similarity, may be beneficial for speaker discrimination as a result of bottom-up processing advantages (Narayan *et al.*, 2017, Quinto *et al.*, 2020). Indeed, our findings support this idea, but highlight that the utility of phonetic similarity is dependent on a number of factors. Having a greater number of overlapping segments generally corresponded to increased speaker discrimination performance; however, quantity was not the only factor at play. Specifically, the vowel overlap provided in our experimental conditions was more useful than the consonant overlap for discriminating speakers, suggesting that the type or quality of the phonetic overlap is often more influential for overall performance. Specifically, having an overlapping vowel is much more useful for discriminating speakers than having multiple overlapping consonants, or more specifically, multiple overlapping stop consonants. Indeed, we can discern this in both the discrimination task, where performance was higher in conditions with overlapping vowels, and in the ratings task, where the same speakers were judged to sound more similar to themselves in conditions with overlapping vowels. Further, performance was consistently higher in the *vowel overlap* condition compared to the *frame overlap* condition, evidencing that a single overlapping vowel segment is more beneficial than multiple overlapping consonantal segments.

It should be noted that the utility of vowel overlap in these instances was not unexpected. Prior to the study, we hypothesized that overlapping vowels would provide significant performance advantages given their cross-linguistic value for speaker discrimination (for review, see Amino *et al.*, 2006). In addition, we opted to utilize only stop consonants in our experimental design given their lack of speaker discriminatory power (see Sec. II A); therefore, it was unsurprising that performance was higher for vowel overlap compared to consonantal overlap. Indeed, our findings strengthen a plethora of literature, which acknowledges the speaker discriminatory power of vowel productions. However, it remains unclear whether this advantage would still be observed had we included a multiple consonantal overlap condition, which contained more speaker-specific segments, such as fricatives or nasals, and how the duration of overlap also contributes to the informativity of the overlap (Amino and Arai, 2007; Andics *et al.*, 2007; Eatock and Mason, 1994). Further research is necessary to discern whether vowel productions retain this importance when compared with other kinds of consonants, or if the apparent utility of the vowel productions in the present study is simply driven by the lack of utility of the stop consonants.

B. Telling speakers together vs telling speakers apart

A further compelling finding is the consistency with which we observe differences in the behavior of participants in *same speaker* compared to *different speaker* trials, especially with regard to their treatment of overlapping vowels. Namely, it seems that overlapping vowel portions were more useful for determining that two productions were from the *same* rather than from *different* speakers. Further, the *phonological rhyme* condition increased perceived similarity in *different speaker* trials compared to all other conditions, but only increased similarity in *same speaker* trials to the same extent as the *same word* condition. This tendency was not unique to our study either; the previous studies have also observed differences in behavior between *same* and *different*

speaker trials. It is possible these findings can be attributed to an emerging idea in voice recognition research, which observes a difference in speakers' abilities to *tell speakers together*, i.e., decide that voice samples belong to the same speaker, and *tell speakers apart*, i.e., decide that voice samples belong to different speakers (Lavan *et al.*, 2019).

The idea originates in the field of facial recognition (Jenkins *et al.*, 2011), whereby studies have observed that, in the presence of substantial variability, participants are more likely to denote two of the same faces as different, but rarely mistake two different faces as being the same. More recently, studies in voice recognition have noted the same effect also occurs with voices as with faces. A typical experiment involves sorting samples of voices into different identities, with participants having complete freedom to create as many identity clusters as they deem necessary. Participants often judge the same speaker as multiple talkers, when presented with variable speech materials, e.g., with highly expressive voices (Lavan *et al.*, 2019) or when having to identify speakers talking vs singing (Stevenage *et al.*, 2023). However, while productions from one identity may be deemed to belong to multiple abstract identities, participants very rarely group different identities within the same cluster, leading researchers to suggest two different abilities in *telling speakers together* and *telling speakers apart*.

Recent studies have also started to consider these phenomena within discrimination tasks by examining *same* and *different speaker* trials separately. Despite limited findings, studies have found differences in confidence levels in responses between *same* and *different speaker* trials. Listeners are substantially more confident in their judgments for *same speaker* trials, compared to *different speaker* trials (Afshan *et al.*, 2022; Stevenage *et al.*, 2021). In these same studies, participants also tended to perform better in *same speaker* trials compared to *different speaker* trials (Afshan *et al.*, 2022) and showed a bias for responding *same speaker* (Stevenage *et al.*, 2021), suggesting an overall greater difficulty with *telling speakers apart* compared to *telling speakers together*.

