Sorry, Not Sorry: The independent role of multiple phonetic cues in signaling the difference between two word meanings

Caitlyn Martinuzzi
University of Toronto, Canada

Jessamyn Schertz
University of Toronto Mississauga, Canada
University of Toronto, Canada

Abstract
We examine the use of multiple subphonemic differences distinguishing homophones in production and perception, through a case study focusing on the distinction between two polysemous senses of the English word “sorry” (apology vs. attention-seeking). An analysis of production data from voice actors revealed significant and substantial durational differences between the two meanings. Tokens expressing an apology were longer than attention-seeking tokens, and the situational intensity of the context also independently affected duration. When asked to identify the meaning in a two-way forced-choice task after hearing each token spliced out of its context, listeners were above chance (64.7% accuracy) in identifying the intended meaning, and their responses were significantly correlated with the duration, intensity, and intonation contour (but not mean F0) of the productions. In a second perception task, listeners heard tokens of “sorry” that had been systematically manipulated to vary in duration, intensity, and intonation contour, with responses indicating that each of these dimensions played an independent role in listeners’ judgments. The results highlight the importance of broadening the scope of research on the use of subphonemic detail during lexical access and considering a wider range of lexical and non-lexical factors that condition variability on multiple acoustic dimensions, in order to work toward a more accurate picture of the systematic variability available in the input and tracked by listeners.

Keywords
Phonetic cue weighting, homophones, subphonemic differences

Corresponding author:
Jessamyn Schertz: Department of Language Studies, University of Toronto Mississauga and Department of Linguistics, University of Toronto, 3359 Mississauga Road North, Mississauga, Canada ON L5L1C6.
Email: jessamyn.schertz@utoronto.ca
Introduction

1.1 Overview

Homophones, by definition, are expected to be pronounced identically; however, they can be characterized by small but systematic differences in their phonetic realization. For example, the more frequent members of a given homophone pair tend to be of shorter duration than their less-frequent homophone twins (e.g., English “fair” is on average shorter than the less-frequent “fare”) (Gahl, 2008; Lohmann, 2018b). A variety of factors, including lexical frequency, underlying phonological representation, and contextual predictability, have been shown to condition systematic variation in homophones, and this variation occurs across multiple acoustic dimensions, as will be reviewed below. Furthermore, listeners can, in some circumstances, identify the intended meaning at greater-than-chance levels, and listeners’ responses have been found to be independently influenced by specific acoustic properties of the sound, such as duration (e.g., Bond, 1973; Warner et al., 2004). These sorts of findings suggest not only that subphonemic differences corresponding to the different meanings or senses must be perceptible in the signal, but also that the relevant information must be encoded in listeners’ speech perception systems at some level of representation and/or processing.

This case study provides an in-depth investigation of listeners’ use of multiple phonetic cues (duration, intensity, and pitch) in differentiating between two meanings of the English word “sorry”: an apology (i.e., “Sorry! I didn’t mean to hurt you,” henceforth Apology) versus an attention-seeking mechanism (i.e., “Sorry—can you move so I can pass by?”, henceforth ExcuseME). We first lay out the acoustic characteristics of productions from a set of four voice actors producing the word “sorry” in different contexts designed to elicit the two meanings. We then examine whether listeners are sensitive to the intended form of the word, as well as which phonetic cues are used to inform their decisions. We test these questions using two perception tasks: the first examines listeners’ use of cues when making judgments about actors’ productions of “sorry,” and the second uses a subset of these stimuli, systematically manipulated across the relevant dimensions, allowing us to assess the independent contribution of each cue.

1.2 Subphonemic variation in production of homophones

Previous investigations of subphonemic variation in homophones have primarily focused on durational differences. Higher-frequency words are, all else being equal, produced with shorter word durations than lower-frequency words, and this is also the case with pairs of homophones (at least for content words; studies involving only function words have not found similar effects, e.g., Jurafsky et al., 2002). In the first large-scale study on this topic, Gahl (2008) examined the durations of 344 homophonous but heterographic word pairs (e.g., “fair” vs. “fare”) in the Switchboard corpus of American English telephone conversations, and found that higher-frequency items were shorter in duration than their lower-frequency counterparts. This result persisted after controlling for factors including speaking rate, contextual predictability, and syntactic category (see Lohmann, 2018b, for a re-analysis and confirmation of this result). Similar patterns were found in a comparison of children and adults (Conwell, 2018) and in Mandarin (Sherr-Ziarko, 2015). The generalization also holds in homographic homophones, as demonstrated by Lohmann (2018a), who examined the durations of 63 homographic noun–verb pair homophones (e.g., “cut” (N) vs. “cut” (V)) in the Buckeye Corpus of spontaneous speech.

Durational differences across members of homophone pairs have also been shown to be conditioned by other factors, including the morphological status of the word (Seyfarth et al., 2018), the
underlying phonological form of the word (“incomplete neutralization,” Port & O’Dell (1985) among many others), and orthography (Warner et al., 2004). Additional factors systematically affecting duration that have not been tested specifically with homophones could also be expected to play a role as well (e.g., contextual predictability: Seyfarth, 2014; Tang & Shaw, 2020). Bell et al. (2009), in a corpus study of conversational English, showed that three separate factors—frequency, contextual predictability, and repetition—all contribute independently to durations of spoken words, even though they are all correlated.

Though duration is the most well-studied property, similar influences might be expected on other acoustic dimensions, particularly those which, like duration, are heavily influenced by the extent of articulatory reduction. A few studies have given evidence for other types of differences: Wright (2004) shows that properties of the lexical neighborhood affect spectral characteristics of vowels, with “easy” lexical items (i.e., those with low neighborhood density and high relative frequency) showing systematically less dispersion in the F1-F2 space than “hard” lexical items, and Tang and Shaw (2020), using a large telephone corpus of Mandarin Chinese, show that contextual predictability has significant effects on maximum pitch and intensity, dimensions related to prosodic prominence.

While all of the phonetic patterns discussed so far are at least indirectly attributable to the extent of articulatory reduction (with shorter durations, smaller pitch range, lower intensity, and less vowel space dispersion), there are many other sources of phonetic variability that could result in systematic differences across members of a homophone pair. For example, if the two meanings tend to be produced in different prosodic environments, they would show different distributions along many phonetic dimensions in the input, particularly for suprasegmental features like F0 or duration. There are a wide range of factors systematically affecting both the phonological (e.g., whether a word is accented or not) and phonetic detail (e.g., F0 alignment and range) of prosody, including discourse status, or topic structure (e.g., whether a topic is given or new; see Hirschberg, 2002, and references therein). Furthermore, intonational patterns may be idiomatically related to certain usages or meanings (Calhoun & Schweitzer, 2012).

Another potential source of differences in a homophone pair is another type of context: the emotion or affect used for each meaning. There is a large body of literature documenting the acoustic correlates of emotion (e.g., Williams & Stevens, 1972; Sauter et al., 2010, among many others; see Bachorowski & Owren, 2008, for a review). One example is work done by Banse and Scherer (1996), who recorded 12 actors asked to portray 14 different emotions in meaningless utterances consisting of non-words. They performed detailed acoustic analyses involving dimensions hypothesized to be relevant to emotion, including multiple measures of F0, intensity, speaking rate, and spectral measures corresponding to voice quality (e.g., spectral tilt). Almost all of the 29 features were statistically significant in explaining some of the emotion-based variance, with pitch and intensity measures capturing the most variance, and discriminant-based classification accuracy was well above chance (40% accuracy given a choice of 14 emotions, with chance being 1/14, or 7%). Relatedly, Ohala (1983) proposes that a biologically based “frequency code” underlies cross-language (and cross-species) regularities in use of higher F0 to signal more submissive, less assertive stances. While the acoustic correlates of emotion or attitude are far from deterministic, it is well established that there are some regularities, and if there are systematic differences in the contexts in which two members of a homophone pair are used, then it follows that there will also be systematic differences in their pronunciation.

Given the many factors at play simultaneously, it is likely that any given pair of words will differ along multiple dimensions for multiple reasons. In a case study, Drager (2011) examined variation across different grammatical uses of the word “like” in a corpus of sociolinguistic interviews in a New Zealand high school. Quotative “like” (as in, “I was like ‘yeah, okay’”) had on average
more monophthongal vowels, shorter initial-consonant-to-vowel ratios, and higher mean pitch than when it was used as a discourse particle or a lexical verb. Differences in intonational patterns were proposed to underlie the pitch differences: “Impressionistically, lexical verb like seemed to be produced in conjunction with a dip in the intonation contour, whereas quotative like rarely was and was sometimes part of a rising contour that raised more steeply after the verb” (Drager, 2011: 702). The different lemmas tend to occur in different prosodic positions, and while Drager argues that prosody is unlikely to account for all of the variation found in the dataset, it does likely have some effect. Though only focusing on a single wordform, this case study demonstrated that a given pair of homophones will be characterized by differences across multiple acoustic dimensions. Whether listeners track these sorts of differences, which dimensions are tracked by listeners, and whether the tracking is different depending on the conditioning factor (e.g., lexical frequency, prosody, emotion) are open questions.

1.3 Listeners’ use of acoustic variation in the perception of homophones

The presence of the systematic acoustic variation discussed above brings up the question of whether listeners make use of this information to inform their perceptual strategies. Listeners have been found to distinguish between the two members of homophone pairs at above-chance accuracy, although these effects tend to be very small, indicating that the cues are far from deterministic. For example, Sanker (2019) showed that listeners were significantly above chance in identifying members of a homophone pair that had been spliced out of sentences (e.g., “a doe is a deer” vs. “a dough is a mixture”) in a forced-choice task. However, accuracy was only very slightly above chance (50.8%), and this effect only held for words that had been produced in a contextually predictable sentence, and not in a different condition where the words were produced in isolation.

Listeners have also been shown to be sensitive to incomplete neutralization: Port and O’Dell (1985) showed that German listeners identified minimal pairs differing in underlying final consonant voicing, like “rat” and “rad” (which are both phonologically devoiced word-finally, such that the broad transcription of both is [ʁɑt]), at 59% accuracy. Looking at the same phenomenon in Dutch, Warner et al. (2004) found that listeners performed above chance in identifying the intended production, but only for stimuli drawn from certain speakers—specifically, those who showed greater durational distinctions between final underlying voiced versus voiceless consonants—suggesting that these durational differences were in fact what listeners were using to inform their decisions. The idea that these sorts of phonetic differences may have only limited usefulness for listeners is supported by the fact that the word-final voicing distinction, which is represented orthographically, is a common source of spelling mistakes in Dutch (Sandra et al., 2001). Finally, there may be cases where these sorts of cues are not used by listeners at all. In Mandarin, Tone 3 (low) is pronounced similarly to Tone 2 (rising) in certain phonological contexts, but systematic phonetic differences indicate that this T3/T2 neutralization is incomplete: however, listeners do not reliably distinguish the two categories (Zhang & Peng, 2013).

