Homophone discrimination based on prior exposure

Chelsea Sanker[*]

Yale University
Dow Hall Room 204, P.O. Box 208366
New Haven, CT 06520, USA

[*] Corresponding author
Email address: chelsea.sanker@yale.edu (Chelsea Sanker)

Homophone discrimination based on prior exposure

Chelsea Sanker*

Yale University
Dow Hall Room 204, P.O. Box 208366
New Haven, CT 06520, 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; exposure to semantically similar words also results in above-chance identification of homophones. These results suggest that listeners can make use of semantically-driven acoustic differences between homophone mates when recent exposure makes these details salient. While this could be explained with word-specific phonetic representations, it could also be explained by listeners becoming attuned to the meaning of each homophone mate and how they align with associations between form and meaning.

Keywords: word-specific phonetics, speaker-specific learning, homophones, perception

1. Introduction

The realization of particular sounds includes phonetic details that must be learned. Under Exemplar Theory, particular words could also have their own distinct phonetic details. 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 detail (Gahl, 2008; Guion, 1995). However, listeners generally cannot discriminate between homophone mates (Bond, 1973; Sanker, 2019). Low accuracy in discrimination tasks might indicate that differences in production are simply an effect of the production

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

context, rather than differences in the representation. However, it is also possible that perceptual discrimination depends on facilitative conditions 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-associated acoustic differences between homophone mates when recent exposure to those words makes such details salient.

1.1. Phonetic details in the phonological representation

Many phonetic details must be learned, rather than being the automatic result of broad 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 thus 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 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 attend to for phonological categorization decisions (Dmitrieva, 2019; Lee et al., 2013) and rating naturalness of tokens (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 is hard of hearing (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). Experimental evidence provides some support for acoustically detailed memories of recent speech. 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).

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 and the probability of flapping. Tang & Shaw (2021) demonstrate that the 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 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 parallel 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. 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 them 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 (e.g. Gahl, 2008; Lohman, 2018), which 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 might suggest that the acoustic differences are 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, without listeners necessarily having word-specific phonetic representations. Bond (1973) found that the duration of a vowel in the stimulus influences decisions between homophone mates, even though the selection preferences did not result in accurate identifications. Duration is substantially influenced by lexical frequency (Gahl, 2008; Guion, 1995), so if the typical difference in duration between two homophone mates is large relative to the variation in duration of each word, listeners might have above chance accuracy in distinguishing between those words based on expectations about how lexical frequency relates to duration.

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 lexical 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 demonstrating gradient effects. 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 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 (Mullennix et al., 1989), and same-different decisions are similarly 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 along with that voice saying that word (Kapnoula & Samuel, 2019). Exemplar memories of particular tokens also include non-linguistic background noise; 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). However, memories for acoustic details of recent tokens do not necessarily indicate that these details ever enter word-specific phonological representations.

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 as the test stimuli 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 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 40 pairs, 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 39.0; 68 male, 59 female, 1 nonbinary) with no reported speech or hearing disorders. 6 participants were excluded and replaced based on having accuracy below 75% for identifications of training items; the training items were decisions between phonologically distinct English words, which should be unambiguous.

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,

[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 some possible sources of variation. For example, 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 (96%-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.

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 *sight* and select either *sight* or *seat* as the written word matching the recording. Both items of each pair were included as training stimuli.

There were two different conditions for the exposure stimuli in this training phase. (1) In the homophone-exposure training condition, the training pairs included all of the words that would subsequently appear in the homophone identification task (e.g. *sight*, *seat*); because the written response options 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 pairs 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. *sight*, *site*). Listeners heard both items of each pair 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 a logistic mixed effects model, calculated with the lme4 package in R (Bates et al., 2015); p-values were calculated by the lmerTest package (Kuznetsova et al., 2015).

2.2. Hypotheses and predictions

There are three main competing hypotheses for whether listeners will be above chance accuracy based on the exposure condition.

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 should also predict above-chance accuracy in all 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). Such an effect is unlikely to interact with the set of exposure stimuli, 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

Results are reported only for the testing phase, in which listeners made decisions between homophone mates.

Table 1 presents the summary of a mixed effects logistic regression model for accuracy. 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); and the interaction between condition and context. There were random intercepts for participant and for homophone pair.

