Homophone discrimination based on prior exposure

Chelsea Sanker*

Stanford University
Margaret Jacks Hall, Building 460
Stanford, CA 94305, USA

*Corresponding author
Email address: sanker@stanford.edu (Chelsea Sanker)

Homophone discrimination based on prior exposure

Chelsea Sanker*

Stanford University
Margaret Jacks Hall, Building 460
Stanford, CA 94305, USA

Abstract

This article presents three studies testing the potential role of word-specific acoustic details in perception, based on how several factors impact listeners' accuracy in identifying homophones. Experiment 1 tests how prior exposure to particular homophones said by the same talker impacts identifications; listeners could discriminate between homophone mates with above chance accuracy after exposure to disambiguated tokens of these words produced by the talker, but not when prior exposure did not include the test words. Experiment 2 tests whether having the same talker in exposure and testing is crucial; accuracy is above chance even when the prior exposure to the homophone mates was from a different talker. Experiment 3 tests whether accuracy in homophone identification might be driven by broad associations between meaning and acoustic form rather than the details of particular words; there is no difference between exposure to the particular homophone mates and exposure to semantically similar words. Just having strong positive or negative emotional valence seems to result in higher accuracy for how homophone mates are identified. These results suggest that listeners can make use of semantically-driven acoustic differences between homophone mates when recent exposure makes these details salient or when the form-meaning associations are already strong. This link to acoustic details can be explained via associations with broad aspects of meaning, rather than depending on word-specific phonetic representations.

Keywords: homophones, perception, word-specific phonetics, talker-specific learning, emotional valence

1. Introduction

The realization of particular sounds includes phonetic details that must be learned, as reflected in differences across languages in the prototypical realization of sounds and the boundaries between them (e.g. Lieberman, 1970; Keating, 1985) and experiments that shift those realizations (e.g. Kraljic & Samuel, 2006; Nielsen, 2011). Under Exemplar Theory, particular words could also have their own distinct phonetic details (Goldinger, 1998; Pierrehumbert,

*Corresponding author
Email address: sanker@stanford.edu (Chelsea Sanker)

2002). Some other approaches similarly associate acoustic details with lexical meanings, though the particular approach is different (e.g. Baayen et al., 2019; Arnold et al., 2017). Homophones provide a test for whether words can have distinct phonetic details despite being phonologically the same. Evidence from production suggests that homophones can differ in their phonetic details (Gahl, 2008; Guion, 1995). However, listeners generally cannot discriminate between homophone mates (Bond, 1973; Sanker, 2019). At-chance accuracy in discrimination tasks might indicate that differences in production are simply an effect of the production context, rather than any differences in the representation. However, it is also possible that perceptual discrimination could reach above-chance accuracy in facilitative conditions that were not present in previous homophone discrimination tasks.

This article presents results from three auditory word identification tasks, testing how accuracy in homophone identification is impacted by prior exposure to those homophones as compared to exposure to unrelated non-homophones, also comparing exposure to a different talker saying the same words and the same talker saying different but semantically similar words. The results suggest that listeners can make use of semantically-driven acoustic differences between homophone mates when recent exposure makes these details salient or when the form-meaning associations are already strong.

1.1. Phonetic details in the phonological representation

Many phonetic details must be learned, rather than being the automatic result of phonological characteristics. These details are apparent from phonetic differences across languages; even when the same contrastive sounds occur in two languages, they can have different prototypical realizations and a different boundary between them (Lieberman, 1970; Keating, 1985). The existence of learned phonetic targets is also reflected in experiments that elicit shifts in those targets. Exposure to altered acoustic characteristics of a sound can change listeners' expectations about pronunciation of that sound and the perceptual boundaries with neighboring categories (e.g. Kraljic & Samuel, 2006). Exposure to altered acoustic characteristics of a sound can also produce a shift in a listener's subsequent pronunciation of that sound, as is illustrated in convergence studies (e.g. Nielsen, 2011). The specific acoustic targets for particular sounds are also reflected in talkers' compensatory response to altered auditory feedback. When talkers' acoustic output is altered in real time, they shift their subsequent productions in compensation; for example, if F1 for a vowel is lowered, participants will raise their F1 (Houde & Jordan, 1998; Rochet-Capellan & Ostry, 2011).

Language-specific acoustic targets are also reflected in how listeners weight perceptual cues that they use for phonological categorization decisions (Dmitrieva, 2019; Lee et al., 2013) and ratings of naturalness (Kong et al., 2012). Prototypicality of a stimulus influences how quickly listeners recognize it; listeners are faster to identify words produced for a real listener than hyperarticulated words produced under instructions to speak clearly or speak for someone who has hearing loss (Scarborough & Zellou, 2013). Phonetic prototypicality also influences degree of lexical activation as reflected in priming (Andruski et al., 1994).

1.2. Word-specific phonetic details

If people have acoustically detailed representations at the phonological level, they might also have acoustically detailed lexical representations. Exemplar models propose detailed representations of this sort; exemplar clouds are linked across memories of particular sounds as well as memories of particular words (Goldinger, 1998; Pierrehumbert, 2002). Another potential approach to word-specific phonetic details is with models that directly associate lexical meanings with acoustic characteristics of speech stimuli, rather than going through an intermediate level of phonemic breakdown (Baayen et al., 2019; Arnold et al., 2017). In either type of model, one of the important considerations is not just being able to account for word-specific acoustic patterns, but also being able to account for phonological patterns across words, including the behavior of homophone mates as compared to words that are phonologically distinct.

Experimental evidence provides some support for acoustically detailed memories of recent speech, though these memories do not necessarily indicate that these details ever enter word-specific phonological representations. Listeners accurately recognize which particular tokens have been presented previously (e.g. Hintzman et al., 1972), and make more accurate phonological decisions when the same particular tokens had been heard previously (e.g. Chiu, 2000). Listeners are also more likely to identify items as having appeared before if they are similar though not identical to previously presented tokens (e.g. Church & Schacter, 1994). Memories of particular tokens are not limited to characteristics of the speech signal; listeners are more accurate in identifying a word under adverse listening condition if it is presented with the same background noise (e.g. phone ringing, dog barking) that was present during prior exposure to that word (Pufahl & Samuel, 2014).

One line of evidence sometimes used in support of word-specific phonetic details is the relationship between lexical frequency and phonetic convergence; several studies have found more convergence in lower frequency words (Goldinger, 1998; Babel, 2010; Nielsen, 2011). This effect is usually explained within Exemplar Theory, as laid out by Goldinger (1998): For lower frequency words, the exemplars from the task are a large proportion of the overall cloud of weighted exemplars, producing strong convergence. For higher frequency words, there are more pre-existing recent exemplars, so the exemplars from the task have a smaller impact in shifting that robust representation. However, a relationship between convergence and lexical frequency is an indirect source of evidence for word-specific details; the relationship might have a different explanation. Sanker (2021) demonstrates that lexical frequency is a predictor of increased similarity between talkers after a simple reading task, in which participants did not hear any input from another talker. Because this pattern of greater increased similarity among lower frequency words can be produced simply by repetition effects, it does not need to be explained by word-specific phonetic details.

Higher frequency words are more reduced than lower frequency words; perception is similarly predicted by lexical frequency. For example, American English listeners have higher accuracy identifying a high-frequency word with a flapped /t/ than a lower-frequency word with a flapped /t/ (Ranbom & Connine, 2007). While these results indicate that the particular outcomes of reduction are part of a listener's phonology, the role of lexical frequency in setting expectations for reduction does not necessarily need to be based on word-specific phonetic

representations and might instead reflect general expectations about lexical frequency as a predictor of flapping. Tang & Shaw (2021) demonstrate that effects of informativity on production of duration, F0, and intensity in Mandarin are apparent for each word even when the environment of each particular token is accounted for, which might suggest that the acoustic effects of predictability have become part of the lexical representation of particular words, rather than existing only as an effect of context. Seyfarth (2014) finds similar effects on duration in English. However, it is possible that these effects could be explained by informativity influencing ease of lexical retrieval, with the speed and strength of activation producing acoustic differences, rather than the representations including distinct acoustic targets (Gahl et al., 2012; Kahn & Arnold, 2012).

If listeners have word-specific phonetic representations, it should be possible to shift the acoustic targets in different ways for different words. A possible example of this comes from work on altered auditory feedback, in which manipulation that differs by word can produce word-specific articulatory shifts. Rochet-Capellan & Ostry (2011) demonstrate that altering subjects' auditory feedback by increasing F1 in "bed" and decreasing F1 for "head" resulted in word-specific compensatory shifts: decreased F1 in "bed" and increased F1 in "head." These results might depend on having a very small number of words with a very large number of repetitions; listeners only produced and heard the altered feedback for three words during the task ("bed", "head", "ted"), each appearing over 100 times. Other types of studies usually use a larger number of words with far fewer repetitions of each one. Sanker (2021) tests whether different convergence can be elicited for words manipulated in opposite directions, either in vowel duration or in F2, and finds no evidence that such word-specific convergence occurs.

1.3. Phonetic details in homophones

If word-specific phonetic details exist, homophones are a key part of the lexicon where it should be possible to clearly distinguish these word-specific details from effects of processes conditioned by the phonological environment. It has been demonstrated that homophone mates can exhibit significant differences in their acoustic details as they are produced in natural speech, based on factors like lexical frequency (Gahl, 2008; Guion, 1995), part of speech (Lohmann, 2017; Conwell, 2017), morphological complexity (Walsh & Parker, 1983; Plag et al., 2017; Seyfarth et al., 2018), and orthographic length (Warner et al., 2004). These patterns of differences between homophone mates could indicate that they have distinct phonetic details in their representations. However, many of the differences in production can be attributed to factors such as position in the sentence (Conwell, 2017) and predictability in context (Jurafsky et al., 2002). The differences between homophone mates are reduced when they are produced in frame sentences or in isolation (Guion, 1995; Sanker, 2019), which is consistent with the acoustic differences being largely an effect of context, rather than being part of the representation.

Perception results provide no clear evidence that listeners have distinct phonetic details in the representation of homophone mate pairs. Bond (1973) found at chance accuracy for identifications of homophone mates. While Sanker (2019) found accuracy above chance for identification of homophone mates in some conditions, the effect was very small. Slightly

above chance accuracy might be explained by expectations based on systematic influences like frequency and morphological complexity, without listeners necessarily having word-specific phonetic representations. Consistent with this explanation, Bond (1973) found that the duration of the vowel in the stimulus influences decisions between homophone mates differing in morphological complexity, even though the selection preferences did not result in above-chance accuracy. If two homophone mates differ in their typical duration or other acoustic characteristics based on factors like lexical frequency or morphological complexity, listeners might distinguish between those words with above chance accuracy based on expectations about how lexical frequency or morphological complexity relate to duration, though the effect is likely to be small.