Prosodic variation has also been shown to be useful to listeners. For example, listeners differentiate short words from longer words containing them (e.g., cap vs. captain) based on their duration (e.g., Davis et al., 2002; Salverda et al., 2003, for English and Dutch listeners, respectively), and Spinelli et al. (2010) showed that listeners use F0 contour as an independent cue to word segmentation in French. Prosodic patterns may also be responsible for listeners’ demonstrated sensitivity to emotion or affect. Multiple studies have shown listeners to be consistently above chance at choosing a speaker’s intended emotion in forced-choice tasks (see Johnstone & Scherer, 2000, for a review). Furthermore, listeners’ perception of different emotions or affects is systematically influenced by specific acoustic properties. For example, intonation contour has been found to independently
influence listeners’ assessment of speaker certainty (Gravano et al., 2008), and listeners tend to associate higher intensity and lower F0 with dominance (e.g., Scherer et al., 1973; Puts et al., 2007). Banse and Scherer (1996) found correlations between listeners’ identification of emotions and acoustic properties of the utterances (see also Sobin & Alpert, 1999). The correspondence is far from deterministic, and listeners’ use of cues does not always reflect the apparent importance of cues in productions (see Johnstone & Scherer, 2000, and Bachorowski & Owren, 2008, for discussion). Nevertheless, as with the other subphonemic differences discussed in this section, they do appear to be able to influence listeners’ perception of emotion in at least some situations.

Finally, the prosody and/or affective tone of an utterance may influence lexical selection or processing. In a study by Nygaard and Lunders (2002), listeners heard words spoken in happy, sad, or neutral voices and were asked to transcribe them. Critical stimuli were members of homophone pairs in which one member had emotional connotations and the other did not (e.g., sad die vs. neutral dye). Listeners were more likely to transcribe the emotional token when it was heard in an emotionally congruent tone of voice, indicating that the emotional information was integrated at some level of linguistic processing, influencing lexical selection.

Taken together, there are a large number of studies demonstrating listeners’ sensitivity to the different types of subphonemic variation discussed above. It is important to remember that many of these effects are very small, inconsistent, or task-dependent. Particularly given the fact that positive results are more likely to be published (“publication bias,” e.g., de Bruin et al., 2015), it should not be concluded that listeners are always able to use this information. Instead, which sources of variation and which acoustic dimensions are tracked by listeners remains an open question.

1.4 Implications for representation, storage, and processing

The examination of subphonemic differences between different meanings of homophones in production, and listeners’ sensitivity to those differences in perception, has been of interest in large part because of its relevance to models of lexical representation. Systematic phonetic variability is pervasive in speech, governed by “contexts” such as the properties of the speaker, the interlocutor, the social/discourse/pragmatic context, and the syntactic/prosodic/semantic context. Since any two meanings are likely to be used in systematically different contexts, they will likely show different distributions of phonetic properties in production. How this variation is tracked, stored, and used by listeners is a question that has implications for the architecture of the lexicon and the structure of the speech processing system.

As discussed above, listeners are sensitive to subphonemic differences in meanings and use it to inform perception, at least in some situations, and this indicates that distinct information must be linked to each meaning. However, this link could take different forms. First, each of the two meanings could be directly linked to separate phonetic representations, either in the form of acoustic “exemplars” or as a summary of distributional statistics of the acoustics. In this situation, listeners would determine the meaning by evaluating the incoming signal against the two possible pronunciations. Alternatively, both meanings could be linked to the same phonetic representation, but each meaning has distinct information about the context in which it tends to be produced. These contexts are in turn linked to general phonetic regularities which also form part of listeners’ knowledge. In this case, the link between meaning and pronunciation is indirect and modulated by context: the phonetic information would suggest a certain context to listeners, and knowledge of the context would in turn inform the decision about the meaning of a word.

There is convincing evidence that both abstraction and acoustic specificity play a role in phonetic representations (e.g., McQueen et al., 2006; Ernestus, 2014; Pierrehumbert, 2016), and teasing apart the relative contributions is not possible with a case study. However, if listeners can make
use of phonetic differences, it is necessary for there to be independent representations of meanings at some level. If they are not separately specified phonetic representations, but rather shared, then it is necessary that the relevant contexts are tracked independently and stored separately for the different meanings. There is evidence that homophones (with unrelated meanings) and polysemes (with related meanings) are represented differently by listeners: specifically, there is psycholinguistic evidence that while the different meanings of a homophone pair have independent semantic representations, the different “senses” of polysemes may fall under the same representation (Rabagliati & Snedeker, 2013; Rodd et al., 2002). If senses do share a representation, it is difficult to see how either option above (separate pronunciations or separate meanings) could be stored.

1.5 The current study

We present a case study examining listeners’ use of multiple phonetic cues in distinguishing between two meanings of “sorry”: when it is used as an apology versus when it is used as an attention-seeking mechanism (EXCUSEME). Restricting the domain to a single word pair allows for an in-depth look at multiple dimensions, including pitch, intensity, and durational measurements. Previous work has shown that listeners make use of phonetic detail to inform perception of homophones; this has primarily been focused on one dimension at a time, focusing either on durational cues or (in a few cases) intonation contour. However, as we have seen above, any pair of words will likely differ on multiple dimensions for multiple reasons, including frequency, contextual predictability, and the prosodic, emotional, and pragmatic contexts they are likely to occur in. In this work, we test whether these other dimensions are used by listeners, and we assess their relative role in cuing the distinction for listeners. Our study is structured around three questions:

1. Which, if any, acoustic characteristics distinguish the Apology vs. EXCUSEME meanings of the English word “sorry”?  
2. Do listeners accurately perceive the difference between the two uses of “sorry,” when hearing isolated tokens removed from their original context?  
3. Which phonetic cues do listeners use when making their judgments, and what is their relative reliance on each cue?

We address these goals in a series of three experiments. In Experiment 1, we analyze productions of “sorry” recorded by voice actors in contexts designed to elicit the two different meanings in different levels of situational intensity, and examine the patterning of several acoustic cues shown to be relevant to homophone distinctions and affect/emotional distinctions in previous work: duration, pitch, and intensity. In Experiment 2, we test listeners’ accuracy in identifying the intended meaning of tokens of “sorry” produced in Experiment 1 and spliced out of their context. We also look at the correlation of listeners’ responses with the acoustic properties of the stimuli to begin to establish which of the cues, if any, might play a role in listeners’ decisions. Based on significant correlations between listeners’ responses and multiple acoustic dimensions in Experiment 2, Experiment 3 examines listeners’ perception of tokens of “sorry” that have been systematically manipulated across three dimensions, to assess whether these cues play an independent role in perception and to quantify listeners’ relative reliance on each cue. Prior to running the experiment, we did not have specific hypotheses about the use of individual cues. However, given the fact that apologies are likely to be said in more submissive, less assertive attitudes, we might expect them to have higher F0 and lower intensity than EXCUSEME. It is also possible that the two meanings are associated with different intonation contours, whether for structural reasons or because they are idiomatically specified (e.g., Calhoun & Schweitzer, 2012).
If listeners associate the two senses with different phonetic realizations, it indicates two things. First, the different representations of the polysemous “sorry” must be independent at some level of processing, regardless of whether they are linked to distinct pronunciations, and/or associated with distinct contexts, which in turn are associated with systematic phonetic patterns. Second, listeners must be sensitive to this acoustic detail and actively make use of it during speech perception. An additional contribution of this study is to examine perception of tokens systematically manipulated along multiple dimensions, as compared to perception of naturally produced tokens, where the dimensions may covary. This allows us to evaluate the extent to which listeners use each of these dimensions independently.

2 Experiment 1—Production

2.1 Overview

The purpose of Experiment 1 was to provide stimuli for perception experiments, as well as to examine the question of which phonetic cues, if any, distinguish productions of the word “sorry” used as an Apology from those meaning ExcuseMe in this dataset. We use recordings from four voice actors producing sentences containing the two types in context as the basis for our analysis. We used scripted dialogues in order to maintain control over the phonetic content of the utterances, and we recruited voice actors because we expected them to be comfortable reading scripted dialogues in a relatively naturalistic way (see Banse & Scherer, 1996, for discussion of the ecological validity of using voice actors to simulate emotional communication).

2.2 Methods

2.2.1 Participants, materials, and procedure. Two male and two female American English voice actors recruited from fiverr.com (a platform for freelance workers) were paid to take part in the experiment. Eighteen scenarios containing the word sorry were used for the production study. Table 1 shows six sample scenarios, and the full set is provided in the supplementary material. Each of these 18 scenarios fit into one of six general situations, half of which included someone apologizing, and the other half of which included someone needing to say “excuse me.” From each of these general scenarios, three specific situations, varying on a three-tier scale of “situational intensity,” were created, resulting in the total of 18 scenarios. For example, the Apology example in Table 1 belongs to the “Book” base scenario grouping. In the low-intensity version of this scenario, only the cover of the book was damaged, while in the medium situational intensity the entire book is significantly damaged, and in the high-intensity setting the book was the friend’s most treasured possession and was completely ruined. The ExcuseMe example belongs to the “Coffee Shop” base scenario. In the low-intensity version of this scenario, the subject needs to pick up coffee while running ahead of schedule, while in the medium and high intensities, this changes to running slightly late in the former situation and running 30 minutes late in the latter.