Although we cannot attest to confidence levels, we did indeed observe higher performance in *same speaker* trials (i.e., higher correct rejection rates than hit rates) and a *same speaker* response bias (see Sec. II C). We can also hypothesize from our findings that listeners are using different strategies in *same* and *different speaker* trials, given the importance of vowel overlap for discriminating the same but not different speakers. Taking these observations together with previous studies, we can strongly infer that *speaker discrimination*, as with *speaker identification*, does not exist as one single process and strategy. Two different processes might also be at play here: *telling speakers together* vs *telling speakers apart*. We recommend in the future that studies continue to explore results from speaker discrimination tasks with separate consideration to *same* and *different speaker* trials. In addition, further consideration into how these processes differ within a discrimination task would be highly beneficial for a greater understanding of the cognitive processing behind human capacities for discriminating speakers.

C. What is so special about the phonological rhyme?

A further finding of note is the vastly different performance in the *phonological rhyme* condition compared to other conditions containing similar quantities of phonetic overlap, i.e., our *same word* and *vowel overlap* conditions. Specifically, we found significantly worse performance in the discrimination task for the *phonological rhyme* condition, likely resulting from a significant bias for participants to respond *same speaker* in this condition. This bias was then corroborated with findings from our Likert ratings task which showed that *different speakers* were judged as sounding significantly *more similar* to one another in the *phonological rhyme* condition.

Previously, the *phonological rhyme* overlap had only been considered with relation to an *unrelated* word pair, in which performance improvements were noted (Narayan *et al.*, 2017; Quinto *et al*, 2020). Indeed, this same tendency was observed when comparing the performance in the *phonological rhyme* condition to the *different word* condition. This finding had previously been tentatively attributed to the additional phonetic overlap in the word pair offering advantages for bottom-up processing and making it somewhat easier to discriminate speakers; however, comparisons with other conditions that had increased phonetic overlap had not been tested. Our findings bridge this gap and show that, while there may be an advantage of the *phonological rhyme* condition when compared to non-linguistically related word pairs, there was equally a disadvantage when compared to similar types of phonetic overlap, i.e., the *same word* condition. This suggests that, while the increased phonetic similarity in the word pair may be a driving force for the previously observed performance improvements in the *phonological rhyme* condition, it does not appear to be the only factor at play.

With regard to their *same speaker* bias findings, Quinto *et al.* (2020) speculated that the *phonological rhyme* condition may offer a situation where top-down (*phonological*) and bottom-up (*phonetic*) processing skills are competing, and the top-down relationship between the words may be overriding the bottom-up relationship. Our findings, along with what is widely known about phonological rhymes, can strongly motivate this hypothesis. In particular, phonological sensitivity is frequently regarded as a continuum, with listeners being more sensitive to and showing earlier development of certain phonological knowledge (Stanovich, 1992). In this continuum, rhyming words are situated at the lower end of the continuum, while skills, such as phoneme segmentation, sit at the higher end of the continuum. Phonological rhymes are regarded as having *shallow phonological sensitivity*, as such, from a very young age, listeners are perceptually aware of them. In fact, rhyming and alliteration are the only measures of phonological sensitivity to reliably produce above chance results with pre-school–aged children (Bryant *et al.*, 1989). Further, pre-school–aged

https://doi.org/10.1121/10.0036562

children will almost always provide rhyming words rather than semantically related words when asked to produce a word related to a target word (Cardoso-Martins and Duarte, 1994).

Seemingly, human listeners are very attuned to rhyming sounds from a very young age, and they are incredibly perceptually salient. It may be that the occurrence of a phonological rhyme temporarily distracts the listeners from the task at hand. The existence of a phonological unit in a word pair means that the utility of the phonetic overlap is substantially lessened, compared to other conditions in which equal amounts of phonetic similarity are incurred, but without this phonological salience. Reiterating the argument by Quinto *et al.* (2020) of competing processing skills, it does appear that top-down processing results in significant interference for bottom-up processing. Specifically, the advantages for bottom-up processing, which are evoked by the increased phonetic similarity in the word-pair, are overpowered by the activation of top-down processing resulting from the presence of the salient phonological unit. Presently, this can only be stated about the phonological rhyme; however, future research could consider other types of phonological relationships that are equally salient, such as alliteration, to strengthen this hypothesis.