Listeners may similarly make use of expectations about how different polysemous uses of a word will be pronounced based on pragmatic factors influencing the prosody. Martinuzzi & Schertz (2021) demonstrate that listeners can distinguish between the apology vs. attention-seeking functions of "sorry", using several prosodic cues. While this result could be interpreted as these two functions including distinct phonetic details for duration and intonational contour, listeners may have distinct phonetic knowledge for pragmatic prosodic factors and phonological factors, and use both when processing incoming speech input. The high accuracy that they found for discriminating between functions of "sorry" might be related to the fact that both functions of this word tend to occur as prosodically isolated units; most word identification tasks use stimuli that are words in isolation (e.g. Bond, 1973; Sanker, 2019), even though the words are normal lexical items that usually appear within sentences in natural speech.

1.4. Phonetic characteristics associated with meaning

Even if listeners do not have word-specific representations that include phonetic detail, broad relationships between meaning and phonetic details may influence perception of phonologically ambiguous items. Meaning is a factor in how talkers produce words. Acoustic characteristics similarly influence how listeners evaluate meaning; listeners are influenced by associations between acoustic form and emotional valence, size, and other characteristics.

Several studies have found acoustic differences based on the emotional valence of the word (e.g. Nygaard et al., 2009) or the emotion being conveyed by the talker (e.g. Nygaard & Lunders, 2002). In a nonce word production task in which words were assigned with positive, negative, or neutral meanings, Nygaard et al. (2009) found that participants produce happy words with higher F0, more variation in F0, higher amplitude, and shorter duration. In a subsequent listening task using these recordings, listeners were more likely to select the meaning that aligned with the meaning assigned to the word when it was produced. Emotional prosody also influences identification of homophones, as is demonstrated by Nygaard & Lunders (2002). They made recordings of a word list produced by actors portraying happy, sad, and neutral emotion; the emotional conditions influenced several acoustic characteristics, including duration, F0 mean, and F0 range. When listeners were asked to identify homophones recorded in these conditions, they were more likely to select the meaning that matched the tone of voice, e.g. selecting *die* in the sad condition and *dye* in the neutral condition.

Work on sound symbolism also demonstrates that size and shape of a referent are associated with acoustic characteristics. Most work on sound symbolism looks across phonological categories, but there is also work that breaks down the effects further. Knoeferle et al. (2017) separate out the phonetic characteristics of each sound that seems to contribute to sound-symbolic associations; longer vowel duration and more compact vowel spaces increase the size that nonce words are rated as indicating. Listeners also learn the meaning of nonce words more quickly when the form of the object aligns with commonly demonstrated associations of the component phonemes, e.g. high unrounded vowel as pointy object, lower round vowel as round object (Kovic et al., 2010). Non-contrastive duration differences also influence expectations. Controlling for vowel height and using gradient duration manipulations, Rojczyk (2011) found that listeners were more likely to assign a nonce word the meaning 'big' when the word had a longer vowel duration. Nonce words are also produced with longer duration when associated with meanings of 'big' rather than 'small' (Nygaard et al., 2009) and lengthening can be used iconically to intensify meaning for existing words (Guerrini, 2020). English speakers also have a higher average F0 for words with small referents than words with large referents (Perlman et al., 2015).

1.5. Talker-specific learning

There is variation in pronunciation across talkers, due to physical differences, dialectal differences, and idiosyncratic habits. Exposure to a particular talker can thus improve familiarity with that talker's phonological system and other characteristics of that individual's speech. Although listeners are substantially above chance accuracy in identifying words and sounds from different talkers and even in the first token produced by a particular talker, accuracy improves with more exposure to a talker (Verbrugge et al., 1976). Word identification is faster and more accurate when the talker is the same across trials or the same in both training and testing (Mullennix et al., 1989; Nygaard et al., 1994), and same-different decisions are slower when the paired items come from different talkers than when they come from the same talker (Cole et al., 1974).

In addition to quickly adapting to natural differences between talkers, listeners can learn artificially manipulated patterns of how particular voices realize particular sounds (e.g. Kraljic & Samuel, 2007). Familiarization with a particular talker might also involve learning other aspects of speech behavior, such as variation in what emotional valence a word has for that talker.

Learning of particular talkers' voices is also reflected in subsequent recognition of particular tokens and preferential looking in eye-tracking studies. Listeners recognize previously presented words more quickly and more accurately when repeated in the same voice than when repeated in a different voice (Goldinger, 1996; Palmeri et al., 1993). The effects of familiarity with the voice are smaller but still present when the specific tokens are distinct (Goh, 2005). Listeners also spend less time looking at competitor images when previous exposure to the target word and competitor word had been in different voices and are presented again in the same voice than when previous exposure had presented both words in the same voice (Creel et al., 2008). After training on nonce words presented with accompanying images, when

listeners hear the nonce words again in the same voice, they spend more time looking at the images originally presented with that voice saying that word (Kapnoula & Samuel, 2019).

1.6. These Studies

This paper presents three studies which look for word-specific acoustic details using homophone identification tasks preceded by different types of exposure. In Experiment 1, the exposure either included the stimulus voice producing the particular homophone mates that would appear during homophone identification or only included the stimulus voice producing unrelated words. In Experiment 2, the exposure either included the same voice producing the homophone mates that would appear during the homophone identification test, a different voice producing these words, or a different voice producing only unrelated words. In Experiment 3, the exposure either included the stimulus voice producing the particular homophone mates that would appear during homophone identification, the stimulus voice producing words that are semantically similar to the target homophones, or the stimulus voice only producing unrelated words.

2. Experiment 1

In Experiment 1, listeners completed a word-identification task with homophones produced by the same talker, in which the response options were homophone mates. The primary variable in the exposure phase was whether listeners heard the stimulus voice producing the particular homophone mates that would appear during homophone identification or only unrelated words. The secondary variable examined was the production environment that the test stimuli were extracted from: a frame sentence or meaningful sentences.

2.1. Methods and Materials

Stimuli were made from recordings of one female American English speaker reading monosyllabic English words, elicited in randomized order with PsychoPy (Peirce, 2007) and recorded in a quiet room with a stand-mounted Blue Yeti microphone in the Audacity software program and digitized at a 44.1 kHz sampling rate with 16-bit quantization. The talker was not naive to the purpose of the experiment; this will be considered further in Experiment 2.

The target words included 20 homophone mate pairs, selected to be similar in frequency as much as possible, to reduce the possibility that listeners might identify homophones with above chance accuracy based on general expectations of frequency-conditioned reduction, rather than word-specific knowledge. All homophone mates were orthographically distinct, e.g. *sight*, *site*. There were two conditions for training words, each containing 80 items (40 pairs of response options), as described below. A list of all words can be found in the appendix.

The words were recorded in two environments: a frame sentence, *The word is* ___, and naturalistic sentences, e.g. *We drove to the site*. The target word was always the last word of the sentence.

Participants were 128 native speakers of American English (mean age 38.4; 64 male, 63 female, 1 nonbinary) with no reported speech or hearing disorders. 15 participants were excluded and replaced based on having accuracy below 80% for identifications of training items or taking longer than 30 minutes to complete the study. The training items were decisions between phonologically distinct English words, which should be unambiguous; low accuracy on these items suggests low attention or poor audio conditions. The median task time was slightly under 10 minutes; long task times suggest potential distractions or interruptions, which could interfere with the relationship between the conditions in the training phase and subsequent performance in the testing phase.

The study was run online, with participants recruited and paid through the Amazon Mechanical Turk system and the experiment presented through Qualtrics.[1]

Participants were instructed that they would hear English words and identify each one as matching one of two associated response options. The stimulus items were presented as a list; listeners clicked on an audio player icon to hear each stimulus. Responses were given by clicking on one of the written words given under the icon for the stimulus. Within a block, the order of items was randomized. The order of the two response options was balanced across participants.

Listeners were allowed to listen to the recordings multiple times if they chose to. The focus of this study is on whether listeners could make use of acoustic details in the stimulus, so ensuring that participants were able to hear the stimulus to their satisfaction was prioritized; given the lack of control over the listening environment in online studies, this reduces the risk of a participant failing to hear an item because of irregular environmental noise or other distractors. Variation in the number of repetitions might have an effect on responses, though it is not clear that this is expected based on previous work; Sanker (2019) found no effect of block on accuracy of homophone identifications for the same stimuli presented in three blocks.

There were two blocks in this experiment: a training block and a testing block. All stimuli were produced by the same individual. However, the test tokens were always different from the tokens heard during training, even when the same word appeared in both phases.

First, listeners completed the training block. They heard a set of 80 items presented individually, all monosyllabic English words which they identified as matching one of two response

[1]There are a range of possible sources of variation across participants, some of which are specific to online studies (e.g. different devices, different listening environment), and some of which are also present for in-person studies (e.g. differences in hearing, differences in attention). Some studies include tests to constrain possible sources of variation, often focusing on headphone use. Woods et al. (2017) investigated how to screen for headphone usage, which can improve performance in some auditory tasks, though they also note that there is additional variation from other sources. The experiments presented in this paper use a relatively large number of participants, which reduces the likelihood that differences across conditions will arise by chance due to a disproportionately large number of listeners in one condition being better or worse at the task due to their listening setup or characteristics like hearing or attention. High accuracy in the training trials (97%-98% across the three experiments) indicates that all listeners were able to hear the stimuli clearly. By-participant intercepts are also included in the models to handle variation in overall accuracy by participant.

options that differed only in the vowel, e.g. hear *sight* and select either "sight" or "seat" as the written word matching the recording. Both items of each pair of response options were included as stimuli, and each training stimulus appeared only once. For example, there would be a trial with a *sight* stimulus and also a trial with a *seat* stimulus, with the same response options "sight" and "seat". There would be another trial in which listeners hear *site* and select either "site" or "set" as the written word matching the recording, and so on. The paired item for each homophone mate was balanced across listeners, e.g. one listener would have "sight" vs "seat" decisions and "site" vs "set" decisions, while another listener would have "site" vs "seat" decisions and "sight" vs "set" decisions.

There were two different conditions for the exposure stimuli in this training phase. (1) In the homophone-exposure training condition, the training items included all of the words that would subsequently appear in the homophone identification task (e.g. *sight*, *seat* providing exposure to *sight*); because the written response options in each training trial included only one of the homophone mates, the meaning of each homophone stimulus is disambiguated by the response options. (2) In the no-homophone-exposure condition, the training items only included only non-homophones (e.g. *pipe*, *peep*), so none of the words in the homophone identification test phase were presented during the training phase for listeners in this condition. Words for the latter condition were selected to be phonologically similar to the words in the first condition.

Second, listeners completed the testing block. They heard a set of 40 items presented individually, all monosyllabic English words which they identified as matching one of two orthographically distinct homophone mates, e.g. hear *sight* and select either "sight" or "site" as the written word matching the recording. Listeners heard both items of each pair of response options during the task.