The script given to the participants included all scenarios. As shown in Table 1, each scenario included a first paragraph to provide context, which was not read aloud, followed by two lines of dialogue, which the participants were asked to read aloud. The first line of dialogue (spoken by the character named Alex) was the sentence containing the target word “sorry,” and it remained identical, including in punctuation, across each of the three intensity levels within each base scenario group. In all sentences, “sorry” was in utterance-initial position, followed by either a stop or affricate (this was done to facilitate clean splicing of the word for use in the subsequent perception experiments). The order of scenarios was pseudo-randomized such that no scenarios of the same base group, or of the same intensity, appeared immediately next to one another.
Table 1. Six of the scenarios used in the production experiment. The “context” was read silently, while the two lines of dialogue, including the target sentence, spoken by “Alex,” were read out loud.

| Apology (“Book” scenario) | ExcuseMe (“Coffee Shop” scenario) |
|---------------------------|----------------------------------|
| **Intensity Level 1**     | **Intensity Level 1**            |
| [Context]: [Alex’s friend has a new book they’ve been interested in reading but haven’t had time yet. Alex wanted to read the book as well so his/her friend lent them their copy. Alex really enjoyed the book and goes to return it and encourage his/her friend to make time to read it so they can discuss it. When s/he leaves his/hers apartment it’s raining pretty hard, so to protect the book s/he puts it in a plastic bag; however, on the way there s/he bumps into someone and the bag rips, dropping the book in a puddle and damaging the cover. Alex feels somewhat responsible and goes to apologize.] | [Context]: [Alex was on his/her way in to work when s/he got a text from his/her co-worker asking him/her to pick up the coffee orders for the office since it was his/her turn. Alex had no problem with this as s/he left home with enough time to make a small detour. Once Alex collects the orders s/he heads over to the self-serve table to add the appropriate amounts of cream and sugar to the various cups. At the counter there’s a woman putting milk in her/his tea who is blocking Alex’s access to the sugar. S/he’s not in a rush but if s/he doesn’t leave soon s/he might be pushing it so s/he decides to reach past the woman to grab the sugar.] |
| Alex: Sorry! Please don’t be mad! It was an accident! I really didn’t mean for it to happen! | Alex: Sorry, just gonna reach past you to grab the sugar real quick, hope you don’t mind. |
| Friend: It’s fine, the cover art was ugly anyway. | Woman: Oh, feel free. |
| **Intensity Level 2**     | **Intensity Level 2**            |
| [Context]: [Alex’s friend has a new book they’ve been interested in reading but haven’t had time yet. Alex agrees to give the book a go so his/her friend lent them their copy. Alex really enjoyed the book and goes to return it and discuss the story with their friend. When s/he leaves his/hers apartment it’s raining pretty hard, so to protect the book s/he puts it in a plastic bag; however, on the way there s/he bumps into someone and the bag rips, dropping the book in a puddle and soaking through all the pages, warping it to twice its normal size even though the text is still readable. Alex feels pretty bad about what happened and goes to apologize.] | [Context]: [Alex was on his/her way in to work when s/he got a text from his/her co-worker asking him/her to pick up the coffee orders for the office since it was his/her turn. Alex was a little behind schedule but figured s/he probably had time if s/he was quick. Once Alex collects the orders s/he heads over to the self-serve table to add the appropriate amounts of cream and sugar to the various cups. At the counter there’s a woman putting milk in her/his tea who is blocking Alex’s access to the sugar. S/he’s in a bit of a rush at this point and if s/he doesn’t leave soon s/he’ll definitely be late so s/he decides to reach past the woman to grab the sugar.] |
| Alex: Sorry! Please don’t be mad! It was an accident! I really didn’t mean for it to happen! | Alex: Sorry, just gonna reach past you to grab the sugar real quick, hope you don’t mind. |
| Friend: That really sucks, but if you buy me a replacement I’ll forgive you. | Woman: Sure. |
| **Intensity Level 3**     | **Intensity Level 3**            |
| [Context]: [Alex’s friend has been going on and on about their favorite book for all the years Alex has known them, their mom gave it to them when they were 14 just before she passed away and they have this book memorized cover to cover. It’s their most treasured possession and if you ever asked them what one thing they would save in a fire, their answer would be the book. They’ve been insisting Alex read it for years so they can talk about the story with each other and as a last resort to get Alex to read it, they lent him/her their copy. After reading the book Alex found s/he really did enjoy it and is going to meet up with his/her friend to discuss the story. When s/he leaves his/hers apartment it’s raining pretty hard, so to protect the book s/he puts it in a plastic bag; however, on the way there s/he bumps into someone and the bag rips, dropping the book in a huge puddle and utterly ruining it. S/he feels awful and immediately goes to apologize to their friend.] | [Context]: [It was already looking like Alex would be late to work when s/he got a text from his/her co-worker asking him/her to pick up the coffee orders for the office since it was his/her turn. Alex cursed his/her bad luck but knew that it was in fact his/her turn. Once Alex collects the orders s/he speed walks over to the self-serve table to add the appropriate amounts of cream and sugar to the various cups. At the counter there’s a woman putting milk in her/his tea who is blocking Alex’s access to the sugar. S/he starts tapping her foot but realizes unless s/he says something s/he could be waiting awhile and s/he’s already been at least 30 minutes late. S/he decides to reach past the woman to grab the sugar.] |
| Alex: Sorry! Please don’t be mad! It was an accident! I really didn’t mean for it to happen! | Alex: Sorry, just gonna reach past you to grab the sugar real quick, hope you don’t mind. |
| Friend: I don’t know what to say. | Woman: Go ahead. |
Participants were sent the script described above and given instructions about the recording procedure. Recordings took place in a quiet environment using personal audio-recording equipment. Participants were instructed to read the context silently to themselves, then to read both lines of dialogue out loud using their regular voices. They were asked not to whisper; this instruction was included in an attempt to avoid breathy voice, which could pose problems for pitch measurement and/or manipulation. Participants recorded two repetitions of the 18 dialogues from the scenarios above, resulting in 144 tokens of “sorry” in context (18 scenarios x 2 repetitions x 4 actors).

In addition to these “contextual” dialogues containing the word “sorry,” participants were also asked to record 20 instances of “sorry” in isolation, 10 as if they were apologizing, and 10 as if they were saying “excuse me.” They were asked to indicate orally when they switched from one set to the other and to feel free to vary the mood however they saw fit. No further specific instructions were given. Five tokens of each Sorry Type from each speaker were chosen for the analysis.

2.2.2 Annotations and measurements. We focused on phonetic dimensions shown in previous work to be relevant to phonetically differentiating between homophones, and/or differentiating expression of emotion: pitch (mean and contour) (ex: Bänziger & Scherer, 2005; Sauter et al., 2010), duration (Gahl, 2008; Lohmann, 2018a; etc.) and intensity (Pereira & Watson, 1998).

All annotations and measurements were done using Praat (Boersma & Weenink, 2018). First, the following landmarks were manually annotated: fricative onset (marked by the onset of frication in /s/), fricative offset (marked by the beginning of visible formants in the vowel following /s/), and word end (the end of stable F2).

F0 (in Hz) was measured at seven time points (0%, 10%, 25%, 50%, 75%, 90%, 100%) across the periodic portion of the word (i.e., everything except the initial fricative). In order to minimize measurement errors, we first determined speaker-specific floors and ceilings based on manual inspection of the data and used these as the basis for automatic pitch measurements. Visible outliers were checked and manually corrected.

Mean intensity (in dB) was measured across the full word.

One additional source of variability that we thought might be relevant was the intonation used to produce the target word. Based on initial inspection of the data, tokens varied in whether they had a final rising or falling contour. Examples of each of these is given in Figure 1, and the corresponding audio files are available in the supplementary materials. Data were coded based on the perceptual judgments of a native speaker of English familiar with the ToBI (Tones and Break Indices) system (Beckman & Ayers, 1997). Final rise tokens fell into the HLH% or LH% categories, and final fall tokens were HLL% or LL%, but were coded into a binary choice of final rise versus final fall for ease of analysis. A second annotator with no knowledge of the study provided independent judgments, and there was 76% agreement. Only tokens which were consistently annotated (n = 140) are included in the intonation contour plots and analyses below.

Based on the landmarks and measurements above, we use the following measures as dependent variables in the statistical analyses below:

1. Total word duration: fricative onset to word end;
2. Mean F0: the average across the seven points measured;
3. Mean intensity;
4. Pitch contour: final rise versus final fall.

2.2.3 Statistical analysis. Our primary question for the production experiment was to determine whether there are phonetic differences between the two types of “sorry” (Apology vs. ExcuseME). To evaluate this, we used mixed-effects regression, using the lme4 (Bates et al., 2015) package in
R (R Core Team, 2019) to assess the effect of Sorry Type on each dimension shown above. For continuous dimensions (all dimensions except pitch contour), we used a linear mixed-effects model (or, for pitch contour, a logistic mixed-effects model) predicting the value of that dimension from Sorry Type. Sorry Type was simple coded (apology: −0.5, excuse me: 0.5). The model included random intercepts for speaker and item (where each of the 18 scenarios represented an item), as well as a by-speaker slope for Sorry Type. P-values were computed using the lmerTest package (Kuznetsova et al., 2017), and an alpha-level of 0.05 was used as the threshold for significance. Although not our primary question of interest, we also tested whether the degree of situational intensity influenced each acoustic dimension (model details described below).

2.3 Results

This experiment sought to determine whether “sorry” as an apology is phonetically distinct from a “sorry” meaning “excuse me,” and if so, which phonetic cues distinguish them. We examined four dimensions that were expected, based on previous work, to be potentially relevant: word duration, mean F0, intensity, and pitch contour. 184 tokens were analyzed (144 “contextual” tokens plus 40 spoken in isolation). Tokens for which more than half of the pitch contour was undefined (n = 9) were omitted from the F0 analysis. We discuss results for each dimension in turn in the following paragraphs. Graphs of all dimensions are provided in Figure 2, and statistical results in Table 2. For continuous dimensions, the graphs show the overall distribution of all tokens (violin plots) for each Sorry Type, as well as speaker-specific means and 95% confidence intervals. For the graph showing the binary variable pitch contour, the proportion of tokens produced with a final rising (vs. falling) tone for each speaker is given.
**Duration:** The first panel of Figure 2 shows the distribution of duration values for *Apology* vs. *ExcuseMe* tokens. Overall, *ExcuseMe* tokens are shorter than *Apology* tokens, and this pattern is consistent across speakers. As indicated in the statistical results (Table 2), tokens are on average 565 ms (as indicated by the estimate for Intercept), and *Apology* tokens are on average 126 ms longer than those intended as *ExcuseMe* (as indicated by the estimate for Sorry Type). This effect of Sorry Type is significant.

**Mean F0:** The second panel of Figure 2 shows the distribution of mean F0 values across Sorry Types. There is no clear difference, nor is there a consistent pattern across speakers. This is reflected in the statistical results: while *ExcuseMe* tokens are on average 13.68 Hz lower than *Apology* tokens (as indicated by the estimate for Sorry Type), this difference is not significant.