An alternative explanation for this *same speaker* response bias may occur from real-life situations where we are more likely to encounter phonological rhymes. It is rare that we encounter phonological rhymes between speakers within a dialogue, aside from perhaps dialogues in plays. Rather, it is substantially more common for listeners to encounter rhymes that are produced by one speaker, whether this is a person reading aloud a story or poem, someone telling a joke, or when listening to songs or rap music. Therefore, it may also be the case that listeners intuitively predict that the counterpart to a rhyming pair will be spoken by the *same speaker* rather than a *different speaker*. This hypothesis can also be tentatively supported by our findings where we observe significantly faster RTs in the *phonological rhyme* condition in *same speaker* trials and equally why this condition is beneficial more so for *same speaker* compared to *different speaker* trials.

D. Avenues for future research

This study provides a foundation for understanding the relationship between linguistic structure and speech processing, but several pathways for future directions can be identified. The present study confirmed that not all phonetic similarity is weighted equally, with vowels showing markedly better utility for speaker discrimination compared to stop consonants. However, future research should other consider other phonetic segments that have been identified as carrying speaker-specific features, such as nasals (Amino and Arai, 2007; Eatock and Mason, 1994) or /s/ (Andics *et al.*, 2007; Eatock and Mason, 1994), to develop our understanding of the relationship between discriminatory power and phonetic similarity.

Finally, we have further evidence of an interference between *top-down* (phonological) and *bottom-up* (phonetic) processing and its possible implications for speaker discrimination capacities. The phonological rhyme posed challenges for speaker discrimination when compared to similar kinds of phonetic overlap, with the comparison of our *rhyme overlap* and *same word* conditions, but advantages when compared to our *different word* condition. However, a substantially larger *same speaker* response bias was observed compared to other conditions with similar phonetic similarity. Therefore, while the *phonological rhyme* may offer advantages for bottom-up processing as a result of the increased phonetic overlap, it may also offer a source of interference from top-down processing as a result of the salience of the phonological unit. Further research is needed to fully disambiguate the role of phonological vs phonetic information in speech on voice processing. This could, for example, be accomplished with a direct comparison of a consonant-vowel (CV) overlap condition compared to the vowel-consonant (VC) overlap offered in the *rhyme overlap* condition, which offers the same degree of linguistic similarity, but without the salient phonological unit. In addition, the extent to which phonological salience interferes with voice processing is currently limited to a focus on phonological rhymes; further exploration into other kinds of phonological knowledge, such as alliteration, will also be necessary to develop further this interpretation.

V. CONCLUSION

Overall, speaker discrimination capacities and voice similarity judgments are highly influenced by the linguistic content of the words spoken, and the degree of phonetic similarity in the word pairs. In particular, having increased phonetic overlap assists with discriminating speakers, but only in conditions where we also have an overlapping vowel. In addition, increased phonetic similarity, leads, to an extent, to increased perceived voice similarity, but with some critical asymmetries regarding whether the judgment is being made for *same* or *different* speakers. We also observed unique behaviors in the *phonological rhyme* condition whereby the salience of the phonological unit, i.e., the rhyme overlap, interfered with the utility of the phonetic overlap, suggesting possible interactions between top-down and bottom-up processing abilities. These findings, when combined with those from previous studies, can substantially enrich what we already know about the relationship between linguistic structure and speech processing.

ACKNOWLEDGMENTS

This work was supported by the Swiss National Science Foundation in the Project No. 185399, "The dynamics of indexical information in speech and its role in speech communication and speaker recognition." The recordings were carried out at the LiRI laboratory at the University of Zürich.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Ethics Approval

All listeners gave their informed consent before participating and received monetary compensation for their participation. The study was approved by the Ethics Committee of the Faculty of Arts and Social Sciences at the University of Zurich and performed in accordance with the Declaration of Helsinki.

DATA AVAILABILITY

Due to the data protection guidelines of the University of Zurich, the raw audio recordings produced are available for scientific purposes only from the first author upon request. Behavioral data and analyses code are made publicly available at https://osf.io/8ay2n/?view_only=93be8c53f9434c48a84356ba35b35a81.

