How to use context to disambiguate overlapping categories: The test case of Japanese vowel length

Kasia Hitczenko[1], Reiko Mazuka[2,3], Micha Elsner[4] & Naomi H. Feldman[1,5]
khit@umd.edu, mazuka@brain.riken.jp, melsner@ling.osu.edu, nhf@umd.edu
[1]Department of Linguistics, University of Maryland
[2]RIKEN Brain Science Institute, Laboratory for Language Development
[3]Department of Psychology and Neuroscience, Duke University
[4]Department of Linguistics, The Ohio State University
[5]Institute for Advanced Computer Studies, University of Maryland

Abstract

Infants learn the sound categories of their language and adults successfully process the sounds they hear, even though sound categories often overlap in their acoustics. Most researchers agree that listeners use context to disambiguate overlapping categories. However, they differ in their ideas about how context is used. One idea is that listeners normalize out the systematic effects of context from the acoustics of a sound. Another idea is that contextual information may itself be an informative cue to category membership, due to patterns in the types of contexts that particular sounds occur in. We directly contrast these two ways of using context by applying each one to the test case of Japanese vowel length. We find that normalizing out contextual variability from the acoustics does not improve categorization, but using context in a top-down fashion does so substantially. This reveals a limitation of normalization in phonetic acquisition and processing and suggests that approaches that make use of top-down contextual information are promising to pursue.

Keywords: speech perception; phonetic category acquisition

One of the first tasks infants face when acquiring their native language is learning what its sound categories are, a task that involves grouping sounds that vary continuously into discrete categories. Even once people have learned their language and its sound categories, they still need to be able to map a particular acoustic pronunciation they hear to one of those categories, in order to process speech effectively. These can be difficult tasks because there is often a lot of overlap between categories in terms of how they are acoustically realized (Bion, Miyazawa, Kikuchi, & Mazuka, 2013), and this overlap can mask which sounds should be grouped together.

A prime example of this is Japanese vowel length, the test case we consider in this paper. In Japanese, vowel length is contrastive: whether a vowel is phonologically short or long can change the meaning of a word (e.g. /biru/ means *building*, but /biːru/ means *beer*). Short and long vowels are separate sound categories, yet analyses have shown that they overlap substantially in their durations.[1] That is, a particular production of a phonologically short vowel can be longer than a particular production of a phonologically long vowel. In fact, because only 9% of Japanese vowels are long, the combined distribution of vowels is unimodal (Figure 1). Cases like this one are problematic for classic distributional learning approaches, which posit that listeners make use of clusters or peaks in the acoustic data they hear to learn and process sounds (Maye, Werker, & Gerken, 2002; Bion et al., 2013).

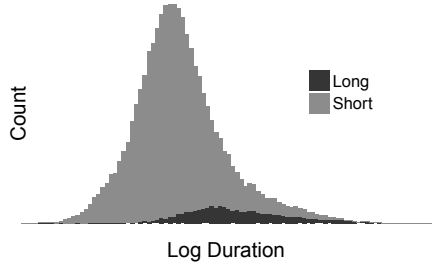

Figure 1: Acoustic distribution of Japanese vowel duration in spontaneously produced infant-directed speech. Data are from the R-JMIC corpus, as described in the Data section.

How do infants learn the sound categories of their language and how do adults process the sounds of their language when there is so much overlap in the speech they hear? A large body of work has argued that listeners use context to disambiguate sound categories, but researchers differ in their ideas about how context is used. One idea is that the context a sound occurs in systematically affects how that sound is produced and causes overlap in sound categories. Listeners then factor out the effect of context from the acoustics ('normalization' or 'adaptation') (e.g. McMurray & Jongman, 2011; Dillon, Dunbar, & Idsardi, 2013; Kleinschmidt & Jaeger, 2015). Another idea, which is not mutually exclusive from the first, is that sound categories differ in the types of contexts or environments they are likely to occur in for phonotactic, lexical, and other reasons. Listeners use this top-down information about which sound is most likely to occur in the context they heard to supplement acoustic cues (e.g. Ganong, 1980; Feldman, Griffiths, Goldwater, & Morgan, 2013). These ideas have been studied extensively; however, we have limited knowledge of the extent to which each one is effective on realistic data.

In this work, we directly contrast these two ways of using context by testing their relative efficacy in separating overlapping vowel length categories in Japanese. We show that factoring out context from the acoustic cues does not improve category separability, while using context as a direct cue parallel to acoustic cues does - and substantially. This result reveals limitations in the efficacy of factoring out systematic variability, and suggests that approaches that make use of top-down contextual information are promising to pursue in future research.