![Figure 2](image-url)
Mean Intensity: The third panel of Figure 2 shows the distribution of mean intensity values across Sorry Types. There is again no consistent pattern, and this is again reflected in the statistical results; numerically, EXCUSEME tokens are on average 0.32 dB lower than APOLOGY tokens, but this difference is not significant.

Pitch contour: As shown in the final panel of Figure 2, the speakers differed in their distribution of pitch contours across the Sorry Types, with three speakers producing more final-rising-tone tokens for EXCUSEME than APOLOGY, and one showing the opposite pattern. Since the pitch contour variable was binary (vs. the three continuous factors discussed above), the statistical model was a mixed-effects logistic regression model. The estimates therefore represent the log odds of a final rising tone (vs. final falling tone). The positive estimate for the intercept indicates that overall, final rising tone is more common than final falling tone, but the lack of significance indicates that this difference is not above chance. The positive estimate for Sorry Type indicates that EXCUSEME tokens are characterized by rising tone more frequently than APOLOGY tokens, but this difference is again not significant.

2.3.1 Situational intensity. Although not our primary question of interest, we examined whether the degree of situational intensity influenced each acoustic dimension. In order to test this, we created four models, one for each dimension, which differed from the models described above in that (1) the 40 tokens produced in isolation were omitted from analysis, and (2) the model structure included an additional predictor variable of Situational Intensity, as well as its interaction with Sorry Type. Situational Intensity was simple-coded as a three-level factor, with Level 1 as the reference level. The only significant effects were found in the model for duration: as above, there was a main effect of Sorry Type, showing longer durations for APOLOGY than EXCUSEME tokens ($\beta = -123.22, SE = 25.63, t = -4.81, p = 0.017$), and there was also a main effect of Situational Intensity, where tokens produced in the Level 3 intensity scenarios (across both Sorry Types) were on average 46 ms longer than those produced in Level 1 ($\beta = 46.22, SE = 19.72, t = 2.34, p = 0.026$). This difference is shown in Figure 3. No other main effects or interactions for any of the models were significant.

2.4 Production: Interim summary

The results above show that APOLOGY tokens in this dataset are significantly longer than EXCUSEME tokens. Numerically, they are also characterized by higher intensity, higher F0, and a lower likelihood of a final rising tone, but these differences were not significant. Durations were also slightly longer for situations with higher emotional intensity (across both Sorry Types).
This is a very small-scale study; therefore, we do not intend to try to interpret these findings as general to a larger population, but rather wanted to provide an overall analysis of the productions that form the basis of the perception experiment. Nevertheless, we can make some observations. First, we found a significant effect of duration, consistent across all speakers. Second, while the other three dimensions were not significant, looking at the by-speaker results suggests that this lack of significance may have different underlying causes for the different dimensions. For intensity and mean F0, none of the speakers showed clear differences, but for pitch contour, we found fairly robust differences within each speaker, but different strategies across speakers. A larger sample size is necessary to determine whether this observation is correct.

3 Experiment 2—Perception of “sorry”

3.1 Overview

In this experiment, listeners completed a forced-choice identification task, classifying tokens of “sorry” produced by voice actors in Experiment 1 as either an Apology or ExcuseMe. The primary purpose of this experiment was to determine whether listeners accurately perceive the speakers’ intended meaning in the absence of context. By examining the acoustic correlates of listeners’ responses, this experiment also allowed for a preliminary investigation of the question of which phonetic cues listeners use in their classification decisions.

3.2 Methods

3.2.1 Participants. 47 listeners residing in Ontario, Canada (7 males and 40 females, age range 18 to 69, mean age 27.5) participated in this experiment. All had learned English as a child (before the age of 10).
3.2.2 Materials. Target stimuli consisted of the 184 tokens of “sorry” analyzed in the production study. These were made up of 144 contextual “sorry” tokens, spliced out of the first line of dialogue shown in Table 1 (recall that the “sorry” was always utterance-initial and followed by a stop or affricate such that the target word was surrounded by silence), as well as 40 tokens produced in isolation.

In addition to these target stimuli, the experiment contained four practice trials and 18 filler trials. Both the practice and filler trials were full sentences with context making it explicit whether the speaker was intending an Apology or ExcuseMe (e.g., Apology: “I’m so sorry for taking advantage of you like that, can you forgive me?” Excuse me: “Sorry, can I just reach past you there for a sec?”). The four practice trials were played at the beginning of the experiment, and the 18 filler trials were interspersed at regular intervals throughout the experiment both in order to ensure that participants were not responding randomly and to break up the monotony of the task.

3.2.3 Procedure. The experiment was run using PsychoPy (Peirce, 2007). Participants were given headphones and were seated in either a soundproof booth (Toronto) or a quiet room (Ottawa). Before the experiment, participants were given an oral explanation of the nature of the experiment. They were first given examples of how the word “sorry” could be used in different ways: to apologize for something, or to get someone’s attention in order to get by (i.e., in place of “excuse me”). Then, they were told that they would be listening to multiple repetitions of the word “sorry” in isolation and asked to decide whether it sounded more like an apology or “excuse me,” indicating their choice via keypress. Short written instructions were also provided at the beginning of the experiment.

The experiment consisted of four practice trials, followed by the 184 target items, randomized by participant, interspersed at regular intervals with the 18 filler trials. The relevant response keys (“a” for Apology and “l” for ExcuseMe) were indicated by stickers on the keyboard, as well as on the experiment presentation screen. Responses were not able to be given until the full sound file was played. The experiment took approximately 15 minutes.

3.2.4 Analysis. We coded listeners’ responses (Apology vs. ExcuseMe), as well as their accuracy in identifying the speaker’s intended meaning. Two analyses were performed. First, logistic mixed-effects models were used to determine whether listeners identified the speakers’ intended meaning at greater than chance level, and whether this differed across Sorry Types or Situational Intensity levels. Second, we examined which acoustic properties of the stimuli were predictive of listeners’ responses, using the acoustic measurements discussed in the production experiment. The lme4 package (Bates et al., 2015) in R (R Core Team, 2019) was used for the statistical analysis.

3.3 Results

3.3.1 Accuracy. Figure 4(a) shows the percentage of time participants accurately classified tokens as Apology vs. ExcuseMe, broken down by intended Sorry Type and by speaker, and Figure 4(b) shows listeners’ responses broken down by Sorry Type and Situational Intensity (excluding the tokens produced in isolation, which did not have a value for Situational Intensity).

Sorry Type: We tested whether listeners performed significantly above chance, and whether performance varied by Sorry Type and speaker. We used a mixed-effects logistic regression model with Accuracy as the response variable and Sorry Type as a fixed predictor (sum-coded, reference level Apology). We also included random intercepts for participant, word, and speaker, as well as by-participant and by-speaker random slopes for Sorry Type. The model estimate for the intercept was significant ($\beta = 0.751, SE = 0.136, z = 5.536, p < .001$), with the positive estimate indicating
that listeners were significantly above chance in this task (64.7% accuracy overall). The main effect for Sorry Type was not significant ($\beta = -0.098$, $SE = 0.486$, $z = -0.202$, $p = 0.84$), as reflected by a lack of clear difference across Sorry Types in Figure 4(a).

**Situational Intensity:** In order to test whether Situational Intensity affected listeners’ responses, we ran a model with only the Contextual stimuli, since the tokens produced in isolation did not have a value for Situational Intensity. The model predicted listeners’ choice of Apology (vs. ExcuseMe) based on the predictor variables of Sorry Type and Situational Intensity (simple-coded, reference level = 1), with random by-participant and by-speaker intercepts and slopes for Sorry Type and Situational Intensity, as well as a random by-word intercept. There was again a significant effect for Sorry Type, reflecting the results above that listeners’ choices matched the actors’ intent ($\beta = -1.312$, $SE = 0.276$, $z = -4.760$, $p < .001$), but Situational Intensity did not have a significant effect on listeners’ responses (Level 2 vs. Level 1: $\beta = 0.055$, $SE = 0.209$, $z = 0.264$, $p = 0.792$; Level 3 vs. Level 1: $\beta = 0.171$, $SE = 0.197$, $z = 0.866$, $p = 0.386$). The intercept of this model was also not significant ($\beta = -0.233$, $SE = 0.249$, $z = -0.935$, $p = 0.350$), indicating that there was no overall bias for either Sorry Type.

In sum, listeners were above chance in classifying Apology vs. ExcuseMe tokens, although performance was still well below ceiling, and there were no systematic differences in accuracy across the two Sorry Types. There was no overall bias in choice of Sorry Type, with listeners choosing Apology 48% of the time.

### 3.3.2 Correlations between responses and acoustic properties

We now turn to the question of which cues listeners use when differentiating between the two Sorry Types, returning to the four acoustic values (duration, mean F0, mean intensity, and pitch contour) described in the production study. For the perception analysis, we used normalized F0 values, scaled to z-scores for each speaker, instead of raw values, since we expect listeners to normalize by speaker. Figure 5 shows how often listeners classified each token as an Apology (vs. ExcuseMe), as a function of each acoustic dimension.

The patterns in Figure 5 suggest that listeners are more likely to perceive tokens as indicating an apology when they are longer in duration, lower in F0, lower in intensity, and characterized by a falling (vs. rising) pitch contour.
In order to determine whether these patterns are significant, we used logistic regression models predicting listeners’ responses from each dimension. Prior to the regression analysis, we examined the interdependencies between the dimensions, testing whether all pairwise relationships were significant using linear regression models, with an alpha-level of 0.05. Three of the six pairwise comparisons showed a significant relationship (F0~intensity: $t = 8.56, p < .001$; F0~pitch contour: $t = 2.59, p = 0.010$; duration~intensity: $t = -3.64, p < .001$) while the other three did not (F0~duration: $t = -0.93, p = 0.355$; duration~pitch contour: $t = -1.06, p = 0.288$; intensity~pitch contour: $t = 1.19, p = 0.235$).