APPENDIX

Trial word pairs for Experiments 1 and 2.

Condition	Trial word pairs			
Rhyme overlap	Peck–Tech	Bop–Cop	Keg–Beg	Bap–Tap
Rhyme overlap	Get–Debt	Tip–Kip	Dot–Pot	Kid–Bid
Rhyme overlap	Bap–Cap	Top–Cop	Tad–Bad	Pip–Dip
Rhyme overlap	Bot–Dot	Kip–Dip	Tac–Back	Pad–Tad
Rhyme overlap	Dig–Big	Gap–Bap	Bat–Cat	Pick–Tick
Rhyme overlap	Bed–Ted	Pop–Cop	Tip–Pip	Gap–Tap
Rhyme overlap	Pop–Top	Get–Pet	Cot–Bot	Kit–Pit
Rhyme overlap	Get–Bet	Pack–Tac	Bop–Top	Bit–Kit
Rhyme overlap	Tap–Cap	Dot–Cot	Kick–Tick	Bag–Tag
Frame overlap	Bat–Bot	Cat–Cot	Tap–Top	Pot–Pit
Frame overlap	Dog–Dig	Pip–Pop	Tick–Tech	Pet–Pat
Frame overlap	Get–Got	Kid–Cod	Bit–Bet	Pack–Peck
Frame overlap	Bad–Bid	Tag–Tog	Kip–Cop	Cob–Cab
Frame overlap	Big–Beg	Debt–Dot	Bap–Bop	Ted–Tad
Frame overlap	Tib–Tab	Cot–Kit	Cap–Cop	Bag–Beg
Frame overlap	Peck–Pick	Pot–Pet	Pet–Pit	Bat–Bet
Frame overlap	Kip–Cap	Tac–Tech	Bed–Bad	Pot–Pat
Frame overlap	Bot–Bit	Tap–Tip	Bot–Bet	Pack–Pick
Vowel overlap	Kit–Dip	Cop–Pot	Pop–Dog	Debt–Keg
Vowel overlap	Pad–Cab	Cod–Tog	Bit–Kip	Got–Pop
Vowel overlap	Cab–Tac	Big–Dip	Bed–Tech	Bid–Dip
Vowel overlap	Peck–Ted	Keg–Ted	Bat–Tap	Pot–Tog
Vowel overlap	Cap–Tad	Bid–Tib	Ted–Beg	Cop–Dot
Vowel overlap	Kip–Big	Cod–Top	Pat–Gap	Kick–Pit
Vowel overlap	Tag–Bat	Bot–Cob	Pat–Tac	Bid–Kick
Vowel overlap	Keg–Bed	Tip–Pick	Tech–Debt	Dot–Bop
Vowel overlap	Dig–Bid	Bag–Cab	Top–Got	Pip–Dig
Same word	Cot–Cot	Pat–Pat	Tag–Tag	Top–Top
Same word	Pick–Pick	Big–Big	Bed–Bed	Gap–Gap
Same word	Pet–Pet	Kit–Kit	Dog–Dog	Cod–Cod
Same word	Dig–Dig	Bat–Bat	Debt–Debt	Bed–Bed
Same word	Bit–Bit	Tip–Tip	Tab–Tab	Peck–Peck
Same word	Bad–Bad	Bot–Bot	Pip–Pip	Pet–Pet
Same word	Pad–Pad	Back–Back	Dog–Dog	Pad–Pad
Same word	Cat–Cat	Pot–Pot	Bag–Bag	Got–Got
Same word	Cob–Cob	Pack–Pack	Tag–Tag	Get–Get
Different word	Bet–Dig	Pit–Gap	Cab–Pit	Tac–Get
Different word	Gap–Dot	Keg–Pit	Cap–Bed	Kick–Debt
Different word	Kid–Bag	Beg–Cod	Tab–Keg	Cat–Pop
Different word	Pick–Tad	Back–Debt	Tap–Bid	Tog–Kid
Different word	Cob–Tick	Pad–Dog	Get–Pack	Cat–Peck
Different word	Pat–Kick	Keg–Bat	Got–Back	Kid–Beg
Different word	Pack–Dot	Ted–Pack	Gap–Dog	Cob–Tac
Different word	Bop–Tick	Kid–Bap	Tag–Cod	Got–Bap
Different word	Cap–Big	Tick–Pat	Cot–Pip	Bit–Tog