There were two conditions for the environments that the stimuli were extracted from. In one condition, the test stimuli had been extracted from a frame sentence. In the other condition, the test stimuli had been extracted from naturalistic sentences. In both cases, the training stimuli came from the opposite environment, i.e. when the test items came from the frame the training items came from the naturalistic sentences, and when the test items came from naturalistic sentences the training items came from the frame. This was done to ensure that the only details that listeners might be using to distinguish between homophone mates had to be due to the lexical items themselves, rather than the environments they occur in.

Statistical results are from logistic mixed effects models, using the lme4 package (Bates et al., 2015) in R (R Core Team, 2022); p-values were calculated by the lmerTest package (Kuznetsova et al., 2015), which uses the Satterthwaite method for approximating degrees of freedom. The details of each model appear below to introduce each model before presenting a summary of its results.

2.2. Hypotheses and predictions

There are three main competing hypotheses for whether listeners will be above chance accuracy based on the exposure condition (whether the exposure phase included the homophone mates or not).

Hypothesis 1a: Listeners may have pre-existing expectations about how homophone mates differ acoustically and will be above chance accuracy in both conditions. Accuracy might also depend on familiarity with the talker; familiarity with the talker's voice may help set expectations about systematic patterns that are present across words. Listeners had prior exposure to the talker in both exposure conditions in this experiment, so an effect of exposure to the talker also predicts above-chance accuracy in both conditions.

Hypothesis 1b: Listeners may identify homophones with above chance accuracy only when they have heard those words produced in the training phase of the experiment. Higher accuracy in this condition could either indicate that listeners are becoming attuned to how the particular talker says these words, or that the recent exposure has drawn listeners' attention to the differences in meaning between these homophone mates and the acoustic characteristics that tend to be associated with each meaning.

Hypothesis 1c: There might be no difference between conditions, with neither condition producing above chance accuracy. This might suggest that listeners do not associate distinct acoustic details with homophone mates or are unable to draw on those associations under the conditions of the task.

There are two competing hypotheses for the possible effects of the original production environment that testing stimuli were extracted from.

Hypothesis 2a: Accuracy might be above chance only when identifying stimuli produced in meaningful sentences, given that acoustic differences between homophone mates are most apparent in words produced in meaningful sentences (Guion, 1995; Sanker, 2019). If there is an effect of production context of the testing stimuli, it is unlikely to interact with the exposure condition, because the production context always differed between training and testing.

Hypothesis 2b: The production context might not influence accuracy. If listeners are sensitive to the word-specific acoustic details produced in meaningful sentences but need exposure to those words to make those details salient, then the training items would not improve accuracy because listeners did not have the same sentential contexts both for the training stimuli and the testing stimuli. This result would also be consistent with listeners not associating distinct acoustic details with homophone mates.

2.3. Results

Accuracy in the training phase was 97.0%; recall that the decisions in the training phase were all choices between unambiguously phonologically distinct words (e.g. *pipe*, *peep*) as matching the auditory stimulus. Results are reported only for the testing phase, in which listeners made decisions between homophone mates (e.g. *sight*, *site*) as matching the auditory stimulus.

Table 1 presents the summary of a mixed effects logistic regression model for accuracy in Experiment 1.[2] The fixed effects were exposure condition (Homophone Exposure, No Ho-

[2] $glmer(accuracy{\sim}Condition * Context + (1|ParticipantID) + (1|pair), data =$

mophone Exposure); original production context of the stimulus items used in homophone identification (Meaningful Sentences, Frame Sentence); and the interaction between condition and context. There were random intercepts for participant and for homophone pair.[3]

	Estimate	Std. Error	z value	p value
(Intercept)	0.167	0.0612	2.72	**0.00652**
Condition NoHomExposure	-0.167	0.0803	-2.08	**0.038**
Context Sentence	-0.041	0.0803	-0.51	0.61
Condition NoHomExposure * Context Sentence	0.0943	0.113	0.831	0.406

Table 1: Logistic regression model for accuracy, Experiment 1. *Intercept: Condition = HomophoneExposure, Context = FrameSentence*

As seen in the intercept, accuracy was significantly above chance (53.6%) when listeners had prior exposure to these particular homophones as said by this talker and when the context was the frame sentence (the latter aspect of the intercept is not crucial, as is discussed below; accuracy in this condition is also significantly above chance if Context is excluded as a factor). This is not merely an effect of familiarization with the talker. Participants in both conditions heard the same number of tokens produced by the talker; in the No Homophone Exposure condition, the training items were phonologically similar but semantically unrelated words. Figure 1 illustrates the overall accuracy in each condition.

Accuracy was significantly lower when the training phase did not include prior exposure to these homophones (50.7%). That is, accuracy was lower in the condition in which the training only included unrelated non-homophones than in the condition in which the training included the homophones that would also appear in testing.

There was no significant effect of the production context of the stimuli, nor an interaction between production context and exposure condition. As described above, the context refers to the original production context of the test items; all words were extracted from their original contexts and presented in isolation. Recall also that the production context always differed between the training items and the test items; the model specifies the production context of the test items used in homophone identification.

Although the model included random intercepts by participant and by homophone mate pair, there was no clear evidence that accuracy depended on either of these factors. This model does not have a significantly better fit than a model without the by-participant intercept (χ^2 = 0.0353, df = 1, $p = 0.851$) and is only marginally better than a model without the by-pair intercept ($\chi^2 = 3.02$, df = 1, $p = 0.082$).

Given that listeners' accuracy in identifying homophones is above chance after exposure to the homophones in disambiguating contexts during training, one question that arises is

$HomophoneData1[which(HomophoneData1\$Task == \text{“}testing\text{”}),], family = binomial)$

[3]Random slopes by participant are impossible because of the between-participants design of the conditions. Including a random slope for condition by pair did not change the results and did not significantly improve the model ($\chi^2 = 1.37$, df = 2, $p = 0.504$); this slope was also strongly correlated with the random intercept by pair. For these reasons, no random slopes were included.

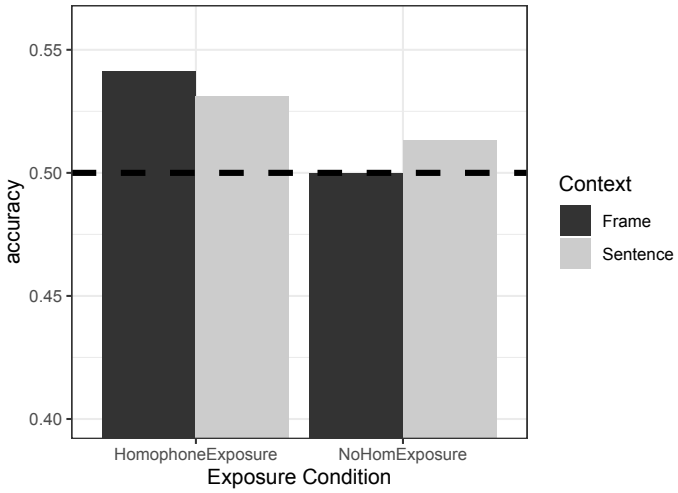

Figure 1: Overall accuracy in each condition for Experiment 1. 50% accuracy is marked with a dashed line.

what cues listeners were using in making these decisions. There is substantial variation in the acoustic characteristics of the stimuli based on the two environments that words were extracted from; however, some characteristics exhibit larger differences between homophone mates (e.g. *sun, son*) than between two instances of the same word (e.g. *sun, sun*). Table 2 presents a summary of differences between items in several acoustic characteristics that have been previously identified as differing between homophone mates.

	Vowel Duration	F0 mean	F0 range	Intensity	Spectral Tilt
Same Word	31.2 ms	24.6 Hz	50.1 Hz	2.6 dB	6.9 dB
Homophone Mates	38.8 ms	28.7 Hz	48.1 Hz	2.5 dB	7.0 dB

Table 2: Acoustic differences between the same word from different contexts (e.g. *sun* from the frame sentence vs *sun* from a naturalistic sentence) and between homophone mates from different contexts (e.g. *sun* from the frame sentence vs *son* from a naturalistic sentence)

If accuracy results from listeners being familiarized with the training stimuli, it might follow that testing stimuli which are more similar to the training stimulus for that word will be identified more accurately. Table 3 presents the summary of a mixed effects logistic regression model for accuracy that includes acoustic predictors based on difference between characteristics of the training stimuli and the testing stimuli.[4] Only the significant predictors are included. The fixed effects were exposure condition (Homophone Exposure, No Homophone Exposure); original production context of the stimulus items used in homophone identification (Meaningful Sentences, Frame Sentence); difference in intensity; and the interaction between condition and context. There were random intercepts for participant and for homophone pair. The continuous variables were centered.

The only measure of acoustic distance that was a significant predictor of accuracy was intensity; greater difference in intensity from the corresponding training item predicted lower

[4] $glmer(accuracy \sim Condition * Context + IntensityDiff + (1|ParticipantID) + (1|pair), data = HomophoneData1[which(HomophoneData1\$Task == \text{“}testing\text{”}),], family = binomial)$

	Estimate	Std. Error	z value	p value
(Intercept)	0.167	0.0682	2.45	**0.0142**
Condition NoHomExposure	-0.167	0.0806	-2.07	**0.0381**
Context Sentence	-0.0411	0.0807	-0.51	0.61
Intensity Difference	-0.0526	0.0181	-2.92	**0.00356**
Condition NoHomExposure * Context Sentence	0.0947	0.114	0.831	0.406

Table 3: Logistic regression model for accuracy, including acoustic differences from the training stimuli as predictors, Experiment 1. *Intercept: Condition = HomophoneExposure, Context = FrameSentence*

accuracy. The lack of clear effects of difference from the training stimulus in other acoustic characteristics may suggest that the effects of training are not simply about familiarization with the particular stimuli and extending that to new tokens, but instead depend on drawing listeners' attention to acoustic details that are already associated with the word.

The main results of this experiment are consistent with Hypothesis 1b in the effect of exposure condition: Listeners are above chance accuracy in identifications of homophones only when they have been exposed to those homophones during the training phase. This might indicate that listeners are becoming attuned to how the particular talker says these words, or that the recent exposure has drawn their attention to the differences in meaning between these homophone mates and the acoustic characteristics that tend to be associated with each meaning. The results are consistent with Hypothesis 2b in the effect of production context: There is no evidence for an effect of the production environment, perhaps suggesting that extracting training and testing stimuli from differing environments focuses listeners' attention on the acoustic cues that are consistent across environments.

3. Experiment 2

In Experiment 1, all stimuli came from a single talker, so it is unclear whether the results reflect talker-specific learning or a more general effect of recent exposure to these homophones. Experiment 2 tests whether identifications of homophones produced by a particular talker are more accurate after exposure to the same talker saying those homophones than after exposure to a different talker saying them.