We then analyzed the relationship between acoustics and listener responses using four separate logistic regression models, one for each dimension, parallel to the production analysis. For each model, the binary response variable was participants’ choice of Apology (vs. ExcuseMe), and the predictor variable was one of the four dimensions shown in the graphs above. Continuous variables (duration, normalized F0, and mean intensity) were scaled to z-scores prior to analysis, and pitch

Figure 5. Percentage Apology (vs. ExcuseMe) choice, as a function of duration, F0 (normed), mean intensity, and final pitch contour. In the first three panels, each point represents one stimulus, and in the final panel, boxplots for pitch contour show the distribution of responses for falling versus rising tokens. Best-fit linear regression lines are shown for the continuous dimensions.
Table 3. Results of logistic mixed-effects regression models predicting Apology choice from four acoustic dimensions. The model structure (example given for duration) is \( \text{glmer(choice} \sim \text{duration} + \text{(duration} \mid \text{participant}) + \text{(duration} \mid \text{speaker}), \text{family=binomial}) \). Significant (\( p < .05 \)) results are bolded.

| Dimension            | Parameter  | \( \beta \) | SE  | \( z \)  | \( p \)  |
|----------------------|------------|-------------|-----|---------|---------|
| Duration             | Intercept  | -0.04       | 0.24| -0.18   | 0.859   |
|                       | Duration   | **0.99**    | **0.08** | **12.40** | **< .001** |
| Mean F0              | Intercept  | -0.11       | 0.18| -0.62   | 0.538   |
|                       | F0 (normed)| -0.26       | 0.16| -1.62   | 0.105   |
| Mean intensity       | Intercept  | -0.02       | 0.15| -0.14   | 0.888   |
|                       | Intensity  | **-0.47**   | **0.12** | **-3.85** | **< .001** |
| Final pitch contour  | Intercept  | -0.30       | 0.24| -1.22   | 0.224   |
|                       | Pitch contour | **-0.53**   | **0.18** | **-2.91** | **0.004** |

contour was simple-coded, with “falling” as the reference level. Each model also included a random by-participant and by-speaker intercept, as well as random by-participant and by-speaker slopes for the fixed predictor (e.g., duration). Statistical results are shown in Table 3.

The estimates for each main effect represent the change in log odds of an Apology response for a one-unit change in the predictor variable: in the case of the continuous predictor variables, which were scaled to z-scores, this represents a one-standard-deviation change, and for pitch contour, this represents the difference between falling and rising pitch contour. The results show that duration, mean intensity,\(^4\) and pitch contour are significantly predictive of listeners’ responses, but mean F0 was not.

These results suggest that multiple acoustic dimensions may inform listeners’ decisions about the meaning of “sorry.” Given that these dimensions were correlated with one another in the natural productions, it is not possible to make claims about the independent role of any single dimension; for example, since duration and intensity are correlated, it could be that listeners are only using duration, but not intensity, with the apparent use of intensity being an artifact of the fact that lower-intensity tokens (which were more likely to be classified as Apology) tend to have longer durations (and were also more likely to be classified as Apology than shorter tokens). Nevertheless, this analysis suggests that one or more of these dimensions is used by listeners; in the final experiment, we manipulate the dimensions independently in order to tease apart their individual contributions to listeners’ decisions.

4 Experiment 3—Perception of systematically manipulated tokens of “sorry”

4.1 Overview

The results of Experiment 2 showed that listeners’ perception of a speaker’s intended meaning of “sorry” is affected by one or more of the acoustic dimensions measured in Experiment 1. The goal of this experiment was to tease apart the independent role of these dimensions: which cues are used, and what is the relative reliance on each cue? We approach this question by examining listeners’ perception of “sorry” tokens that have been systematically manipulated to vary along each dimension.

4.2 Methods

4.2.1 Participants. 47 listeners residing in Ontario, Canada (17 males and 30 females, age range 18 to 73, mean age 27.2) participated in this experiment. All participants had learned English as a child (before the age of 10).
4.2.2 Materials. The stimuli for this experiment again consisted of individual tokens of “sorry” in isolation. Stimuli were created from 12 baseline tokens that had been used as stimuli in Experiment 2. These baseline tokens were subsequently manipulated on the three acoustic dimensions found in Experiment 2 to play a role in predicting listeners’ responses: duration, intensity, and pitch contour.

**Baseline tokens:** Baseline tokens, a subset of the stimulus set of Experiment 2, were selected to serve as the basis for subsequent manipulations. To increase generalizability and to allow for the fact that there are almost certainly other dimensions beyond those we were manipulating that affect listeners’ perceptions, we selected 12 baseline tokens, three from each of the four speakers in Experiment 1. In order to maximize the range of variability in the natural tokens, we chose the token from each speaker that had elicited the highest, lowest, and most ambiguous Apology responses in Experiment 2 (across the four speakers, these tokens averaged 94%, 9%, and 51% Apology responses respectively). Each of these 12 tokens then served as the baseline for the manipulations described below. A summary of the stimulus set used in this experiment is given in Table 4.

**Pitch contour manipulation:** We created two stylized pitch contours, one with a falling and one with a rising tone. These two contours (which approximate HLH% and HLL% contours in the ToBI system, Beckman & Ayers, 1997), were chosen as models for manipulation because they were the most common contours seen in the data. The parameters for the stylized contours were chosen by trial and error, with the goal of creating contours that (1) approximated the patterns seen in the production data and (2) sounded natural to native speakers of English. An example of the manipulation is shown in Figure 6.

Pitch contour manipulations were done using the “Manipulation” interface in Praat. Each stylized pitch contour was created by setting pitch points at three landmarks: the first at the end of the

---

**Table 4. Summary of factors varying in the stimuli for Experiment 3.**

| Dimension          | Num. levels | Levels                                                                 |
|--------------------|-------------|------------------------------------------------------------------------|
| **Baseline tokens**| Speaker 4    | The four speakers from Experiment 1                                     |
|                    | Base token 3 | Heard in Exp2 as Apology vs. ExcuseMe vs. ambiguous                   |
| **Manipulations**  | Pitch contour 2 | Falling, rising                                                         |
|                    | Intensity 2   | 65 dB, 75 dB                                                            |
|                    | Duration 5    | 350 ms, 475 ms, 600 ms, 725 ms, 850 ms                                |
| **Total tokens**   | 240          |                                                                        |

**Figure 6.** Examples of falling (a) and rising (b) stylized pitch contours. The visible pitch range is 150 to 400 Hz (audio files available in the supplementary materials).
first vowel of “sorry,” the second at the onset of the second vowel, and the final one at the end of the word. Contours are shown in Figure 6. For the falling contour, the pitch point fell eight semitones from the first to the second landmark, then fell an additional two semitones to the third landmark. For the rising contour, the pitch point fell five semitones from the first to the second landmark, then rose five semitones to the third landmark. Identical contours were superimposed on all baseline tokens, but the raw pitch values were speaker-specific: the first landmark was always set as the speaker’s average F0 value at the 10% point across the full production dataset. The pitch contour manipulation resulted in two pitch files for each of the 12 baseline tokens (24 stimuli).

**Intensity manipulation:** Using the “Scale Intensity” function in Praat, we created two levels of intensity: 65 and 75 dB for each of the previously manipulated tokens, resulting in 48 tokens. These values were chosen to be centered around 70 dB and to allow for a perceptible difference in volume while remaining within a natural-sounding range so as not to be overly salient or distracting.

**Duration manipulation:** Since results from both Experiments 1 and 2 suggested an important role for duration, we chose to create a five-step continuum to allow for more precision in the analysis. The endpoints were set at 350 ms and 850 ms, based on the range of duration in the production dataset (after removing some apparent outliers). Using the PSOLA-based manipulation algorithm in Praat (Moulines & Charpentier, 1990), we manipulated each of the previously created 48 tokens to have each of the five duration values (350, 475, 600, 725, and 850 ms), resulting in a final total of 240 stimuli.

### 4.2.3 Procedure

The procedure was identical to that of Experiment 2, except the 240 manipulated tokens were used as target stimuli in lieu of the 184 naturally produced target items from Experiment 2. As in Experiment 2, 18 filler trials consisting of complete sentences with context were interspersed at regular intervals, and the order of the 240 target tokens was randomized by participant.

### 4.2.4 Analysis

We used a logistic mixed-effects model to model listeners’ response (Apology vs. ExcuseMe) as a function of the manipulated acoustic dimensions. Since the dimensions were manipulated independently and therefore uncorrelated, we were able to analyze all predictors in a single model. The response variable was listeners’ choice of Apology (vs. ExcuseMe), with fixed predictors of Pitch Contour (falling vs. rising), Duration (350–850 ms), Intensity (65 dB vs. 75 dB), and Base Type (elicited mostly ExcuseMe vs. mostly Apology vs. ambiguous responses in Experiment 2), along with random by-participant intercepts and slopes for all the fixed predictors, and random by-speaker intercepts. No interactions were included. Duration was centered and analyzed as a continuous predictor. Categorical predictors were sum-coded (reference levels in italics above). As above, the lme4 package (Bates et al., 2015) in R (R Core Team, 2019) was used for the statistical analysis.

### 4.3 Results

Listeners’ responses as a function of each manipulated dimension are shown in Figure 7, and statistical results are given in Table 5. As shown by the non-significant intercept in the model, there was no overall bias for Apology vs. ExcuseMe response in this dataset. We discuss the effect of each dimension in turn in the following paragraphs.

**Base token:** As shown in Figure 7, listeners were more likely to identify tokens created from baselines that elicited more Apology responses in Experiment 2 as an apology, even when all of the dimensions considered in this work were controlled for by manipulating in the same way for all
Figure 7. Listeners’ choice of Apology (vs. ExcuseMe) as a function of (a) base token, (b) duration, (c) mean intensity, and (d) pitch contour. Violin plots are based on the distribution of by-participant means; error bars represent 95% confidence intervals of these by-participant means.

Table 5. Results of logistic mixed-effects regression models predicting Apology choice from baseline token, duration, intensity, and pitch contour. Reference levels are in italics. The model structure is glmer(choice ~ base + duration + intensity + pitch.contour + (base + duration + intensity + pitch.contour | participant) + (1 | speaker), family=binomial). Significant (p < .05) results are bolded.

|                          | β   | SE  | z     | p     |
|--------------------------|-----|-----|-------|-------|
| Intercept                | −0.22 | 0.44 | −0.50 | 0.617 |
| Base (AMBIG. vs EXC.)    | 0.14 | 0.08 | 1.80  | 0.072 |
| **Base (APOL. vs. EXC.)**| **0.75** | **0.09** | **8.00** | **< .001** |
| Duration                 | 1.00 | 0.09 | 11.59 | **< .001** |
| Intensity (75 dB vs. 65 dB) | −0.15 | 0.06 | −2.63 | 0.008 |
| Pitch contour (rising vs. falling) | −0.72 | 0.17 | −4.28 | **< .001** |
baseline tokens. The statistical results support this, showing that baseline Apology tokens elicited more Apology responses (53%) than baseline ExcuseMe tokens (41%). Baseline ambiguous tokens were intermediate (43%) but were not significantly different from ExcuseMe.