As seen in the intercept, accuracy was significantly above chance 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;

	Estimate	Std. Error	z value	*p* value
(Intercept)	0.15	0.058	2.6	**0.0097**
Cond NoHomExposure	-0.19	0.08	-2.4	**0.016**
Context Sentence	-0.035	0.08	-0.43	0.66
Cond NoHomExposure * Cont Sentence	0.082	0.11	0.73	0.47

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

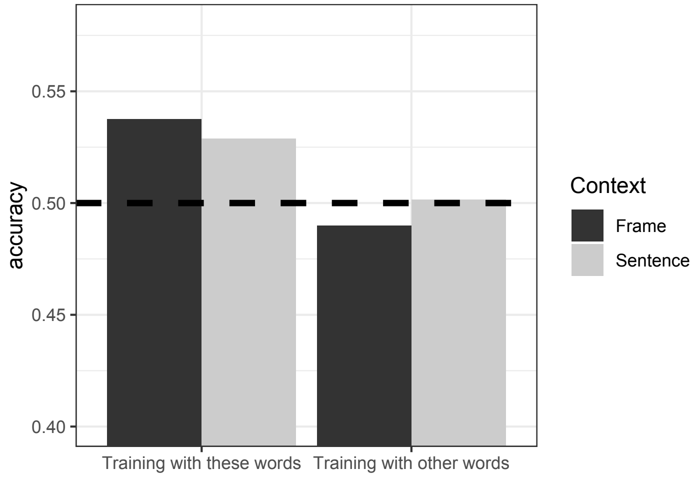

Figure 1: Mean accuracy in each condition for Experiment 1. 50% accuracy is marked with a dashed line.

accuracy in this condition is also significantly above chance if Context is excluded as a factor). Figure 1 illustrates the mean accuracy in each condition.

Accuracy was significantly lower when the training phase did not include prior exposure to these homophones. 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 these 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; the results are not clearly driven by particular words or particular participants. This model does not have a significantly better fit than a model without the by-participant intercept (χ^2 = 0.0049, df = 1, p = 0.944) or without the by-pair intercept (χ^2 = 0.598, df = 1, p = 0.44).

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 exposure to a particular talker saying the target homophones results in higher accuracy in identifications of homophones produced by the same talker than exposure to a different talker saying the same words.

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.

The target words included 20 homophone mate pairs, all orthographically distinct, e.g. *chord*, *cord*. There were three conditions for training words, each containing 40 pairs, 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 28.8; 80 male, 110 female, 2 nonbinary) with no reported speech or hearing disorders. 3 participants were excluded and replaced based on having accuracy below 75% for identifications of training items; the training items were decisions between phonologically distinct English words, which should be unambiguous.

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. 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 testing 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 were included as training stimuli.

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

Statistical results are from a logistic mixed effects model, calculated with the lme4 package in R (Bates et al., 2015); p-values were calculated by the lmerTest package (Kuznetsova et al., 2015).

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

Results are reported only for the testing phase, in which listeners made decisions between homophone mates.

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

	Estimate	Std. Error	z value	*p* value
(Intercept)	0.16	0.0562	2.85	**0.00431**
Condition DifferentTalker	-0.0254	0.0655	-0.388	0.698
Condition Unrelated Words	-0.124	0.0654	-1.89	**0.0583**

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

As seen in the intercept, accuracy was significantly above chance when listeners had prior exposure to semantically related words. Accuracy did not significantly differ between the same talker and different talker exposure conditions, both of which exposed listeners to the

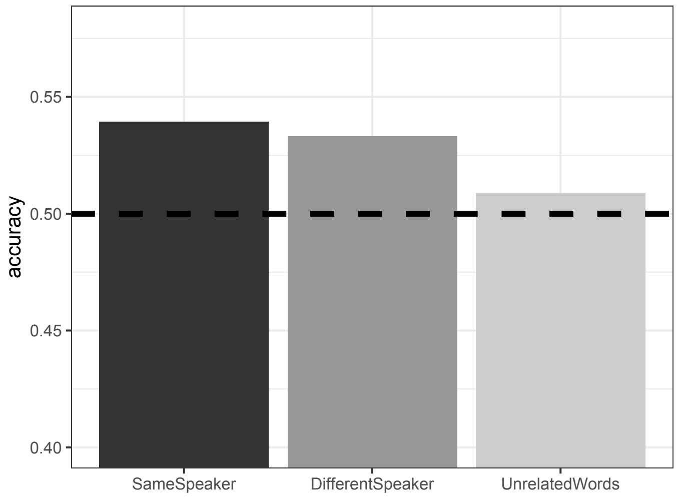

Figure 2: Mean accuracy in each condition for Experiment 2. 50% accuracy is marked with a dashed line.

target homophone mates in training. In the condition where listeners only heard unrelated words during training, accuracy was lower; the difference approached significance. Figure 2 illustrates the mean 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 ($\chi^2 = 9.2$, df = 1, $p = 0.00241$) or without the by-pair intercept ($\chi^2 = 16.6$, df = 1, $p < 0.001$).