3.1. Methods and Materials

Stimuli were made from recordings of two female American English speakers reading monosyllabic English words, elicited in randomized order with PsychoPy (Peirce, 2007) and recorded in a quiet room with a stand-mounted Blue Yeti microphone in the Audacity software program and digitized at a 44.1 kHz sampling rate with 16-bit quantization.

One of the talkers was naive to the purpose of the experiment, but the other was not. While there is a risk that some acoustic cues were emphasized in some way by the non-naive talker, it is unclear what acoustic effects would be expected or how listeners would respond to such patterns. If there were an effect of particular patterns produced by the non-naive talker,

higher accuracy might be predicted in the same-talker condition; however, as can be seen in the results below, there was little difference in accuracy between the same-talker and different-talker conditions.

The target words included 20 homophone mate pairs, all orthographically distinct, e.g. *chord*, *cord*. There were three conditions for training words, each containing 80 items (40 pairs of response options), as described below. A list of all words can be found in the appendix. All items were recorded in a frame sentence, *The word is* ___, and the target word was extracted to be presented in isolation.

Participants were 192 native speakers of American English (mean age 28.6; 80 male, 110 female, 2 nonbinary) with no reported speech or hearing disorders. 5 participants were excluded and replaced based on having accuracy below 80% for identifications of training items or taking longer than 30 minutes to complete the study.

The study was run online, with participants recruited and paid through the Prolific system and the experiment presented through Qualtrics.

Participants were instructed that they would hear English words and identify each one as matching one of two associated response options. The stimulus items were presented as a list; listeners clicked on an audio player icon to hear each stimulus. Listeners were allowed to listen to the recordings multiple times if they chose to. Responses were given by clicking on one of the written words given under the icon for the stimulus. Within a block, the order of items was randomized. The order of the two response options was balanced across participants.

There were two blocks: a training block and a testing block. As described in the conditions below, there were two different talkers whose voices appeared in the training phase for different conditions. The test tokens were always different from the tokens heard during training, even when the same word appeared in both phases.

First, listeners completed the training block. They heard a set of 80 items presented individually, all monosyllabic English words which they identified as matching one of two response options that differed only in the vowel, e.g. hear *chord* and select either "chord" or "card" as the written word matching the recording. Both items of each pair of response options were included as stimuli, and each training stimulus appeared only once. For example, there would be a trial with a *chord* stimulus and also a trial with a *card* stimulus, with the same response options "chord" and "card". There would be another trial in which listeners hear *cord* and select either "cord" or "cared" as the written word matching the recording, and so on. The paired item for each homophone mate was balanced across listeners, e.g. one listener would have "chord" vs "card" decisions and "cord" vs "cared" decisions, while another listener would have "cord" vs "card" decisions and "chord" vs "cared' decisions.

There were three different conditions for the exposure stimuli in this training phase. (1) In the same-talker training condition, the training items included all of the words that would subsequently appear in the homophone identification task, produced by the same talker (e.g. *chord*, *card* providing evidence for *chord*). (2) In the different-talker condition, the training items were the same words but produced by a different talker. (3) In the unrelated-training

condition, the training items only included words that would not appear in the homophone identification task, produced by a different talker than the one who produced the test stimuli. These were selected to be a relatively close phonological match to the items in the other conditions (e.g. *spore*, *spar*).

Second, listeners completed the testing block; listeners in all conditions heard the same test stimuli. They heard a set of 40 items presented individually, all monosyllabic English words which they identified as matching one of two orthographically distinct homophone mates, e.g. hear *chord* and select either "chord" or "cord" as the written word matching the recording. Listeners heard both items of each pair during the task.

Statistical results are from logistic mixed effects models, using the lme4 package (Bates et al., 2015) in R (R Core Team, 2022); p-values were calculated by the lmerTest package (Kuznetsova et al., 2015), which uses the Satterthwaite method for approximating degrees of freedom. The details of each model appear below to introduce each model before presenting a summary of its results.

3.2. Hypotheses and predictions

There are two main competing hypotheses for whether listeners will be above chance accuracy.

Hypothesis 1a: Listeners become familiar with how a talker says particular words, including differences between homophone mates, which could produce above-chance discrimination of homophone mates only with the same-talker training, when listeners have previously heard the talker saying those particular words.

Hypothesis 1b: Exposure to any talker saying these homophones might draw listeners' attention to the acoustic details that characterize them in this context, resulting in above-chance accuracy both in the same-talker condition and the different-talker condition.

3.3. Results

Accuracy in the training phase was 98.3%; recall that the decisions in the training phase were all choices between unambiguously phonologically distinct words (e.g. *spore*, *spare*) as matching the auditory stimulus. Results are reported only for the testing phase, in which listeners made decisions between homophone mates (e.g. *chord*, *cord*) as matching the auditory stimulus.

Table 4 presents the summary of a mixed effects logistic regression model for accuracy in Experiment 2.[5] The fixed effect was exposure condition (Same Talker, Different Talker, Unrelated Words). There were random intercepts for participant and for homophone pair.[6]

[5] $glmer(accuracy \sim Condition + (1|ParticipantID) + (1|pair), data = HomophoneDataSpeaker[which(HomophoneDataSpeaker\$Task == \text{“}testing\text{”}),], family = binomial)$

[6] Random slopes by participant are impossible because of the between-participants design of the conditions. Including a random slope for condition by pair produces a singularity error because there is little difference between two of the conditions and the existing variance is highly correlated with the random intercept by pair. For these reasons, no random slopes were included.

	Estimate	Std. Error	z value	p value
(Intercept)	0.165	0.0567	2.91	**0.00358**
Condition DifferentTalker	-0.0302	0.0657	-0.459	0.646
Condition UnrelatedWords	-0.129	0.0657	-1.96	**0.05**

Table 4: Logistic regression model for accuracy, Experiment 2. *Intercept: Condition = SameTalker*

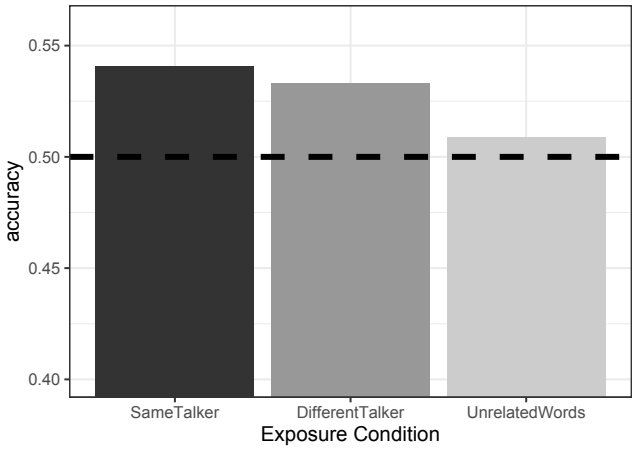

Figure 2: Overall accuracy in each condition for Experiment 2. 50% accuracy is marked with a dashed line.

As seen in the intercept, accuracy was significantly above chance when listeners had prior exposure to the homophones. Accuracy did not significantly differ between the same talker (54.1%) and different talker exposure conditions (53.3%), both of which exposed listeners to the target homophone mates in training. In the condition where listeners only heard unrelated words during training, accuracy was lower (50.9%). Figure 2 illustrates the overall accuracy in each condition.

There was substantial variation by participant and by homophone mate pair. The model with random intercepts for both of these has a significantly better fit than a model without the by-participant intercept ($\chi^2 = 9.69$, df $= 1$, $p = 0.00186$) or without the by-pair intercept ($\chi^2 = 17.7$, df $= 1$, $p < 0.001$).

What cues were listeners using to distinguish between homophone mates? In addition to differences between the two talkers, there were differences between homophone mates (e.g. *chord*, *cord*), which were larger than than between two instances of the same word (e.g. *chord*, *chord*). Table 5 presents a summary of differences between items in several acoustic characteristics that have been previously identified as differing between homophone mates.

Given the greater differences between homophone mates than between two instances of the same word, listeners are likely to be using some combination of these characteristics when they accurately identify homophones. Table 6 presents the summary of a mixed effects logistic regression model for accuracy that includes acoustic predictors based on difference between characteristics of the training stimuli and the testing stimuli.[7] Only the significant predictors

[7] $glmer(accuracy \sim Condition * VowelDurDiff + F0meanDiff + (1|ParticipantID) + (1|pair), data =$

	Vowel Duration	F0 mean	F0 range	Intensity	Spectral Tilt
SameWord, SameTalker	23.1 ms	19.4 Hz	53.8 Hz	1.2 dB	6.2 dB
HomophoneMates, SameTalker	52.7 ms	27.6 Hz	58.6 Hz	1.6 dB	6.5 dB
SameWord, DifferentTalker	29.4 ms	29.4 Hz	66.1 Hz	2.2 dB	9.6 dB
HomophoneMates, DifferentTalker	44.0 ms	31.0 Hz	72.3 Hz	2.1 dB	9.7 dB

Table 5: Acoustic differences between the same word (e.g. *chord*, *chord*) as produced by the same talker or different talkers, and between homophone mates (e.g. *chord*, *cord*) as produced by the same talker or different talkers

are included. The unrelated-words condition was excluded, as there was no corresponding training stimulus for the homophones. The fixed effects were exposure condition (Same Talker, Different Talker), vowel duration difference, F0 mean difference, and the interaction between exposure condition and vowel duration difference. There were random intercepts for participant and for homophone pair. The continuous variables were centered.

	Estimate	Std. Error	z value	p value
(Intercept)	0.0811	0.0652	1.24	0.214
Condition DifferentTalker	0.0585	0.0738	0.793	0.428
Vowel Duration Diff	-0.00553	0.00261	-2.12	**0.0345**
F0 mean Diff	-0.0037	0.00174	-2.12	**0.0337**
Condition DifferentTalker * Vowel Duration Diff	0.00705	0.00284	2.49	**0.0129**

Table 6: Logistic regression model for accuracy, including acoustic differences from the training stimuli as predictors, Experiment 2. *Intercept: Condition = SameTalker*

For both the different-talker condition and the same-talker condition, the difference in F0 mean was a significant predictor; the likelihood of an accurate response decreased with larger differences in F0. The difference in vowel duration was a significant predictor just for the same-talker condition; if the training stimuli and testing stimuli were produced by the same talker, the likelihood of an accurate response decreased with larger differences in vowel duration. The effect of vowel duration only in the same-talker condition may reflect adaptation to by-talker differences in speech rate; similarity to the duration of an earlier stimulus is a more reliable cue if that stimulus was produced by the same talker. These results suggest that similarity to the particular training stimuli does matter, though the relationship to listeners' pre-existing expectations might also matter.