**Duration**: Reflecting the results of Experiment 2, as well as the production study, there is a remarkably systematic relationship between duration and listeners’ Apology responses, with longer tokens eliciting significantly more Apology responses: looking at the endpoints, the shortest duration (350 ms) elicited on average 20% Apology responses, whereas the longest duration (850 ms) elicited 66% Apology responses.

**Intensity**: As in Experiment 2, tokens with lower intensity (i.e., softer tokens) elicited more Apology responses (47%) than louder tokens (44%), and although small, this difference was significant.

**Pitch contour**: As in Experiment 2, tokens with a falling contour were more likely to be considered as an Apology than those with a rising contour (52% for falling vs. 39% for rising contour), and this difference was significant.

Overall, significant differences were found for all factors tested, and paralleled the results found in Experiment 2. This suggests an independent role for each of these dimensions: tokens that were longer in duration, lower in intensity, and with a final falling pitch tended to be classified more often as Apology than ExcuseMe. This was also the case for tokens that elicited more Apology responses in Experiment 2, suggesting there are cues other than those manipulated here that inform listeners’ decisions.

We also wanted to assess the relative importance of these cues in predicting listeners’ responses. While the question of how to quantify cue use is complex (see Schertz & Clare, 2020, for discussion), one metric for how predictive a cue is in this sort of paradigm is simply by assessing the difference in predicted responses across the range of each dimension. These differences are summarized in Table 6. Based on this metric, duration is the most predictive of the factors examined in this work, followed by pitch contour, followed by difference in base token, followed by intensity (which only elicited a 3% difference in responses). However, it should be kept in mind that this can only be interpreted within the stimulus set used in this experiment. With different parameter values (e.g., a larger difference between the two intensity values), the apparent “use” of the cue could differ.

Overall, the results of Experiment 3 strengthen the findings of Experiment 2, suggesting that multiple cues inform listeners’ perception of Sorry Type: using manipulated stimuli that varied orthogonally on the dimensions of interest, this experiment provided evidence for an independent

| Factor       | Levels       | % Apology Response | Difference |
|--------------|--------------|--------------------|------------|
| Duration     | 350 ms       | 20%                | 46%        |
|              | 850 ms       | 66%                |            |
| Pitch contour | falling      | 52%                | 13%        |
|              | rising       | 39%                |            |
| Base token   | ExcuseMe     | 41%                | 12%        |
|              | Apology      | 53%                |            |
| Intensity    | 65 dB        | 47%                | 3%         |
|              | 75 dB        | 44%                |            |

**Table 6.** Comparison of differences in percentage Apology response across the levels (or extreme values, in the case of continuous factors or factors with more than two levels) of each factor.
role for each of these dimensions. Specifically, tokens showing longer duration, lower intensity, and falling pitch are more likely to be classified as apologies, and that, at least in the acoustic space of the stimulus set, duration exerts the greatest influence. Furthermore, the baseline token played a role, with tokens created from baselines that had been perceived as Apology in their natural form eliciting more Apology responses even with duration, intensity, and pitch controlled, indicating the presence of other information in the signal that is relevant to the perception of this distinction.

5 Discussion

5.1 Summary of results

This work examined the use of multiple acoustic dimensions to cue different meanings of the word “sorry” in production and perception. In Experiment 1, tokens of “sorry” produced by voice actors differed in duration based on whether they were in the contexts suggesting an Apology versus an ExcuseMe, with longer durations for Apology tokens. Duration was also longer in contexts of greater situational intensity for both Sorry Types. No significant differences in other dimensions (mean F0, intonation contour, intensity) were found, although given the small number of talkers, these null results cannot be interpreted as a strong claim for a lack of difference in the general population. In Experiment 2, listeners were able to identify the actors’ intended meaning (Apology vs. ExcuseMe) at above-chance levels (64.7% accuracy overall), and their responses were correlated with duration, intonation contour, and intensity (but not mean F0). Finally, we tested whether each of these cues contributed independently to listeners’ judgments in Experiment 3, where listeners identified tokens drawn from a controlled acoustic space varying systematically along the relevant dimensions. All three manipulated dimensions did contribute independently to listeners’ response patterns: duration had the largest effect of the three manipulated cues, with longer duration cuing Apology, and with pitch contour and intensity (where a final F0 fall and lower intensity cued Apology) playing smaller but still significant roles. The baseline token from which the tokens were created also influenced listeners’ responses in the expected direction, indicating that other dimensions other than those that were manipulated also play a role. Not only were all manipulated dimensions significant, but they also explained a fairly large amount of the variability: tokens with short duration, rising final tone, and high intensity (collapsing over all baseline tokens) elicited 15% Apology response, whereas those tokens with long duration, falling final tone, and low intensity elicited 73% Apology response.

While listeners’ use of duration reflected the distribution of durations in the production dataset, two of the other cues (intonation contour and intensity) were found to influence listeners’ judgments, despite the fact that there was no significant difference in the values of these two dimensions across the two Sorry Types in production. On the surface, this may appear to be a case of acoustic cues being used in perception even when they are not informative in production (e.g., Warner et al., 2004). However, we think it is more plausible that these cues are indeed informative in the input, but that our production study simply did not have enough power, or was not representative enough, to reflect the true input. Our sample size was quite small, and given the usually very small size of subphonemic differences, we expect that a larger sample size would be necessary to detect an effect. Furthermore, we used a nonrepresentative sample of talkers (voice actors) and a non-naturalistic production setting (asking actors to read a script), such that it is difficult to assess the extent to which this reflects the distribution of productions that listeners would hear in real life. Further studies using corpus work, or more naturalistic production experiments with larger, more representative samples of speakers, are necessary to answer questions about the relationship between the distribution of cues in the input and listeners’ use of these cues in perception.
5.2 Use of individual cues

In the following sections, we discuss the findings for each acoustic dimension in the context of previous work, and we lay out potential explanations for their use.

5.2.1 Duration. Given the many factors shown to influence duration, it is perhaps unsurprising to find a durational difference between the two forms in production, with Apology tokens being on average 126 ms longer than ExcuseMe tokens. We do not have independent lexical statistics for the two meanings of “sorry,” so we did not have specific predictions about durational differences based on factors shown to affect duration in previous work (e.g., frequency: Gahl, 2008; Lohmann, 2018b; or contextual predictability: Seyfarth, 2014). Given that these effects are well established, we expect that they could contribute to the difference found in these productions. Systematic differences in prosodic phrasing may provide another explanation for the difference: as pointed out by a reviewer, all of our Apology tokens were followed by an exclamation point, while ExcuseMe tokens were followed by a comma. This may have made it more likely for speakers to produce the apology tokens as an independent intonational phrase, and therefore lengthened. While this was not an intentional element of our design, we speculate that this difference reflects the typical prosodic environments of these meanings in real-world discourse, and that these results are not simply an artefact of punctuation. The perception results showing listeners’ use of duration are consistent with the idea that these prosodic regularities are linked to each sense, either via a direct link between each sense and its typical prosodic realization, or via an indirect link in which each sense is linked to an abstract prosodic context.

We also found systematic durational variability within each meaning as a function of situational intensity. The fact that higher-intensity situations (with identical utterances) elicited longer durations provides evidence that factors unrelated to lexical statistics or prosodic structure must play a role, and it suggests that emotion/affect exerts an independent influence on production.

5.2.2 Pitch contour. There were no statistically significant differences between the pitch contours of the two Sorry Types in production, but this lack of result needs to be interpreted with particular caution for several reasons. First, our analyses were based on perceptual judgments using the ToBI coding system. These judgments can differ substantially across coders, even in “best-case” scenarios with highly trained coders and clear speech (e.g., Syrdal & McGory, 2000). We only reported tokens that were agreed-upon by our two coders, which reduces the amount of data (although the results were the same when using the full dataset of each coder individually). Second, the different speakers in our study showed different patterns, with three of the four speakers showing more falling tones for Apology productions, and the other showing the opposite. These speaker-specific production strategies may have obscured the group-level results, and point to the need to look at a wider range of speakers, as well as the specific phonetic realizations of the production patterns, to get an accurate view of how this difference is manifested in production.

While it is not possible to make strong generalizations about how this contrast is realized in production, we did see a clear influence of intonation contour in both perception experiments, with final falling tone eliciting more Apology responses. We again consider potential sources of this difference. Intonation contour has been found to be an acoustic correlate for different emotions: for example, Pereira and Watson (1998) found that falling contours are acoustic correlates of sad utterances. On a linguistic level, F0 as a source of subphonemic variation has been less well studied than duration, but Tang and Shaw (2020) found F0 differences paralleling the contextual-predictability effects on duration discussed above: items that tend to be more predictable have lower F0 peaks. While it is possible that this factor is at play, this is a different question than categorically different intonation
language and speech 65(1)

contours, which is what we were testing. We think the most likely explanation is that there may be systematic differences in the contour used for Apology vs. ExcuseMe tokens, and this may be due to pragmatic and/or linguistic context. A corpus or larger-scale production experiment is necessary to test whether this is the case.

It might seem surprising that the overall F0 level (mean F0) was not found to play a role in either production or perception, given the fact that it has been a relatively consistent feature found to vary based on emotion (Banse & Scherer, 1996, among others). Furthermore, we might expect Apology tokens to be lower in F0, as they presumably tend to be more submissive and less assertive than attention-seeking ExcuseMe tokens, based on Ohala’s (1983) frequency code. However, when there are overall differences in intonation contour, as we saw in the current set of tokens analyzed in production and used for perception, the intonation-based variability in F0 would make it difficult to detect a difference in overall mean F0 even if it existed. Furthermore, we did not test an independent effect of mean F0 in Experiment 3. More sensitive tests would be necessary to make strong claims about the absence of an overall F0 effect.

5.2.3 Intensity. Intensity was found to play a small but significant role in both perception experiments, with lower intensity eliciting more Apology responses, but no effect was found in production. As with F0, Tang and Shaw (2020) found decreased intensity corresponding to greater contextual predictability, and based on the fact that lexical frequency is also expected to result in overall lower prominence, decreased intensity might also be expected in lower-frequency forms. At the same time, lower intensity has been shown to be associated with sadness (Pereira & Watson, 1998) and lack of assertiveness/dominance (Puts et al., 2007), both of which may be expected in Apology compared to ExcuseMe productions. Therefore, as with the other dimensions, both lexical and non-lexical factors potentially contribute to the current results.