[1]While we acknowledge that the phonetic similarity discussed in this study, and in the previous studies, is a product of phonological similarity (i.e., overlapping segments), we will use the term *phonetic similarity* throughout this paper to describe our experimental conditions. This decision is motivated by previous studies (Narayan *et al.*, 2017; Quinto *et al.*, 2020) which distinguish between phonological and phonetic similarity to evoke discussions about bottom-up and top-down processing. Our study aims to develop these discussions but with greater attention to the role of bottom-up processing, i.e., the utility of the phonetic similarity (overlapping segments that allow access and comparisons of speaker-specific realizations) as a result of phonological similarity (the abstract category representations).

[2]We also considered the inclusion of an initial CV_ overlap condition to offer a direct comparison to the *rhyme overlap* condition which contains the same quantity of overlapping segments, but without the phonological unit. This was considered in the early stages of the experiment design but was left out of the current study due to the already lengthy experiment time for Experiment 2 (~25 min) raising concerns of participant fatigue. This condition should be considered in future research.

Abercrombie, D. (**1967**). *Elements of General Phonetics* (Edinburgh University Press, Edinburgh, UK).

Abu El Adas, S., and Levi, S. V. (**2022**). "Phonotactic and lexical factors in talker discrimination and identification," Atten. Percept. Psychophys. **84**(**5**), 1788–1804.

Afshan, A., Kreiman, J., and Alwan, A. (**2022**). "Speaker discrimination performance for 'easy' versus "hard" voices in style-matched and mismatched speech," J. Acoust. Soc. Am. **151**(**2**), 1393–1403.

Amino, K., and Arai, T. (**2007**). "Contribution of consonants and vowels to the perception of speaker identity," in *Japan-China Joint Conference of Acoustics, Acoustical Society of Japan*, June 4–6, Sendai, Japan.

Amino, K., Sugawara, T., and Arai, T. (**2006**). "Effects of the syllable structure on perceptual speaker identification," IEICE Tech. Rep **105**(**685**), 109–114.

Andics, A., McQueen, J. M., and Van Turennout, M. (**2007**). "Phonetic content influences voice discriminability," in *16th International Congress of Phonetic Sciences, ISCA*, August 6–10, Saarbrücken, Germany, pp. 1829–1832.

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., and Evershed, J. K. (**2020**). "Gorilla in our midst: An online behavioral experiment builder," Behav. Res. **52**(**1**), 388–407.

Bates, D., Maechler, M., Bolker, B., and Walker, S. (**2003**). "lme4: Linear mixed-effects models using 'Eigen' and S4 [computer software]," pp. 1.1–35.5.

Binder, J. R., Frost, J. A., Hammeke, T. A., Bellgowan, P. S., Springer, J. A., Kaufman, J. N., and Possing, E. T. (**2000**). "Human temporal lobe activation by speech and nonspeech sounds," Cereb. Cortex **10**(**5**), 512–528.

Brooks, M., Bolker, B., Kristensen, K., Maechler, M., Magnusson, A., McGillycuddy, M., Skaug, H., Nielsen, A., Berg, C., van Bentham, K., Sadat, N., Lüdecke, D., Lenth, R., O'Brien, J., Geyer, C. J., Jagan, M., Wiernik, B., and Stouffer, D. B. (**2024**). "glmmTMB: Generalized linear mixed models using Template Model Builder (1.1.9) [computer software]," available at https://cran.r-project.org/web/packages/glmmTMB/index.html (Last viewed July 2024).