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

As in the preceding experiments, 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 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 40 pairs, 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 38.7; 114 male, 78 female) with no reported speech or hearing disorders. 11 participants were excluded and replaced based on having accuracy below 75% for identifications of training items; the training items were decisions between phonologically distinct English words, which should be unambiguous.

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. 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 were included as training stimuli.

There were three different conditions for the exposure stimuli in this training phase. (1) In the same-word training condition, the training pairs included all of the words that would subsequently appear in the homophone identification task (e.g. *great*, *greet*). (2) In the semantically-related condition, the training pairs included words that were semantically similar to the target homophone (e.g. *best* as an item matched with *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 pairs 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. *great*, *grate*). Listeners heard both items of each pair during the task.

Statistical results are from a logistic mixed effects model, calculated with the lme4 package in R (Bates et al., 2015); p-values were calculated by the lmerTest package (Kuznetsova et al., 2015).

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 a talker says particular words, 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

Results are reported only for the testing phase, in which listeners made decisions between homophone mates.

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

	Estimate	Std. Error	z value	*p* value
(Intercept)	0.16	0.0678	2.37	**0.018**
Condition UnrelatedTraining	-0.0594	0.0672	-0.884	0.377
Condition SameWordTraining	0.0143	0.0673	0.213	0.832

Table 3: 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. Accuracy did not significantly differ between conditions. Figure 3 illustrates the mean 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 ($\chi^2 = 12.4$, df = 1, $p < 0.001$) or without the by-pair intercept ($\chi^2 = 53.0$, df = 1, $p < 0.001$).

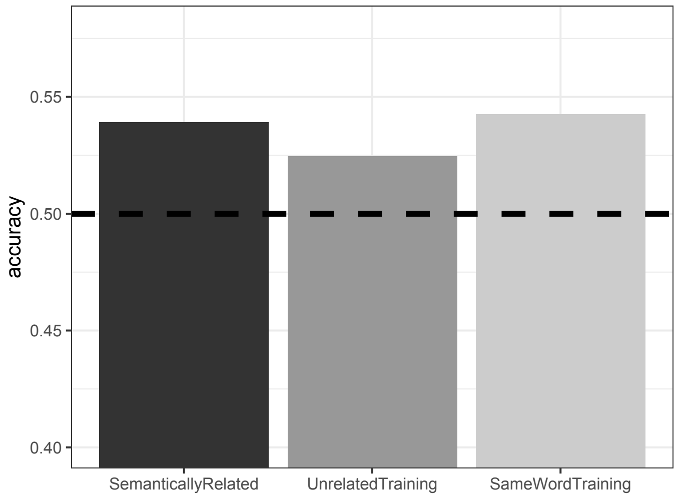

Figure 3: Mean accuracy in each condition for Experiment 3. 50% accuracy is marked with a dashed line.

5. Discussion

Accuracy of homophone identifications was significantly above chance (53%) 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 Experiments 2 and 3. This accuracy suggests that listeners are learning acoustic characteristics of the words that cannot be reliably predicted by other words and which are not salient without the training context.

One potential way to account for this result is with word-specific acoustic details. Exemplar models can predict word specific detail, and even memories for a specific talker producing a specific word (cf. Pierrehumbert, 2002; Goldinger, 1998). In such a model, clouds of exemplar memories also exist for phonological categories, which connect to all of the clouds for particular words and particular talkers. Given the robust shared representation at the phonological level, expectations about the realization of a sound will primarily be established by the phonological category, except when consistent phonologically-conditioned patterns arise that split the category. The acoustic details of particular words are not likely to move away from their category or include obligatory characteristics that are not otherwise contrastive in the language. However, there are tendencies in how specific words are produced based on the contexts that they tend to occur in (Jurafsky et al., 2002; Seyfarth, 2014; Tang & Shaw, 2021), which may be part of talkers' memories of those words.