5.3 Implications for models of speech perception and processing

Listeners’ systematic use of multiple phonetic dimensions to differentiate two polysemous uses of “sorry” (Apology vs. ExcuseMe) indicates that these two senses must be associated with different representations at some level of processing. Previous work has presented evidence that the different senses of polysemes share a semantic representation, whereas the different (unrelated) meanings of homophones have independent semantic representations (Rabagliati & Snedeker, 2013; Rodd et al., 2002). Although our results point to separate representations for the different senses of the polysemous “sorry,” we do not think that they are inconsistent with these previous findings. The evidence for shared senses in Rabagliati and Snedeker (2013) was for very closely related, “regular” polysemes (e.g., “chicken” the animal vs. “chicken” the food); in a separate condition testing less closely related, “irregular” polysemes (e.g., “sheet of glass” vs. “drinking glass”), results suggested separately stored meanings. The two meanings of “sorry” used in this case study fall toward the less-related end of the continuum (see Moldovan, 2019, for definitions and discussion of the gradient nature of the polysemy–homophony continuum). Although we are not aware of any previous work along these lines, this brings up the interesting possibility that evidence of listeners’ systematic phonetic differentiation of different senses/meanings could potentially be used as a supporting diagnostic for separate semantic representations (taking into account the extent of phonetic differences that actually exist in the production of a given word-pair).

It is clear from our results that the two senses of “sorry” must be independent at some level of representation. However, this study cannot provide a definitive answer to how sense-specific phonetic information is represented and linked to the two meanings. The fact that listeners differentiate the two senses based on acoustic information could be explained straightforwardly by each sense
having an independently specified phonetic representation (e.g., a separate phonetic prototype or exemplars associated with each sense). Under this view, listeners’ choice of Apology vs. ExcuseMe would be determined via direct comparison with the phonetic representations of each sense (e.g., “This production had a long duration, and Apology tokens are characterized by longer duration”). However, another possibility is that the two senses could also be linked to a single, shared phonetic representation. Under this model, each sense could also be associated with distinct “contexts,” which are in turn associated with systematic phonetic properties, as discussed in the Introduction. When distinguishing between the two senses, the phonetic information would inform listeners’ judgment about the context, which would then inform their choice of category (e.g., “This production had a long duration, which means it’s likely to have been produced in a single prosodic phrase, and Apology tokens are more likely to be produced in a single prosodic phrase”).

The findings that listeners use a given dimension is consistent with either of these scenarios. More broadly, an accurate model of speech perception likely includes both components as a means to represent pronunciation variation. As discussed in Pierrehumbert (2016), there is evidence that (at least some) subphonemic detail must be able to be stored independently for (at least some) lexical representations. However, more abstract knowledge of phonetic patterns also influences lexical decisions. For example, Shatzman and McQueen (2006) found that listeners used duration information to distinguish short versus embedded novel words (e.g., “bap” vs. “baptoe”) just as they did for real words, even though durational information across the two syllables was identical during an exposure phase. If the representations of these new words were simply made up of acoustic exemplars, duration would not be an informative cue for listeners, and the fact that listeners were using it indicates that listeners were drawing on more general or abstract information. The fact that listeners are sensitive to phonetic properties of emotion or affect, even in non-words, provides further support for use of non-lexically-specific phonetic detail. Taken together, it appears that the appropriate question is not whether lexical representations are purely abstract or fully phonetically specified, but rather how word-specific and abstract components are integrated during perception and processing (see Ernestus, 2014, for discussion).

5.4 Lessons from a case study

As discussed in the sections above, there are several plausible explanations for the effects found in this work. It is likely that differences in prosodic phrasing play a large role across several of the dimensions examined here; this was not one of our primary considerations when designing the study, but its importance was highlighted by reviewers of an initial version of this work. For example, an Apology may be more likely to be produced in a single intonational phrase than an ExcuseMe, affecting both duration and F0, and/or the two senses might be associated with different intonation contours. In either case, listeners’ use of these dimensions could then be explained by a model of perception which incorporates prosodic knowledge as part of the process of lexical competition (e.g., a “Prosody Analyzer,” as proposed by Cho et al., 2007), or by independently specified intonation contours for each sense (e.g., Calhoun & Schweitzer, 2012; Tang & Shaw, 2020). Emotion or affect could also be a contributing factor: tokens identified as Apology by listeners were characterized by features found to correspond to sadness (falling F0 contour, lower intensity), and the fact that the situational intensity exerted an independent influence on duration suggests a role for emotion/affect (though it is not necessarily sadness per se). Finally, although difficult to evaluate given the lack of frequency measures for the two different senses, the role of frequency also likely plays a role, given the findings of previous work (e.g., Lohmann, 2018b). We think it most likely that our results indicate listeners’ tracking and use of the systematic variation conditioned by both linguistic and affective factors. However, in the presence of multiple possible
interpretations, it is not possible to decide between them in a case study using a single word. Instead, determining the relative roles needs to be investigated separately in studies with multiple words.

The effects found in this case study were quite large in comparison to previous work. For example, consider the 126 ms average difference between Apology and ExcuseMe tokens in the production study, compared to a vowel duration difference of 3.5 or 10 ms for differences based on underlying voicing in Dutch and German (Port & O’Dell, 1985; Warner et al., 2004), or an average of 15 ms and 21 ms difference between high- versus low-frequency homophones analyzed in Gahl (2008) and noun–verb homophone pairs (Lohmann, 2018b). Listeners’ accuracy in correctly identifying the intended sense for naturally produced tokens was also quite high compared to very small effects in previous work (just over 50% in Sanker, 2019), and in our manipulated task, a large amount of the variance was explained by our manipulations, as opposed to very weak correlations between acoustic features and listeners’ responses in previous work (Sanker, 2019; Drager, 2010).

There are multiple possible reasons for the relatively large effects found here. First, the studies about durational differences cited above were examining homophonous contrasts which differed along scalar variables (e.g., frequency, contextual predictability) or which were actually expected to be phonologically neutralized. In contrast, the senses of “sorry” used in the current study are commonly used in social routines. Their frequency and the social context of their use may form the basis of a robust contrast in production, and/or greater sensitivity to these differences by listeners; indeed, these are the types of words proposed by Calhoun and Schweitzer (2012) to be the most likely to have lexicalized or idiomatic intonational contours. If so, this would account for the particularly strong listener judgments. Second, as discussed at the beginning of this section, there are multiple factors, including prosodic regularities, emotion/affect, and lexical statistics, that could potentially play additive roles. Determining the relative roles of each of these factors is only possible through larger-scale studies examining multiple words; however, the strength of the effects in the current work highlights the importance of examining word-specific effects in these larger-scale studies. This is crucial because a particularly robust contrast for a single word could skew group-level effects, but also because examining the properties of words which have stronger/weaker differences could shed light on which factors are most important.

Finally, our findings from Experiment 3 show an independent contribution of different acoustic dimensions, and the pattern of responses to the five-step duration continuum indicates a gradient use of this cue (the other dimensions only had two levels, so we are unable to make observations about gradience). This suggests that dimensions are stored and tracked independently. The question of which dimensions are tracked, and the details of their use, is another factor that needs to be considered in models of speech perception.

5.5 Limitations and future directions

This case study of a single word pair allows for simultaneous investigation of listeners’ use of multiple acoustic dimensions, and for a comparison of perception of naturally produced and manipulated tokens. The results highlight the fact that listeners rely on several acoustic dimensions to inform their perception of these words, and while the design of this study does not allow for a definitive answer about why these cues are used, it is plausible that the regularities underlying the use of these cues stem from both lemma-specific and general contextual factors.

Any affective/emotive elements of the utterances produced in our production study are simulated, due to the fact that they are based on reading of scripted productions by voice actors. We make the assumption, following Banse and Scherer (1996) and many other researchers, that these utterances contain properties of “real-life” speech, and we believe that this assumption was supported by the findings that listeners did indeed identify the intended meanings at above-chance
levels. Nevertheless, as with any laboratory-based study, we do expect that there will be systematic differences between lab-based and spontaneous productions. In this case, we might expect phonetic distinctions to be exaggerated, because of the read speech, because of the use of voice actors, and because the word of interest was likely clear to them. In the future, corpus work could be used to test to what extent the patterns found here are reflected in naturalistic speech.

As with any case study, this is a starting point for formulating and testing more general hypotheses about models of speech processing and cue use. One question, in terms of the perception–production interface, is how faithfully cue use mirrors input distributions. This has been examined quite frequently in terms of perceptual cue-weighting for phonetic categorization (see Schertz & Clare, 2020, for discussion), the question being to what extent the relative informativity of various cues in distinguishing two members of a phonetic contrast predicts listeners’ relative reliance on these cues in perception. Expanding the scope of this inquiry to lemma-level lexical contrasts can help provide information about the level on which listeners are computing input statistics, and how these statistics are used. A second, related question is which factors underlie the use of these cues. As discussed above, in the perception of a given word meaning, listeners likely draw on both lemma-specific knowledge (i.e., information encoded in the lexical entry) and their knowledge of the contexts in which the word meaning is likely to occur, as well as the phonetic regularities associated with those contexts. The range of potential “contexts” is vast, encompassing syntactic, prosodic, discourse, and emotional domains. The extent to which listeners track statistics along these dimensions, and whether they do it in similar ways for different contexts, has, to our knowledge, been largely unexplored.

Moving forward in these areas will require work on two methodological fronts: first, corpus work and/or larger-scale elicitation-based production studies should be done to get a better idea of the distribution of cues in the input as a function of a broader range of factors. Second, perception patterns should be explored in controlled experimental work using comparable methodologies as previous work to tease apart and quantify the relative roles of word-specific versus abstract information, considering a broader range of contexts over which listeners might generalize (e.g., emotion/affect). While these are broad and complex questions, answering them will help build accurate and computationally viable models of how information is stored and used.

Acknowledgements
We would like to thank Yoonjung Kang, the Associate Editor Holger Mitterer, and two anonymous reviewers for helpful comments on earlier drafts of this work. We would also like to thank Olivia Marasco for help with prosodic annotations.