[1]We use *vowel length* to refer to the phonological status of a vowel and *vowel duration* to refer to the physical acoustic property of a vowel (i.e. how long the speaker took to produce that sound). It is thought that vowel duration is the main acoustic cue to vowel length.

Context in Phonetic Perception

This paper contrasts two ideas for how context could be used in the acquisition and processing of overlapping sound categories. The first is based on the idea that contextual factors (broadly construed to include speaker, neighboring sounds, speech rate, etc.) systematically and predictably affect the acoustic realization of particular sounds, causing the observed overlap between different categories. The idea is that context can be "factored out" of the acoustics, in order to reduce category overlap. Listeners might do this either by learning the structure of the variability and undoing its effect ('normalization'), or by building a separate model of the mapping between acoustics and categories for each context a sound occurs in ('adaptation'). Both normalization and adaptation have considerable scientific support. A body of experimental work has convincingly shown that listeners' perception of a particular sound can be changed by modifying the speaker (Nearey, 1978), the neighboring sounds (Mann & Repp, 1980), or the speech rate (Fujisaki, Nakamura, & Imoto, 1975) of the surrounding utterance. This set of findings has generally been interpreted as support for the idea that listeners take into account systematic variability when making categorization decisions - at least on controlled lab or synthetic speech. These findings have been supplemented by computational work, which has found that models that take into account systematic variability achieve better matches with human performance than models that do not, both in adult categorization (McMurray & Jongman, 2011) and sound category acquisition (Dillon et al., 2013).

The second idea, which is not mutually exclusive from the first, is that sound categories differ in the types of contexts or environments they are likely to occur in for phonotactic, lexical, historical and other reasons. Just knowing the context of a target sound, then, could be informative about what category it is likely to be. Listeners might supplement bottom-up acoustic cues with this type of top-down contextual information, when learning and categorizing the sounds they hear. Indeed, experimental work has shown that participants in speech perception experiments are biased to choose sound categorizations that result in words over non-words (Ganong, 1980), as well as phonotactically legal sequences over sequences that violate phonotactic constraints (Brown & Hildum, 1956). Feldman et al. (2013) showed that a computational model that used information about the word frames that sounds occurred in resulted in an improvement in sound category learning over models that did not incorporate lexical information.

The literature on these two ideas is extensive, but is not conclusive on what role each of these strategies plays in acquisition and processing. Many of the studies that are cited as classical evidence for factoring out systematic variability are consistent with using top-down linguistic information, and vice versa. As an example, Port and Dalby (1982) showed that whether participants perceived a particular stimulus as being the word *rapid* or *rabid* changed depending on the duration of the vowel that preceded the /p/ or /b/. This finding was originally taken as evidence that participants were normalizing the acoustics for speech rate. However, it was later considered evidence that participants were using the duration of the vowel as a direct cue to determining the identify of the consonant (Toscano & McMurray, 2012). Because these ideas have been somewhat conflated in the literature, it is hard to evaluate their relative contribution to acquisition and processing. In addition, because most of the evidence for these ideas comes from work on synthetic or controlled lab speech, we have limited knowledge about whether they are also effective on more naturalistic speech. This paper isolates the two ideas and tests their relative efficacy in separating overlapping categories, by applying them to the Japanese vowel length contrast.

Japanese vowel length is an ideal test case because there is evidence that both of these strategies could be helpful in overcoming the overlap between short and long vowels. On the one hand, research has shown that factors such as vowel quality (Hirata, 2004; Bion et al., 2013), speech rate (Hirata, 2004), prosodic position (Martin, Igarashi, Jincho, & Mazuka, 2016), and neighboring sounds (Hirata & Whiton, 2005) all affect the duration of Japanese vowels. It is possible that these factors could cause overlap between short and long vowels, in which case factoring out the effect of context would be effective. On the other hand, there is also evidence that there are systematic differences between short and long vowels in the types of contexts and environments that they occur in. For example, different vowel qualities (a, e, i, o, u) have different relative proportions of short and long vowels, short and long vowels differ in the types of sounds they co-occur with (Hirata, 2004), and long vowels are less likely to occur phrase-finally in some strata of the Japanese lexicon (Moreton & Amano, 1999). Listeners could exploit these contextual patterns in a top-down process to better process and learn the contrast.

In what follows, we compare the relative efficacy of these two strategies in separating overlapping categories by testing how well each of them categorizes Japanese vowels as short or long. Analysis 1 tests whether normalization improves categorization performance. Following Cole, Linebaugh, Munson, and McMurray (2010) and McMurray and Jongman (2011), and Nearey (1990), we implement the idea of factoring out systematic variability by regressing out contextual variability from acoustic cues. We then test whether a logistic regression categorization model that uses normalized cues outperforms ones that use unnormalized cues. Analysis 2 tests whether using top-down contextual information improves categorization performance, by comparing a logistic regression that only uses acoustic cues to ones that also use contextual factors as direct predictors of category membership.