The main results of this experiment are consistent with Hypothesis 1b: The effect of exposure to the homophones in training improves accuracy in testing regardless of whether the training talker and testing talker were the same. This suggests that accuracy of homophone

$HomophoneDataSpeaker[which(HomophoneDataSpeaker\$Task == \text{“}testing\text{”}),], family = binomial)$

identifications is not due to talker-specific learning, but instead increased attention to the acoustic details that characterize each word based on associations with its meaning.

4. Experiment 3

From the previous results, it seems that exposure to particular homophones in disambiguating contexts improves listeners' ability to subsequently identify other recordings of the same homophones. While this learning might depend on exposure to these particular words, it is possible that the same learning could be elicited based on exposure to words with similar meanings and thus similar emotional valence or other semantic associations that have acoustic correlates. Experiment 3 tests whether exposure to semantically similar words improves listeners' accuracy in identifying homophones.

4.1. Methods and Materials

Stimuli were made from recordings of one female American English speaker reading monosyllabic English words, elicited in randomized order with PsychoPy (Peirce, 2007) and recorded in a quiet room with a stand-mounted Blue Yeti microphone in the Audacity software program and digitized at a 44.1 kHz sampling rate with 16-bit quantization.

The talker was not naive to the purpose of the experiment, so there is a risk that some acoustic cues were emphasized in some way. However, the focus of this study is whether listeners are able to use acoustic cues to emotional valence to identify homophone mates; it has previously been established that differences based on valence exist (e.g. Nygaard et al., 2009; Nygaard & Lunders, 2002), and the acoustic characteristics of the stimuli are consistent with the previously reported patterns. The results of Experiment 2, in which one talker was naive to the purpose of the experiment and one was not, provide no evidence that this factor influences results.

The target words included 20 homophone mate pairs, selected such that one item of each pair had substantially more positive associations than the other (e.g. *die-dye* and *great-grate*). All homophone mates were orthographically distinct. There were three conditions for training words, each containing 80 items (40 pairs of response options), as described below. A list of all words can be found in the appendix. All items were recorded in a frame sentence, *The word is* ___, and the target word was extracted to be presented in isolation.

Participants were 192 native speakers of American English (mean age 38.1; 114 male, 78 female) with no reported speech or hearing disorders. 20 participants were excluded and replaced based on having accuracy below 80% for identifications of training items or taking longer than 30 minutes to complete the study.

The study was run online, with participants recruited and paid through the Amazon Mechanical Turk system and the experiment presented through Qualtrics.

Participants were instructed that they would hear English words and identify each one as matching one of two associated response options. The stimulus items were presented as a list; listeners clicked on an audio player icon to hear each stimulus. Listeners were allowed

to listen to the recordings multiple times if they chose to. Responses were given by clicking on one of the written words given under the icon for the stimulus. Within a block, the order of items was randomized. The order of the two response options was balanced across participants.

There were two blocks: a training block and a testing block. All stimuli were produced by the same individual. However, the test tokens were always different from the tokens heard during training, even when the same word appeared in both phases.

First, listeners completed the training block. They heard a set of 80 items presented individually, all monosyllabic English words which they identified as matching one of two response options that differed only in the vowel, e.g. hear *great* and select either "great" or "greet" as the written word matching the recording. Both items of each pair of response options were included as stimuli, and each training stimulus appeared only once. For example, there would be a trial with a *great* stimulus and also a trial with a *greet* stimulus, with the same response options "great" and "greet". There would be another trial in which listeners hear *grate* and select either "grate" or "grit" as the written word matching the recording, and so on. The paired item for each homophone mate was balanced across listeners, e.g. one listener would have "great" vs "greet" decisions and "grate" vs "grit" decisions, while another listener would have "grate" vs "greet" decisions and "great" vs "grit" decisions.

There were three different conditions for the exposure stimuli in this training phase. (1) In the same-word training condition, the training items included all of the words that would subsequently appear in the homophone identification task (e.g. *great*, *greet* providing evidence for *great*). (2) In the semantically-related condition, the training items included words that were semantically similar to the target homophone (e.g. *best*, *beast*; *best* is semantically related to *great*); this set of training stimuli was also designed to be phonologically similar to the same-word training. (3) In the unrelated-training condition, the training items only included words that would not appear in the homophone identification task and had neutral associations as much as possible, selected to be a relatively close phonological match to the items in the semantically-related condition (e.g. *guess*, *geese*).

Second, listeners completed the testing block; listeners in all conditions heard the same test stimuli. They heard a set of 40 items presented individually, all monosyllabic English words which they identified as matching one of two orthographically distinct homophone mates, e.g. hear *great* and select either "great" or "grate" as the written word matching the recording. Listeners heard both items of each pair during the task.

Statistical results are from logistic mixed effects models, using the lme4 package (Bates et al., 2015) in R (R Core Team, 2022); p-values were calculated by the lmerTest package (Kuznetsova et al., 2015), which uses the Satterthwaite method for approximating degrees of freedom. The details of each model appear below to introduce each model before presenting a summary of its results.

4.2. Hypotheses and predictions

There are three main competing hypotheses for whether listeners will be above chance accuracy.

Hypothesis 1a: Listeners become familiar with how particular words are said in this context, including differences between homophone mates, which could produce above-chance discrimination of homophone mates only with the same-word training, when listeners have previously heard the talker saying those particular words.

Hypothesis 1b: Exposure might make broad associations between meaning and acoustic form salient, resulting in above-chance accuracy in the semantically-related condition in addition to the same-word condition.

Hypothesis 1c: Because the homophone mates were selected to have strong positive or negative emotional valence, listeners might already have expectations based on broad associations between meaning and phonetic form. In this case, exposure might be unnecessary for drawing listeners' attention to these associations, resulting in above-chance accuracy in all conditions.

4.3. Results

Accuracy in the training phase was 98.2%; recall that the decisions in the training phase were all choices between unambiguously phonologically distinct words (e.g. *best*, *beast*) as matching the auditory stimulus. Results are reported only for the testing phase, in which listeners made decisions between homophone mates (e.g. *great*, *grate*) as matching the auditory stimulus.

Table 7 presents the summary of a mixed effects logistic regression model for accuracy in Experiment 3.[8] The fixed effect was exposure condition (Semantically Related, Unrelated Training, Same-Word Training). There were random intercepts for participant and for homophone pair.[9]

	Estimate	Std. Error	z value	p value
(Intercept)	0.173	0.0664	2.6	**0.00925**
Condition UnrelatedTraining	-0.0578	0.0674	-0.858	0.391
Condition SameWordTraining	-0.00818	0.0674	-0.121	0.903

Table 7: Logistic regression model for accuracy, Experiment 3. *Intercept: Condition = SemanticallyRelated*

As seen in the intercept, accuracy was significantly above chance when listeners had prior exposure to words that were semantically similar to the homophone mates that appeared in testing (54.2%). Accuracy did not significantly differ between conditions; accuracy was also significant in the condition with same word training (54.0%) and marginally significant in the condition with unrelated word training (52.8%). Figure 3 illustrates the overall accuracy in each condition.

[8] $glmer(accuracy{\sim}Condition + (1|ParticipantID) + (1|pair), data = HomophoneDataSemantic[which(HomophoneDataSemantic\$Task == \text{“}testing\text{”}),], family = binomial)$

[9] Random slopes by participant are impossible because of the between-participants design of the conditions. Including a random slope for condition by pair produces a singularity error because there is little difference between the conditions and the existing variance is highly correlated with the random intercept by pair. For these reasons, no random slopes were included.

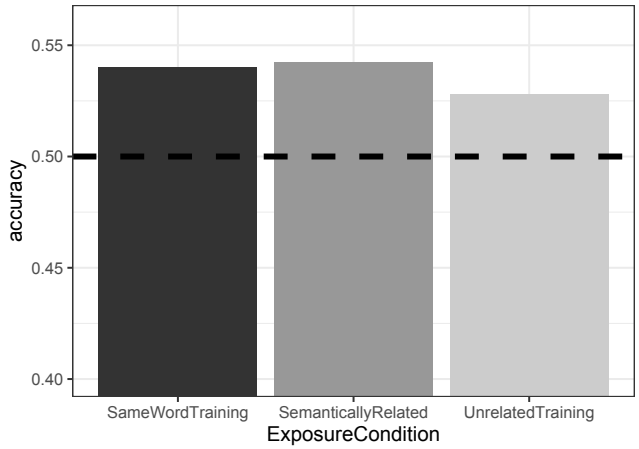

Figure 3: Overall accuracy in each condition for Experiment 3. 50% accuracy is marked with a dashed line.

There was substantial variation by participant and by homophone mate pair. The model with random intercepts for both of these has a significantly better fit than a model without the by-participant intercept ($\chi^2 = 12.9$, df $= 1$, $p < 0.001$) or without the by-pair intercept ($\chi^2 = 47.4$, df $= 1$, $p < 0.001$).

What cues were listeners using to distinguish between homophone mates? The homophone mates were designed so that the words in each pair differed in their emotional valence; one was always more positive than the other. Table 8 presents a summary of several acoustic characteristics that have been previously identified as differing between homophone mates, divided into the more positive and more negative item of each pair.

	Vowel Duration	F0 mean	F0 range	Intensity	Spectral Tilt
More positive	174.5 ms	246.2 Hz	71.1 Hz	63.1 dB	6.4 dB
More negative	186.3 ms	218.7 Hz	63.5 Hz	61.7 dB	2.8 dB

Table 8: Acoustic characteristics of the homophone mate of each pair with the more positive meaning and more negative meaning. Note that for half of the pairs the difference was positive vs neutral (e.g. *great, grate*) and for the other half the difference was neutral vs negative (e.g. *steel, steal*); what is being considered here is the valence relative to the word's homophone mate.

Given the acoustic differences between more positive and more negative items, listeners are likely to be using some combination of these characteristics when they accurately identify homophones. Table 9 presents the summary of a mixed effects logistic regression model for whether an item was identified as the more positive homophone mate, using acoustic characteristics of the testing stimuli.[10] Only the significant predictors are included. The fixed effects were F0 mean and Intensity. There were random intercepts for participant and for homophone pair. The continuous variables were centered.

Higher F0 mean predicted more responses of the more positive homophone mate. Higher

[10] $glmer(ResponsePositivity \sim F0mean + Intensity + (1|ParticipantID) + (1|pair), data = HomophoneDataSemantic[which(HomophoneDataSemantic\$Task == \text{“}testing\text{”}),], family = binomial)$

	Estimate	Std. Error	z value	p value
(Intercept)	0.0692	0.0917	0.755	0.45
F0 mean	0.00393	0.00115	3.41	**0.000651**
Intensity	0.0507	0.0142	3.58	**0.000345**

Table 9: Logistic regression model for whether an item was identified as the more positive homophone mate (vs the more negative one), Experiment 3.

intensity also predicted more responses of the more positive homophone mate. Both of these patterns are consistent with the previously observed relationships between acoustic characteristics and emotional valence, as well as the characteristics of this particular set of stimuli.