Funding
The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was conducted with the financial support of the Joseph Armand Bombardier Canada Graduate Scholarship, awarded to Caitlyn Martinuzzi by the Social Sciences and Humanities Research Council of Canada.

ORCID iD
Jessamyn Schertz https://orcid.org/0000-0003-2492-8498

Supplementary material
Supplementary material for this article, including the full dataset (acoustic measurements from production study and listener responses from perception studies) and sample sound files, is available at https://osf.io/m9qya/.
Notes
1. In this work, we use the word “homophone” to encompass polysemous words (i.e., words with distinct but related senses) as well as “true” homophones (i.e., words whose meanings are unrelated); we address this distinction and the potential relevance for our study below.
2. The word “sorry” was chosen because (1) it is frequently used and encountered (particularly in the Canadian context of this experiment); (2) it has distinct meanings that we expected to be easily recognizable and explainable to participants; and (3) according to the impressionistic judgment of the authors, we thought it likely that the uses would be phonetically distinct in production.
3. Thanks to Holger Mitterer and an anonymous reviewer for these suggestions.
4. In response to a question from an anonymous reviewer, we confirmed that the relationship between responses and intensity was still significant when the lowest-intensity token was removed ($\beta = -0.54$, $SE = 0.12$, $z = -4.62$, $p < .001$).

References
Bachorowski, J. A., & Owren, M. J. (2008). Vocal expressions of emotion. In M. Lewis, et al. (eds.), Handbook of emotions (pp. 196–210). Guilford Press.
Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. Journal of Personality and Social Psychology, 70(3), 614–636.
Bänziger, T., & Scherer, K. R. (2005). The role of intonation in emotional expressions. Speech Communication, 46(3–4), 252–267. https://doi.org/10.1016/j.specom.2005.02.016
Bates, D., Maechler, M., Bolker, B. M., & Walker, S. C. (2015). Lme4: Linear mixed-effects models using eigen and S4. https://cran.r-project.org/web/packages/lme4/index.html
Beckman, M. E., & Ayers, G. (1997). Guidelines for ToBI labelling. The OSU Research Foundation, 3, 30.
Bell, A., Brenier, J. M., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. Journal of Memory and Language, 60(1), 92–111.
Boersma, P., & Weenink, D. (2018). Praat: Doing phonetics by computer [Computer program]. Version 6.0.43. http://www.praat.org/
Bond, Z. S. (1973). The perception of sub-phonemic phonetic differences. Language and Speech, 16(4), 351–355.
Calhoun, S., & Schweitzer, A. (2012). Can intonation contours be lexicalised? Implications for discourse meanings. In G. Elordieta Aclibar & P. Prieto (eds.), Prosody and meaning (pp. 271–328). Interface Explorations.
Cho, T., McQueen, J. M., & Cox, E. A. (2007). Prosodically driven phonetic detail in speech processing: The case of domain-initial strengthening in English. Journal of Phonetics, 35(2), 210–243.
Conwell, E. (2018). Token frequency effects in homophone production: An elicitation study. Language and Speech, 61(3), 466–479.
Davis, M. H., Marslen-Wilson, W. D., & Gaskell, M. G. (2002). Leading up the lexical garden path: Segmentation and ambiguity in spoken word recognition. Journal of Experimental Psychology: Human Perception and Performance, 28(1), 218.
De Bruin, A., Treccani, B., & Della Sala, S. (2015). Cognitive advantage in bilingualism: An example of publication bias?. Psychological Science, 26(1), 99–107.
Drager, K. K. (2010). Sensitivity to grammatical and sociophonetic variability in perception. Laboratory Phonology, 1(1), 93–120.
Drager, K. K. (2011). Sociophonetic variation and the lemma. Journal of Phonetics, 39(4), 694–707.
Ernestus, M. (2014). Acoustic reduction and the roles of abstracts and exemplars in speech processing. Lingua, 142, 27–41.
Gahl, S. (2008). Time and thyme are not homophones: The effect of lemma frequency on word durations in spontaneous speech. Language, 84(3), 474–496.
Gravano, A., Benus, S., Hirschberg, J. B., German, E. S., & Ward, G. (2008). The effect of contour type and epistemic modality on the assessment of speaker certainty. Speech Prosody. https://doi.org/10.7916/D8J96FVZ
Hirschberg, J. (2002). Communication and prosody: Functional aspects of prosody. *Speech Communication, 36*(1–2), 31–43.

Johnstone, T., & Scherer, K. R. (2000). Vocal communication of emotion. In M. Lewis & J. Haviland (eds.), *Handbook of emotions* (pp. 220–235). Guilford.

Jurafsky, D., Bell, A., & Girard, C. (2002). The role of the lemma in form variation. In C. Gussenhoven, et al. (eds.), *Papers in Laboratory Phonology, 7*, 1–34.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*(13).

Lohmann, A. (2018a). Cut (n) and cut (v) are not homophones: Lemma frequency affects the duration of noun–verb conversion pairs. *Journal of Linguistics, 54*(4), 753–777.

Lohmann, A. (2018b). Time and thyme are not homophones: A closer look at Gahl’s work on the lemma-frequency effect, including a reanalysis. *Language, 94*(2), e180–e190.

McQueen, J. M., Cutler, A., & Norris, D. (2006). Phonological abstraction in the mental lexicon. *Cognitive Science, 30*(6), 1113–1126.

Moldovan, A. (2019). Descriptions and tests for polysemy. *Axiomathes*, 1–21.

Moulines, E., & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication, 9*, 453–467.

Nygaard, L. C., & Lunders, E. R. (2002). Resolution of lexical ambiguity by emotional tone of voice. *Memory & Cognition, 30*(4), 583–593.

Ohala, J. J. (1983). Cross-language use of pitch: An ethological view. *Phonetica, 40*(1), 1–18.

Peirce, J. W. (2007). PsychoPy—psychophysics software in Python. *Journal of Neuroscience Methods, 162*(1–2), 8–13.

Pereira, C., & Watson, C. (1998). Some acoustic characteristics of emotion. In *Fifth International Conference on Spoken Language Processing*.

Pierrehumbert, J. B. (2016). Phonological representation: Beyond abstract versus episodic. *Annual Review of Linguistics, 2*, 33–52.

Port, R. F., & O’Dell, M. L. (1985). Neutralization of syllable-final voicing in German. *Journal of Phonetics, 13*(4), 455–471.

Puts, D. A., Hodges, C. R., Cárdenas, R. A., & Gaulin, S. J. (2007). Men’s voices as dominance signals: Vocal fundamental and formant frequencies influence dominance attributions among men. *Evolution and Human Behavior, 28*(5), 340–344.

R Core Team. (2019). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. http://www.R-project.org

Rabagliati, H., & Snedeker, J. (2013). The truth about chickens and bats: Ambiguity avoidance distinguishes types of polysemy. *Psychological Science, 24*(7), 1354–1360.

Rodd, J., Gaskell, G., & Marslen-Wilson, W. (2002). Making sense of semantic ambiguity: Semantic competition in lexical access. *Journal of Memory and Language, 46*(2), 245–266.

Salverda, A. P., Dahan, D., & McQueen, J. M. (2003). The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. *Cognition, 90*(1), 51–89.

Sandra, D., Brysbaert, M., Frisson, S., & Daems, F. (2001). Paradoxen van de Nederlandse werkwoordspelling: Een confrontatie tussen taalkundige logica, problemen voor spellers en bruikbaarheid voor lezers. *De Psycholoog, 30*(3), 282–287.

Sankar, C. (2019). Effects of lexical ambiguity, frequency, and acoustic details in auditory perception. *Attention, Perception, & Psychophysics, 81*(1), 323–343.

Sauter, D. A., Eisner, F., Calder, A. J., & Scott, S. K. (2010). Perceptual cues in nonverbal vocal expressions of emotion. *The Quarterly Journal of Experimental Psychology, 63*(11), 2251–2272.

Scherer, K. R., London, H., & Wolf, J. J. (1973). The voice of confidence: Paralinguistic cues and audience evaluation. *Journal of Research in Personality, 7*(1), 31–44.

Schertz, J., & Clare, E. J. (2020). Phonetic cue weighting in perception and production. *Wiley Interdisciplinary Reviews: Cognitive Science, 11*(2), e1521.

Seyfarth, S. (2014). Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. *Cognition, 133*(1), 140–155.
Seyfarth, S., Garellek, M., Gillingham, G., Ackerman, F., & Malouf, R. (2018). Acoustic differences in morphologically-distinct homophones. *Language, Cognition and Neuroscience, 33*(1), 32–49.

Shatzman, K. B., & McQueen, J. M. (2006). Prosodic knowledge affects the recognition of newly acquired words. *Psychological Science, 17*(5), 372–377.

Sherr-Ziarko, E. (2015). Word frequency effects on homophonous words in Mandarin Chinese. In *Proceedings of the International Congress on Phonetic Sciences*.

Sobin, C., & Alpert, M. (1999). Emotion in speech: The acoustic attributes of fear, anger, sadness, and joy. *Journal of Psycholinguistic Research, 28*(4), 347–365.

Spinelli, E., Grimault, N., Meunier, F., & Welby, P. (2010). An intonational cue to word segmentation in phonemically identical sequences. *Attention, Perception, & Psychophysics, 72*(3), 775–787.

Syrdal, A. K., & McGory, J. (2000). Inter-transcriber reliability of ToBI prosodic labeling. In *Sixth International Conference on Spoken Language Processing*.

Tang, K., & Shaw, J. A. (2020). Prosody leaks into the memories of words. *arXivLabs*: https://arxiv.org/ftp/arxiv/papers/2005/2005.14716.pdf

Warner, N., Jongman, A., Sereno, J., & Kemps, R. (2004). Incomplete neutralization and other sub-phonemic durational differences in production and perception: Evidence from Dutch. *Journal of Phonetics, 32*(2), 251–276.

Williams, C. E., & Stevens, K. N. (1972). Emotions and speech: Some acoustical correlates. *The Journal of the Acoustical Society of America, 52*(4B), 1238–1250.

Wright, R. (2004). Factors of lexical competition in vowel articulation. In J. Local, et al. (eds.), *Papers in laboratory phonology VI* (pp. 75–87). Cambridge University Press.

Zhang, C., & Peng, G. (2013). Productivity of Mandarin third tone sandhi: A wug test. In G. Peng & F. Shi (eds.), *Eastward flows the great river: Festschrift in honor of Prof. William S-Y. Wang on his 80th birthday* (pp. 256–282). City University of Hong Kong Press.