Brooks, E., and Gotcher, P. (**2023**). "Pro Tools (version 2023) [computer program]," https://www.avid.com/pro-tools (Last viewed August 2023).
Bryant, P. E., Bradley, L., Maclean, M., and Crossland, J. (**1989**). "Nursery rhymes, phonological skills and reading," J. Child Lang. **16**(**2**), 407–428.
Cardoso-Martins, C., and Duarte, G. A. (**1994**). "Preschool children's ability to disregard meaning and focus attention on the phonological properties of speech: Some discrepant findings," Br. J. Dev. Psychol. **12**(**4**), 429–438.
Dellwo, V., Kathiresan, T., Pellegrino, E., He, L., Schwab, S., and Maurer, D. (**2018**). "Influences of fundamental oscillation on speaker identification in vocalic utterances by humans and computers," in *Proceedings of Interspeech 2018, ISCA*, September 2–6, Hyderabad, India, pp. 3795–3799.
Eatock, J. P., and Mason, J. S. (**1994**). "A quantitative assessment of the relative speaker discriminating properties of phonemes," in *Proceedings of ICASSP '94. IEEE International Conference on Acoustics, Speech and Signal Processing*, April 19–22, Adelaide, Australia, pp. 133–136.
Fleming, D., Giordano, B. L., Caldara, R., and Belin, P. (**2014**). "A language-familiarity effect for speaker discrimination without comprehension," Proc. Natl. Acad. Sci. U.S.A. **111**(**38**), 13795–13798.
Goggin, J. P., Thompson, C. P., Strube, G., and Simental, L. R. (**1991**). "The role of language familiarity in voice identification," Mem. Cogn. **19**(**5**), 448–458.
Goldinger, S. D., Pisoni, D. B., and Logan, J. S. (**1991**). "On the nature of talker variability effects on recall of spoken word lists," J. Exp. Psychol. Learn. Mem. Cogn. **17**(**1**), 152–162.
Goldstein, A. G., Knight, P., Bailis, K., and Conover, J. (**1981**). "Recognition memory for accented and unaccented voices," Bull. Psychon. Soc. **17**(**5**), 217–220.
Grier, J. B. (**1971**). "Nonparametric indexes for sensitivity and bias: Computing formulas," Psychol. Bull. **75**(**6**), 424–429.
Hollien, H., Majewski, W., and Doherty, E. T. (**1982**). "Perceptual identification of voices under normal, stress and disguise speaking conditions," J. Phon. **10**(**2**), 139–148.
Imai, S., Walley, A. C., and Flege, J. E. (**2005**). "Lexical frequency and neighborhood density effects on the recognition of native and Spanish-accented words by native English and Spanish listeners," J. Acoust. Soc. Am. **117**(**2**), 896–907.
Jenkins, R., White, D., Van Montfort, X., and Mike Burton, A. (**2011**). "Variability in photos of the same face," Cognition **121**(**3**), 313–323.
Kerstholt, J. H., Jansen, N. J. M., Van Amelsvoort, A. G., and Broeders, A. P. A. (**2006**). "Earwitnesses: Effects of accent, retention and telephone," Appl. Cogn. Psychol. **20**(**2**), 187–197.
Lavan, N., Burston, L. F. K., and Garrido, L. (**2019**). "How many voices did you hear? Natural variability disrupts identity perception from unfamiliar voices," Br. J. Psychol. **110**(**3**), 576–593.
Lenth, R. V. (**2017**). "emmeans: Estimated marginal means, aka least-squares means [computer software]," p. 1.10.3 (Last viewed July 2024).
Levi, S. V. (**2019**). "Methodological considerations for interpreting the Language Familiarity Effect in talker processing," WIREs Cognit. Sci. **10**(**2**), e1483.
Levi, S. V., Winters, S. J., and Pisoni, D. B. (**2011**). "Effects of cross-language voice training on speech perception: Whose familiar voices are more intelligible," J. Acoust. Soc. Am. **130**(**6**), 4053–4062.
Loakes, D. (**2004**). "Front vowels as speaker-specific: Some evidence from Australian English," in *Proceedings of the Australian International Conference on Speech Science*, December 8–10, Sydney, Australia, pp. 289–294.
Macmillan, N. A. (**2002**). "Signal detection theory," in *Stevens' Handbook of Experimental Psychology: Methodology Experimental Psychology*, edited by H. Pashler and J. Wixted (John Wiley & Sons, Hoboken, NJ), Vol. 4, pp. 43–90.
Macmillan, N. A., and Creelman, C. D. (**1990**). "Response bias: Characteristics of detection theory, threshold theory, and 'nonparametric' indexes," Psychol. Bull. **107**(**3**), 401–413.