However, not all characteristics exhibited a strong effect on identification decisions, even those which exhibited a strong effect of valence in production. The lack of clear effects might be due to other factors like lexical frequency also influencing identification decisions. Additionally, the positive/negative divide is probably an oversimplification, so individual words might not fit into the general pattern if they are positive or negative in different ways.

The main results of this experiment are consistent with Hypothesis 1c: Listeners exhibited above chance accuracy in all conditions, suggesting that there are pre-existing associations between meaning and phonetic form. Those associations are likely to have a stronger effect in this experiment than in the others because the homophone mates were selected to have strong positive or negative emotional valence.

5. Discussion

These experiments provide evidence that listeners can attend to acoustic differences between homophone mates under some conditions. It is important to note that the behavior of homophone mates is not equivalent to the behavior of words with categorical phonological contrasts; accuracy is low but significantly above chance in several conditions of the experiments presented here.

Accuracy of homophone identifications was significantly above chance (53.6%) in Experiment 1 when listeners had been exposed to the same words produced by the talker during the training phase; this is significantly higher than accuracy in the condition when listeners had only been exposed to unrelated words. That is, when listeners had heard all homophones produced by the talker in association with written forms to disambiguate them, their subsequent accuracy in identifying new tokens of these homophones was higher. Accuracy is similarly above chance in the same condition in Experiment 2. This accuracy suggests that listeners are making use of acoustic characteristics of the words that cannot be reliably predicted by other words and which are only made salient by the training context.

There are a range of acoustic characteristics that listeners might be using to make these decisions for particular words. For Experiments 1 and 2, the difference between the training

stimulus and testing stimulus for each word was considered for several acoustic characteristics: Vowel duration, F0 mean, F0 range, intensity, and spectral tilt. While many measures exhibit more similarity between instances of the same word (e.g. *sun, sun*) than between homophone mates (e.g. *sun, son*), not all of them were significant predictors of responses. In Experiment 1, only the difference in intensity from the training stimulus was a significant predictor for accuracy of identifications of testing items. In Experiment 2, the difference in vowel duration and difference in mean F0 were significant predictors of accuracy. The variation in which predictors are significant probably indicates that the salient cues differ based on what additional sources of acoustic differences are present (different production environments that stimuli were extracted from in Experiment 1, different speakers in Experiment 2). Some variation is also likely due to the specific set of homophone mates and which acoustic characteristics listeners might attend to in those items based on what semantic, syntactic, or other factors are salient.

Previous work has identified several different sources of differences between homophone mates in production; particular homophone mate pairs will vary in which factors will be relevant. Lexical frequency has an impact on several phonetic characteristics, including duration and degree of vowel reduction (Gahl, 2008; Guion, 1995; Lohmann, 2017), though these effects are smaller outside of meaningful sentences (Guion, 1995; Sanker, 2019). The homophone mate pairs in the current study were selected to have broadly similar frequency as much as possible, but not all items were closely matched in frequency. Part of speech is another factor that predicts acoustic differences between homophone mates, such as duration and F0 (Lohmann, 2017). Some studies find that effects of part of speech disappear when other factors like position in the sentence are controlled for (Conwell, 2017; Lohmann, 2020), but the typical contexts that a word is often produced in may have an impact on production even when the word is produced in other contexts (Sóskuthy & Hay, 2017; Seyfarth, 2014; Tang & Shaw, 2021). Most of the homophone mates in the current study differed in part of speech, though they were all elicited in the same position, sentence-final. Morphological complexity can influence segmental durations in speech production (Walsh & Parker, 1983; Plag et al., 2017; Seyfarth et al., 2018), though the effects are variable. In perception, Bond (1973) found that listeners were not better than chance at identifying homophone mates, but the duration of the vowel was a predictor of whether a stimulus was identified as monomorphemic or bimorphemic (e.g. *wade* vs *weighed*). In the current study, the only homophone that is morphologically complex in a way that can be broken down segmentally was *heard*; whether or not *weight* is treated as monomorphemic may vary by individual, and the vowel quality which marks number in *feet* may produce different effects than have been observed for segmentally separable morphemes. Orthographic length of a word can influence duration (Warner et al., 2004), which may be particularly relevant when responses are given by selecting a written form to match the stimulus. Size and emotional valence also influence acoustic characteristics (e.g. Nygaard et al., 2009; Nygaard & Lunders, 2002); this factor was specifically examined in Experiment 3.

In Experiment 3, listeners identified homophone mates with significantly above chance accuracy when they had been exposed to semantically related words produced by the same talker. Accuracy was not significantly higher for listeners exposed to the target homophone mates during training. Accuracy in these conditions also was not significantly higher than

in the condition with unrelated-word training, so the set of homophone mates may itself be responsible for the results; the homophones were selected so that one item of each pair had strong positive or negative emotional valence. When listeners already have expectations about the pronunciation of a homophone based on associations between meaning and acoustic form, recent exposure to utterances of those words may have less of an impact on expectations. Consistent with previous work, more positive homophone mates had shorter duration, higher mean F0, larger F0 range, and higher intensity than more negative homophone mates. More positive homophone mates also had higher spectral tilt, suggesting phonation differences, with breathier vowels in more positive items. Mean F0 and intensity were both significant predictors of which item listeners selected as a response. The lack of clear effects for other acoustic characteristics might be due to other factors also influencing identification decisions (e.g. positivity predicts shorter duration for *feat*, but lexical frequency predicts shorter duration for *feet*). There is also variation in semantic associations beyond a simple positive/negative divide, e.g. positive words can be calm (e.g. *peace*) or exciting (e.g. *great*).

One potential way to account for this result is with exemplar models, in which speakers have acoustically detailed memories of particular utterances of words (Pierrehumbert, 2002; Goldinger, 1998). Hybrid exemplar models have exemplars structured both by words and by phonological categories (Pierrehumbert, 2002). Expectations about the realization of a sound will primarily be established by the phonological category, because there are far more exemplars supporting the realization of the component elements of a word at the phonological level than exemplars supporting the realization of the particular word. Unless the realization of a particular word is highly consistent and distinct from what is otherwise phonologically expected, the acoustic details of that word are not likely to move away from their phonological categories or develop obligatory characteristics that are not otherwise contrastive in the language. There is some evidence that acoustic characteristics which systematically occur in productions of a word based on its informativity or the contexts that it tends to occur in can enter the representation of that word, and thus be apparent even when the words are produced in contexts that would not cause the phonetic effects (Sóskuthy & Hay, 2017; Seyfarth, 2014; Tang & Shaw, 2021). However, some of these effects might be due to ease of access rather than those phonetic details being part of the representation of the word (Gahl et al., 2012; Kahn & Arnold, 2012).

I propose an exemplar model that includes an additional dimension of structure that groups words based on their semantic associations, e.g. words with positive or negative emotional valence. Like phonologically structured exemplar clouds, semantic clouds are supported by more exemplars than any individual word, allowing stronger associations between meaning and acoustic details to develop. Aspects of meaning like size and emotional valence influence acoustic characteristics in production (e.g. Nygaard et al., 2009; Nygaard & Lunders, 2002), and these characteristics also influence listeners' decisions about meaning (e.g. Nygaard & Lunders, 2002; Knoeferle et al., 2017). Listeners may be able to identify homophones in these experiments with above chance accuracy because of these broad associations between acoustic cues and semantic characteristics like size of the referent or emotional valence. This usage of associations between acoustics and semantic characteristics does not necessarily need to involve associations between particular words and specific acoustic details. Even if the

realization of a particular word is not consistent enough in regular usage to establish distinct word-specific acoustic details, listeners can have expectations for that word based on knowing what it means, consistent with Nygaard et al. (2009) finding that the pronunciation of nonce words exhibited significant effects of each word's assigned meaning. Exemplar models might also have additional dimensions of structure, e.g. for part of speech. Given that part of speech is associated with some acoustic differences (Lohmann, 2017), acoustic characteristics like duration and F0 may help set listeners' expectations for a word's part of speech even outside of syntactic contexts that would disambiguate.

For words with meanings that fall neatly into semantic categories that have relatively strong associations with acoustic characteristics, listeners may make use of those acoustic characteristics without any additional exposure, as is seen in Experiment 3. Under this analysis, the effect of the exposure phase in Experiments 1 and 2 was important because many homophones have meanings which do not have such strong associations with phonetic form. Thus, the training draws listeners' attention to the acoustic details and provides them with evidence for what acoustic cues to meaning will be present for these words in this context. Most words are not heard in isolation very often in natural speech, which may contribute to the effects of recent exposure in this particular context. Even when the semantic links are more complex, use of acoustic details is supported by pre-existing expectations based on words with similar semantic associations and phonetic details. When there is no pre-existing association to reinforce, listeners are likely to require substantial experimental exposure to produce a link between a lexical item and particular acoustic details, perhaps to a degree that does not occur in natural language usage. In a convergence task with different lexical items manipulated in opposite directions, Sanker (2021) found no evidence that listeners learn arbitrary word-specific phonetic characteristics. However, Rochet-Capellan & Ostry (2011) provide evidence that a word-specific shift is possible with sufficient exposure, using an altered auditory feedback experiment. While a key aspect of their experiment is probably the large amount of exposure to a small number of words, the method of shifting pronunciations might also be relevant; shifts elicited by altered auditory feedback might differ from shifts elicited in convergence or perceptual learning based on fundamental differences in what is being targeted.

The results of these experiments do not seem to depend on talker-specific learning. Experiment 2 found no significant difference between hearing the same talker in exposure and testing or different talkers in exposure and testing; accuracy was above chance in both conditions when the exposure included the particular homophone mates that would appear in testing.

In Experiment 1, the original sentential context that the words were extracted from was not a predictor of accuracy of homophone identifications. Previous work has demonstrated that some differences between homophone mates in natural sentences are eliminated when words are produced in isolation or in frame sentences (Guion, 1995; Sanker, 2019), which might predict higher accuracy for words extracted from natural sentences. Several factors might contribute to the lack of effect in this study. First, the environment always differed between training and testing. Distinct environments were used to avoid the possibility that listeners would make use of characteristics caused by the syntactic, semantic, or phonological environment; as Jurafsky et al. (2002) demonstrate, most differences between homophone

mates can be attributed to the context. Second, the words were separated from their original environments, which may obscure some prosodic patterns. Third, the homophone mate pairs in the study were selected so that the majority of them were similar in frequency. Lexical frequency is a major predictor of acoustic differences between homophone mates produced in natural context (Gahl, 2008); context effects may be reduced if the interaction between frequency and context is obscured by lack of variation in frequency.[11] This aspect of the study was aimed just at testing whether the production environment of the testing stimuli impacted accuracy, and not whether production environment of the test stimuli interacts with the production environment of the training stimuli. It is possible that exposure to homophones extracted from meaningful contexts in both training and testing would produce higher accuracy than homophones extracted from frame sentences in both conditions.