If talkers have detailed memories of word-specific patterns in production, the typical acoustic details of particular words might be accessible to play a role in perception when made salient by context. Setting word-specific expectations may depend on pre-existing associations that listeners can draw on. 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. On the other hand, Rochet-Capellan & Ostry (2011) elicited distinct formant shifts in distinct words in an altered auditory feedback experiment. The 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 or differences in the number of words and amount of exposure to each word.

The results of the experiments presented here might alternatively be explained by broad associations between meaning and acoustic characteristics. Even if word-specific acoustic details do exist in the representation, such details are likely to be supported by broader expectations. While some short-term learning of arbitrary details might be possible with extensive exposure, listeners already have expectations about non-arbitrary acoustic characteristics associated with meaning. 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). Acoustic cues to semantic characteristics do not necessarily need to involve associations between particular words and specific acoustic details at all. Listeners may be able to identify homophones in these tasks with above chance accuracy because of these broad associations between acoustic cues and semantic characteristics like size of the referent or emotional valence, which have been made salient by the presence of these homophones in the training phase.

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. This result is consistent with accuracy being driven by expectations of broad semantic influences on phonetic characteristics rather than word-specific phonetic knowledge. However, 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 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.

Under this analysis, the effect of the exposure phase in Experiment 1 was important because many homophones have meanings which do not have similarly strong associations with phonetic form. Thus, the training draws listeners' attention to the acoustic details and provides the listeners with evidence for what acoustic cues to meaning will be present for 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. The presence of many homophones in the homophone-exposure training condition might make the differences between homophone mates particularly salient and make listeners more attuned to the relationship between form and meaning.

The results 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, either as a main effect or in interaction with the training condition. Previous work has demonstrated that some of the differences between homophone mates that are found in natural sentences are eliminated when words are produced in isolation or in frame sentences (Guion, 1995; Sanker, 2019), which might predict that accuracy would be higher 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. The distinct environments were used to avoid the possibility that listeners would learn distinctive characteristics caused by the syntactic, semantic, or phonological environment of surrounding words. As Jurafsky et al. (2002) demonstrate, most differences between homophone mates can be entirely attributed to the context. Second, the words were separated from their original environments; altering the environment 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 as produced in natural context (Gahl, 2008); the appearance of context effects may be reduced if the interaction between frequency and context is obscured by lack of variation in frequency.[2] This aspect of the study 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

Listeners could discriminate between homophone mates with above chance accuracy after exposure to disambiguated tokens of these words produced by the same talker, but they were not above chance accuracy 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 also suggest that identifications are based on broad expectations about how meaning is reflected in form more than word-specific phonetic memories.

Listeners are sensitive to semantically-driven acoustic differences between homophone mates when they are made salient by the training, and can use those differences to distinguish

[2] 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; the overall accuracy cannot be attributed to sensitivity to frequency.

between homophone mates with slightly higher than chance accuracy. However, associations between form and meaning do not necessarily reflect word-specific acoustic details; effects of meaning on acoustic characteristics are reflected similarly across words, which establishes broad associations between form and meaning that can be accessed to discriminate between homophone mates.

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.

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.

Lohman, A. (2018). *Cut*(n) and *cut*(v) are not homophones: Lemma frequency affects the duration of noun-verb conversion pairs. *Journal of Linguistics*, *54*, 1–25. doi:10.1017/S0022226717000378.

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.

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.

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.

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.

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.

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 4: Same-word (Homophone) Training Pairs, Experiment 1

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

Table 5: Non-Homophone Training Pairs, Experiment 1

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 6: Homophone Mate Pairs, Experiment 1

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 7: Same-word (SameTalker and DifferentTalker) Training Pairs, Experiment 2

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

Table 8: Unrelated Training Pairs, Experiment 2

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 9: Homophone Mate Pairs, Experiment 2

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 10: Same-word Training Pairs, Experiment 3

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

Table 11: Semantically Related Training Pairs, Experiment 3

snap - snip	rein - ran	fond - fanned	goat - got
hurt - heart	tint - taunt	kind - coined	toll - toil
skill - scale	leg - lag	bloom - blame	wheat - wit
best - beast	grind - grinned	mend - mind	toe - two
bliss - bless	part - port	wreck - rake	lift - left
true - tree	note - knot	kill - keel	sled - slid
pain - pin	flight - flit	rob - robe	tin - teen
nice - niece	room - roam	bad - bed	tube - tub
weep - whip	moose - moss	dump - damp	hip - heap
frail - frill	days - doze	champ - chomp	four - fur

Table 12: Unrelated Training Pairs, Experiment 3

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 13: Homophone Mate Pairs, Experiment 3

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