Mullennix, J. W., and Pisoni, D. B. (**1990**). "Stimulus variability and processing dependencies in speech perception," Percept. Psychophys. **47**(**4**), 379–390.
Narayan, C. R., Mak, L., and Bialystok, E. (**2017**). "Words get in the way: Linguistic effects on talker discrimination," Cogn. Sci. **41**(**5**), 1361–1376.
Nygaard, L. C., and Pisoni, D. B. (**1998**). "Talker-specific learning in speech perception," Percept. Psychophys. **60**(**3**), 355–376.
Nygaard, L. C., Sommers, M. S., and Pisoni, D. B. (**1994**). "Speech perception as a talker-contingent process," Psychol. Sci. **5**(**1**), 42–46.
Perrachione, T. K. (**2019**). "Speaker recognition across languages," in *The Oxford Handbook of Voice Perception*, edited by S. Frühholz and P. Belin (Oxford University Press, Oxford, UK), pp. 515–538.
Perrachione, T. K., Dougherty, S. C., McLaughlin, D. E., and Lember, R. A. (**2015**). "The effects of speech perception and speech comprehension on talker identification," in *Proceedings of the 18th International Congress of Phonetic Sciences*, Glasgow, Scotland (University of Glasgow, Glasgow, UK).
Perrachione, T. K., and Wong, P. C. M. (**2007**). "Learning to recognize speakers of a non-native language: Implications for the functional organization of human auditory cortex," Neuropsychologia **45**(**8**), 1899–1910.
Philippon, A. C., Cherryman, J., Bull, R., and Vrij, A. (**2007**). "Earwitness identification performance: The effect of language, target, deliberate strategies and indirect measures," Appl. Cogn. Psychol. **21**(**4**), 539–550.
Quinto, A., Abu El Adas, S., and Levi, S. V. (**2020**). "Re-examining the effect of top-down linguistic information on speaker-voice discrimination," Cogn. Sci. **44**(**10**), e12902.
Scott, D. R., and Cutler, A. (**1984**). "Segmental phonology and the perception of syntactic structure," J. Verbal Learn. Verbal Behavior **23**(**4**), 450–466.
Souza, P., Gehani, N., Wright, R., and McCloy, D. (**2013**). "The advantage of knowing the talker," J. Am. Acad. Audiol. **24**, 689–700.
Stanislaw, H., and Todorov, N. (**1999**). "Calculation of signal detection theory measures," Behav. Res. Methods, Instrum. Comput. **31**(1), 137–149.
Stanovich, K. E. (**1992**). "Speculations on the causes and consequences of individual differences in early reading acquisition," in *Reading Acquisition*, edited by R. B. Gough, L. C. Ehri, and R. Treiman (Erlbaum, Mahwah, NJ), pp. 307–342.
Stevenage, S. V., Clarke, G., and McNeill, A. (**2012**). "The 'other-accent' effect in voice recognition," J. Cogn. Psychol. **24**(**6**), 647–653.
Stevenage, S. V., Singh, L., and Dixey, P. (**2023**). "The curious case of impersonators and singers: Telling voices apart and telling voices together under naturally challenging listening conditions," Brain Sci. **13**(**2**), 358.
Stevenage, S. V., Tomlin, R., Neil, G. J., and Symons, A. E. (**2021**). "May I speak freely? The difficulty in vocal identity processing across free and scripted speech," J. Nonverbal Behav. **45**(**1**), 149–163.
Sumner, M., and Samuel, A. G. (**2009**). "The effect of experience on the perception and representation of dialect variants," J. Mem. Lang. **60**(**4**), 487–501.
Thompson, C. P. (**1987**). "A language effect in voice identification," Appl. Cogn. Psychol. **1**(**2**), 121–131.
Vanags, T., Carroll, M., and Perfect, T. J. (**2005**). "Verbal overshadowing: A sound theory in voice recognition?," Appl. Cogn. Psychol. **19**(**9**), 1127–1144.
Woods, K. J. P., Siegel, M. H., Traer, J., and McDermott, J. H. (**2017**). "Headphone screening to facilitate web-based auditory experiments," Atten. Percept. Psychophys. **79**(**7**), 2064–2072.
Xie, X., and Myers, E. B. (**2015**). "General language ability predicts talker identification," Proc. Ann. Mtg. Cognit. Sci. **37**, 2697–2702.
Zarate, J. M., Tian, X., Woods, K. J. P., and Poeppel, D. (**2015**). "Multiple levels of linguistic and paralinguistic featres contribute to voice recognition," Sci. Rep. **5**, 11475.