6. Conclusions

In Experiments 1 and 2, listeners could discriminate between homophone mates with above chance accuracy after exposure to disambiguated tokens of these words, but their accuracy was not above chance when prior exposure did not include the test words. There was no significant difference between hearing the same talker in exposure and testing or different talkers; accuracy was above chance when the exposure included the particular homophone mates that would appear in testing, regardless of the talker, indicating that the results are not due to talker-specific learning.

The particular set of lexical items used for homophone identification is important. In Experiment 3, using homophones with strongly positive or negative emotional valence resulted in higher accuracy in the control condition (exposure to unrelated words) than was found in the other experiments; these results suggest that identifications are based on broad expectations about how meaning is associated with acoustic details.

Some acoustic differences are tied to relatively strong semantically-driven patterns, which are accessible without any training. For items with more complicated or variable semantic associations, listeners can make use of acoustic cues when they are made salient by training. These results can be explained with an exemplar model that includes a semantic dimension that links words with similar emotional valence; meanings which are frequently produced with similar acoustic characteristics (e.g. higher F0 in happy words) will become associated with those acoustic characteristics.

References

Andruski, J. E., Blumstein, S. E., & Burton, M. (1994). The effect of subphonetic differences on lexical access. *Cognition*, *52*, 163–187. doi:10.1016/0010-0277(94)90042-6.

[11] The one pair that differed substantially in lexical frequency was *knight-night*. While this pair was one of the ones with higher accuracy, it was not an outlier.

Arnold, D., Tomaschek, F., Sering, K., Lopez, F., & Baayen, R. H. (2017). Words from spontaneous conversational speech can be recognized with human-like accuracy by an error-driven learning algorithm that discriminates between meanings straight from smart acoustic features, bypassing the phoneme as recognition unit. *PloS One*, *12*, Article e0174623. doi:10.1371/journal.pone.0174623.

Baayen, R. H., Chuang, Y.-Y., Shafaei-Bajestan, E., & Blevins, J. P. (2019). The discriminative lexicon: A unified computational model for the lexicon and lexical processing in comprehension and production grounded not in (de)composition but in linear discriminative learning. *Complexity*, *2019*, Article 4895891. doi:10.1155/2019/4895891.

Babel, M. (2010). Dialect divergence and convergence in New Zealand English. *Language in Society*, *39*, 437–456. doi:10.1017/S0047404510000400.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48. doi:10.18637/jss.v067.i01.

Bond, Z. S. (1973). The perception of sub-phonemic phonetic differences. *Language and Speech*, *16*, 351–355. doi:10.1177/002383097301600405.

Chiu, C.-Y. P. (2000). Specificity of auditory implicit and explicit memory: Is perceptual priming for environmental sounds exemplar specific? *Memory & Cognition*, *28*, 1126–1139. doi:10.3758/BF03211814.

Church, B. A., & Schacter, D. L. (1994). Perceptual specificity of auditory priming: Implicit memory for voice intonation and fundamental frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 521–533. doi:10.1037/0278-7393.20.3.521.

Cole, R. A., Coltheart, M., & Allard, F. (1974). Memory of a speaker's voice: Reaction time to same-or different-voiced letters. *Quarterly Journal of Experimental Psychology*, *26*, 1–7. doi:10.1080/14640747408400381.

Conwell, E. (2017). Prosodic disambiguation of noun/verb homophones in child-directed speech. *Journal of Child Language*, *44*, 734–751. doi:10.1017/S030500091600009X.

Creel, S. C., Aslin, R. N., & Tanenhaus, M. K. (2008). Heeding the voice of experience: The role of talker variation in lexical access. *Cognition*, *106*, 633–664. doi:10.1016/j.cognition.2007.03.013.

Dmitrieva, O. (2019). Transferring perceptual cue-weighting from second language into first language: Cues to voicing in Russian speakers of English. *Journal of Phonetics*, *73*, 128–143. doi:10.1016/j.wocn.2018.12.008.

Gahl, S. (2008). Time and thyme are not homophones: The effect of lemma frequency on word durations in spontaneous speech. *Language*, *84*, 474–496. doi:10.1353/lan.0.0035.

Gahl, S., Yao, Y., & Johnson, K. (2012). Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of Memory and Language*, *66*, 789–806. doi:10.1016/j.jml.2011.11.006.

Goh, W. D. (2005). Talker variability and recognition memory: Instance-specific and voice-specific effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 40. doi:10.1037/0278-7393.31.1.40.

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 1166–1183. doi:10.1037/0278-7393.22.5.1166.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*, 251–279. doi:10.1037/0033-295X.105.2.251.

Guerrini, J. (2020). Vowel quality and iconic lengthening. In M. Franke, N. Kompa, M. Liu, J. L. Mueller, & J. Schwab (Eds.), *Proceedings of Sinn und Bedeutung* (pp. 242–255). volume 24. doi:10.18148/sub/2020.v24i1.864.

Guion, S. G. (1995). Word frequency effects among homonyms. In *Texas Linguistic Forum* (pp. 103–116). volume 35.

Hintzman, D. L., Block, R. A., & Inskeep, N. R. (1972). Memory for mode of input. *Journal of Verbal Learning and Verbal Behavior*, *11*, 741–749. doi:10.1016/S0022-5371(72)80008-2.

Houde, J. F., & Jordan, M. I. (1998). Sensorimotor adaptation in speech production. *Science*, *279*, 1213–1216. doi:10.1126/science.279.5354.1213.

Jurafsky, D., Bell, A., & Girand, C. (2002). The role of the lemma in form variation. In C. Gussenhoven, & N. Warner (Eds.), *Laboratory Phonology VII* (pp. 3–34). Berlin: Mouton de Gruyter.

Kahn, J. M., & Arnold, J. E. (2012). A processing-centered look at the contribution of givenness to durational reduction. *Journal of Memory and Language*, *67*, 311–325. doi:10.1016/j.jml.2012.07.002.

Kapnoula, E. C., & Samuel, A. G. (2019). Voices in the mental lexicon: Words carry indexical information that can affect access to their meaning. *Journal of Memory and Language*, *107*, 111–127. doi:10.1016/j.jml.2019.05.001.

Keating, P. (1985). Universal phonetics and the organization of grammars. In V. A. Fromkin (Ed.), *Phonetic linguistics: Essays in honor of Peter Ladefoged* (pp. 115–131). Academic Press.

Knoeferle, K., Li, J., Maggioni, E., & Spence, C. (2017). What drives sound symbolism? Different acoustic cues underlie sound-size and sound-shape mappings. *Scientific Reports*, *7*, Article 5562. doi:10.1038/s41598-017-05965-y.

Kong, E. J., Beckman, M. E., & Edwards, J. (2012). Voice onset time is necessary but not always sufficient to describe acquisition of voiced stops: The cases of Greek and Japanese. *Journal of Phonetics*, *40*, 725–744. doi:10.1016/j.wocn.2012.07.002.

Kovic, V., Plunkett, K., & Westermann, G. (2010). The shape of words in the brain. *Cognition*, *114*, 19–28. doi:10.1016/j.cognition.2009.08.016.

Kraljic, T., & Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin & Review*, *13*, 262–268. doi:10.3758/BF03193841.

Kraljic, T., & Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. *Journal of Memory and Language*, *56*, 1–15. doi:10.1016/j.jml.2006.07.010.

Kuznetsova, A., Bruun Brockhoff, P., & Haubo Bojesen Christensen, R. (2015). *lmerTest: Tests in Linear Mixed Effects Models*. URL: https://CRAN.R-project.org/package=lmerTest r package version 2.0-29.

Lee, H., Politzer-Ahles, S., & Jongman, A. (2013). Speakers of tonal and non-tonal Korean dialects use different cue weightings in the perception of the three-way laryngeal stop contrast. *Journal of Phonetics*, *41*, 117–132. doi:10.1016/j.wocn.2012.12.002.

Lieberman, P. (1970). Towards a unified phonetic theory. *Linguistic Inquiry*, *1*, 307–322.

Lohmann, A. (2017). *Cut*(n) and *cut*(v) are not homophones: Lemma frequency affects the duration of noun-verb conversion pairs. *Journal of Linguistics*, *54*, 753–777. doi:10.1017/S0022226717000378.

Lohmann, A. (2020). Nouns and verbs in the speech signal: Are there phonetic correlates of grammatical category? *Linguistics*, *58*, 1877–1911. doi:10.1515/ling-2020-0249.

Martinuzzi, C., & Schertz, J. (2021). Sorry, not sorry: The independent role of multiple phonetic cues in signaling the difference between two word meanings. *Language and Speech*, (p. OnlineFirst). doi:10.1177/0023830921988975.

Mullennix, J. W., Pisoni, D. B., & Martin, C. S. (1989). Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America*, *85*, 365–378. doi:10.1121/1.397688.

Nielsen, K. (2011). Specificity and abstractness of VOT imitation. *Journal of Phonetics*, *39*, 132–142. doi:10.1016/j.wocn.2010.12.007.

Nygaard, L. C., Herold, D. S., & Namy, L. L. (2009). The semantics of prosody: Acoustic and percpetual evidence fo prosodic correlates to word meaning. *Cognitive Science*, *33*, 127–146. doi:10.1111/j.1551-6709.2008.01007.x.

Nygaard, L. C., & Lunders, E. R. (2002). Resolution of lexical ambiguity by emotional tone of voice. *Memory & Cognition*, *30*, 583–593. doi:10.3758/BF03194959.

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science*, *5*, 42–46. doi:10.1111/j.1467-9280.1994.tb00612.x.

Palmeri, T. J., Goldinger, S. D., & Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*, 309. doi:10.1037/0278-7393.19.2.309.

Peirce, J. W. (2007). PsychoPy–Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*, 8–13. doi:10.1016/j.jneumeth.2006.11.017.

Perlman, M., Clark, N., & Johansson Falck, M. (2015). Iconic prosody in story reading. *Cognitive Science*, *39*, 1348–1368. doi:10.1111/cogs.12190.

Pierrehumbert, J. (2002). Word-specific phonetics. In C. Gussenhoven, & N. Warner (Eds.), *Laboratory Phonology VII* (pp. 101–140). Berlin: Mouton de Gruyter.

Plag, I., Homann, J., & Kunter, G. (2017). Homophony and morphology: The acoustics of word-final S in English. *Journal of Linguistics*, *53*, 181?216. doi:10.1017/S0022226715000183.

Pufahl, A., & Samuel, A. G. (2014). How lexical is the lexicon? Evidence for integrated auditory memory representations. *Cognitive Psychology*, *70*, 1–30. doi:10.1016/j.cogpsych.2014.01.001.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria. URL: https://www.R-project.org/.

Ranbom, L. J., & Connine, C. M. (2007). Lexical representation of phonological variation in spoken word recognition. *Journal of Memory and Language*, *57*, 273–298. doi:10.1016/j.jml.2007.04.001.

Rochet-Capellan, A., & Ostry, D. J. (2011). Simultaneous acquisition of multiple auditory-motor transformations in speech. *Journal of Neuroscience*, *31*, 2657–2662. doi:10.1523/JNEUROSCI.6020-10.2011.

Rojczyk, A. (2011). Sound symbolism in vowels: Vowel quality, duration and pitch in sound-to-size correspondence. *Poznán Studies in Contemporary Linguistics*, *47*, 602–615. doi:10.2478/psicl-2011-0030.

Sanker, C. (2019). Effects of lexical ambiguity, frequency, and acoustic details in auditory perception. *Attention, Perception, & Psychophysics*, *81*, 323–343. doi:10.3758/s13414-018-1604-x.

Sanker, C. (2021). Convergence doesn't show lexically-specific phonetic detail. In R. Bennett, R. Bibbs, M. L. Brinkerhoff, M. J. Kaplan, S. Rich, A. Rysling, N. Van Handel, & M. W. Cavallaro (Eds.), *Supplemental Proceedings of the 2020 Annual Meeting on Phonology*. doi:10.3765/amp.v9i0.4896.

Scarborough, R., & Zellou, G. (2013). Clarity in communication: "clear" speech authenticity and lexical neighborhood density effects in speech production and perception. *Journal of the Acoustical Society of America*, *134*, 3793–3807. doi:10.1121/1.4824120.

Seyfarth, S. (2014). Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. *Cognition*, *133*, 140–155. doi:10.1016/j.cognition.2014.06.013.

Seyfarth, S., Garellek, M., Gillingham, G., Ackerman, F., & Malouf, R. (2018). Acoustic differences in morphologically-distinct homophones. *Language, Cognition and Neuroscience*, *33*, 32–49. doi:10.1080/23273798.2017.1359634.

Sóskuthy, M., & Hay, J. (2017). Changing word usage predicts changing word durations in New Zealand English. *Cognition*, *166*, 298–313. doi:10.1016/j.cognition.2017.05.032.

Tang, K., & Shaw, J. A. (2021). Prosody leaks into the memories of words. *Cognition*, *210*, Article 104601. doi:10.1016/j.cognition.2021.104601.

Verbrugge, R. R., Strange, W., Shankweiler, D. P., & Edman, T. R. (1976). What information enables a listener to map a talker's vowel space? *Journal of the Acoustical Society of America*, *60*, 198–212. doi:10.1121/1.381065.

Walsh, T., & Parker, F. (1983). The duration of morphemic and non-morphemic /s/ in English. *Journal of Phonetics*, *11*, 201–206. doi:10.1016/S0095-4470(19)30816-2.

Warner, N., Jongman, A., Sereno, J., & Kemps, R. (2004). Incomplete neutralization and other sub-phonemic duration differences in production and perception: Evidence from Dutch. *Journal of Phonetics*, *32*, 251–276. doi:10.1016/S0095-4470(03)00032-9.

Woods, K. J., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, & Psychophysics*, *79*, 2064–2072. doi:10.3758/s13414-017-1361-2.

Appendix

Table 10: Same-word (Homophone) Training Items, Experiment 1. The items are grouped by the response options that were given together (e.g. hear *brick*, decide between "brick" and "brake"). The items paired with each homophone mate were balanced across participants (e.g. half of participants had "brick" vs "brake" and "brook" vs "break" response options, while the other have had "brick" vs "break" and "brook" vs "brake").

brick - brake	brook - break
crook- creak	crack - creek
flaw - flea	fly - flee
hide - herd	hoard - heard
mud - maid	mad - made
meal - mail	mile - male
might - meet	moat - meat
neat - knight	net - night
pass - piece	pace - peace
poor - pier	pair - peer
root - write	rat - right
seal - sail	soul - sale
set - sight	seat - site
sour - soar	sir - sore
stick - stake	stack - steak
stall - steal	still - steel
sign - sun	sane - son
vine - vain	van - vein
will - wail	wall - whale
walk - weak	work - week

Table 11: Non-Homophone Training Items, Experiment 1. The items are grouped by the response options that were given together.

cake - kick	moan - moon	heap - hip	code - could
map - mop	file - fall	fur - fair	car - core
match - much	keep - cape	rail - real	tap - top
team - tame	role - rule	hill - heel	mess - mass
cheek - check	rage - ridge	hum - home	lid - led
mode - mood	lack - luck	chain - chin	lean - lane
shell - shall	pipe - peep	job - jab	cup - cope
peach - pitch	sheep - ship	feel - fell	main - mean
nut - knot	tug - tag	bake - back	ride - raid
lace - less	lawn - loan	wall - wool	wait - wheat

Table 12: Homophone Mate Items, Experiment 1. The items are grouped by the response options that were given together.

brake - break	creak - creek	flea - flee	herd - heard
maid - made	mail - male	meet - meat	knight - night
piece - peace	pier - peer	write - right	sail - sale
sight - site	soar - sore	stake - steak	steal - steel
sun-son	vain - vein	wail - whale	weak - week

Table 13: Same-word (SameTalker and DifferentTalker) Training Items, Experiment 2. The items are grouped by the response options that were given together (e.g. hear *owed*, decide between "owed" and "ad"). The items paired with each homophone mate were balanced across participants (e.g. half of participants had "owed" vs "ad" and "odd" vs "add" response options, while the other have had "owed" vs "add' and "odd" vs "ad").

owed - ad	odd - add
aisle - ale	all - ail
brick - brake	broke - break
card - cord	cared - chord
flow - flea	fly - fly
hard - herd	horde - heard
hall - whole	hail - hole
lit - loot	late - lute
meal - male	mile - mail
met - meat	might - meet
pass - piece	pus - peace
rate - right	rat - write
soil - sail	seal - sale
sane - seen	sign - scene
suit - sight	sit - site
swat - sweet	sweat - suite
vine - vain	van - vein
wheat - weight	wit - wait
wick - weak	wake - week
wall - whale	wheel - wail

Table 14: Unrelated Training Items, Experiment 2. The items are grouped by the response options that were given together.

lad - load	stack - stock	lane - line	trait - trite
lake - lick	cakes - coax	coal - call	spore - spar
tea - toe	pleat - plight	fur - far	curse - course
door - dare	troll - trail	doom -dim	room - rhyme
tame - team	mill - mile	dean - den	bite - bet
keep - cap	heat - hut	wide - wade	kite - cat
fail - foil	same - seem	meek - make	sheen - shine
bright - brute	lime - lamb	sweep - swap	bead - bed
brain - brine	stare - star	raid - ride	late - lit
seek - sick	feed - fade	hail - hall	cane - keen

Table 15: Homophone Mate Items, Experiment 2. The items are grouped by the response options that were given together.

ad - add	ale - ail	brake - break	cord - chord
flea - flee	herd - heard	whole - hole	loot - lute
male - mail	meat - meet	piece - peace	right - write
sail - sale	seen - scene	sight - site	sweet - suite
vain - vein	weight - wait	weak - week	whale - wail

Table 16: Same-word Training Items, Experiment 3. The items are grouped by the response options that were given together (e.g. hear *brick*, decide between "brick" and "brake"). The items paired with each homophone mate were balanced across participants (e.g. half of participants had "brick" vs "brake" and "broke" vs "break" response options, while the other have had "brick" vs "break" and "broke" vs "brake").

brick - brake	broke - break
dare - deer	dire - dear
due - dye	day - die
fear - fare	fur - fair
fit - feet	fate - feat
floor - flour	flair - flower
greet - grate	grit - great
hill - heel	hail - heal
pace - piece	pass - peace
rat - write	rate - right
rise - raise	rose - raze
sear - soar	sir - sore
slow - sleigh	sly - slay
stale - steel	still - steal
swat - suite	sweat - sweet
veal - vial	veil - vile
wake - week	wick - weak
wan - one	wine - won
west - waist	worst - waste
wheel - whale	will - wail

Table 17: Semantically Related Training Items, Experiment 3. The items are grouped by the response options that were given together. Recall that one of the items from each pair of response options was semantically related to one of the homophones that would appear in testing; the related homophone is given in parentheses after each item.

snap (cf. *break*) - snip	rein (cf. *brake*) - ran	fond (cf. *dear*) - fanned	goat (cf. *deer*) - got
hurt (cf. *die*) - heart	tint (cf. *dye*) - taunt	kind (cf. *fair*) - coined	toll (cf. *fair*) - toil
skill (cf. *feat*) - scale	leg (cf. *feet*) - lag	bloom (cf. *flower*) - blame	wheat (cf. *flour*) - wit
best (cf. *great*) - beast	grind (cf. *grate*) - grinned	mend (cf. *heal*) - mind	toe (cf. *heel*) - two
bliss (cf. *peace*) - bless	part (cf. *piece*) - port	wreck (cf. *raze*) - rake	lift (cf. *raise*) - left
true (cf. *right*) - tree	note (cf. *write*) - knot	kill (cf. *slay*) - keel	sled (cf. *sleigh*) - slid
pain (cf. *sore*) - pin	flight (cf. *soar*) - flit	rob (cf. *steal*) - robe	tin (cf. *steel*) - teen
nice (cf. *sweet*) - niece	room (cf. *suite*) - roam	bad (cf. *vile*) - bed	tube (cf. *vial*) - tub
weep (cf. *wail*) - whip	moose (cf. *whale*) - moss	dump (cf. *waste*) - damp	hip (cf. *waist*) - heap
frail (cf. *weak*) - frill	days (cf. *week*) - doze	champ (cf. *won*) - chomp	four (cf. *one*) - fur

Table 18: Unrelated Training Items, Experiment 3. The items are grouped by the response options that were given together.

nap - nip	main - man	bond - band	coat - caught
dirt - dart	flint - flaunt	lines - loins	coal - coil
still - stale	beg - bag	moon - mane	seat - sit
guess - geese	dine - din	bend - bind	show - shoe
miss - mess	cart - court	tech - take	knit - net
new - knee	tote - taught	pill - peel	red - rid
chain - chin	light - lit	lob - lobe	bin - bean
pipe - peep	choose - chose	lad - led	cube - cub
sleep - slip	boot - bought	lump - lamp	ship - sheep
trail - trill	graze - grows	ramp - romp	horde - herd

Table 19: Homophone Mate Items, Experiment 3. The items are grouped by the response options that were given together.

brake - break	deer - dear	dye - die	fare - fair
feet - feat	flour - flower	grate - great	heel - heal
piece - peace	raise - raze	write - right	sleigh - slay
soar - sore	steel - steal	suite - sweet	vial - vile
whale - wail	waist - waste	week - weak	one - won