We choose to implement factoring out systematic variability as normalization rather than adaptation because normalization allows us to isolate the two ways of using context in a way that adaptation does not.[2]

[2]Adaptation builds separate models for each context a sound occurs in, allowing it to make use of top-down information about how likely each category is to occur in a particular context, in addition to factoring out systematic acoustic variability.

Data

The data we use come from the RIKEN Japanese Mother-Infant Conversational Corpus (Mazuka, Igarashi, & Nishikawa, 2006). The data were originally collected by recording the speech of 22 mothers who visited the lab with their 18- to 24-month old children. The mothers first played with their child with picture books. They then played with their child with toys. Speech by the mother in these two sessions was labelled as infant-directed. In the final session, the mothers talked to a female experimenter and the mother's speech in this session was labelled as adult-directed. The corpus consists of about 14 total hours of speech, and is hand-labelled for both phonetic and prosodic information.

We extracted information about each of the vowels produced by the mothers, but excluded singing, coughing, devoiced vowels, diphthongs, and any segments that the researchers could not transcribe. This left 92003 total vowels, 30035 of which were in the adult-directed section of the corpus and 61968 of which were in the infant-directed section of the corpus. All of the analyses we report were run on the infant-directed part of the corpus, plotted in Figure 1.[3]

Acoustic cues

We extracted acoustic information about each vowel:

- **Duration**: We extracted vowel duration in seconds.
- **Formants**: Although, up to this point, we have only discussed duration, previous work has shown that spectral information can improve categorization performance (Hirata, 2004). As a result, we used the first three formants at the vowel midpoint which were automatically extracted using Praat (Boersma, 2001) by Antetomaso et al. (2017).

Contextual information

We also extracted a set of contextual factors about each vowel:

- **Vowel quality**: This was a categorical variable that took one of five values: /a/, /e/, /i/, /o/, /u/.
- **Speaker**: This was a categorical variable with one of 22 different possible values.
- **Accented?**: This was a binary variable that took a value of 1 if the vowel was accented and 0 if it was not.
- **Condition**: This variable indicated whether the mother uttered the vowel to their child while they were playing with books or with toys.
- **Prosodic position**: We extracted a categorical variable that indicated whether the word that the vowel occurred in was at the end of an accentual phrase (AP), at the end of an intonational phrase (IP), at the end of an utterance, or none of the above. We extracted a second categorical variable, which indicated whether the word that the vowel was in was AP-initial, IP-initial, utterance-initial, or none of the above. Finally, we extracted a vector of binary variables, of which the first three elements indicated the position of the vowel in the word (initial, medial, final), the next three its position in its AP, the next three its position in its IP, and the last three its position in the utterance. Unlike the previous two variables, this one marked the position of the vowel itself rather than its containing word.
- **Speech rate**: We extracted the average syllable duration of both the word and utterance the vowel was in (pauses were excluded in calculations). We also extracted the duration of the previous (and following) sound. For vowels without immediately preceding (or following) sounds, we used the overall average previous (and following) duration.
- **Neighboring sound**: We extracted the quality of the previous and following sounds, as well as whether or not they were geminate consonants.
- **Part-of-speech**: We coded whether each vowel was in a function or a content word based on part-of-speech annotations in the corpus.

[3]We also ran these analyses on the adult-directed part of the corpus and found comparable results.

Analysis 1: Removing Systematic Variability

In this section, we test to what extent normalizing out systematic variability from acoustics can help disambiguate short and long vowels.

Methods

We compare categorization models that make use of normalized acoustic cues to ones that make use of unnormalized acoustic cues. The unnormalized cues are the duration and formants taken directly from corpus annotations. To obtain the normalized acoustic cues, we train a linear regression model to predict each vowel's acoustic cues (duration and formants) from its context (speech rate, neighboring sounds, etc., as listed under Contextual Information). The model's prediction represents what we expect the vowel's acoustic cues to be based on its context. Once we have these predictions, we calculate each vowel's normalized acoustic cues by subtracting that vowel's predicted acoustic cues from its actual acoustic cues (i.e. by taking the residuals). This step effectively subtracts out the influence the context had on each vowel's acoustic cues. We vary whether or not we factor out the effect of part-of-speech from the acoustics. We want our analyses to apply to both acquisition and adult speech perception and because infants probably do not have access to part-of-speech information when learning about vowel length, we test how effective normalization is both with and without part-of-speech.

Once we have the normalized cues, we train logistic regression models to predict each vowel's length either from its unnormalized or its normalized acoustic cues. These logistic regressions take the input acoustic cues and output the relative probability that the vowel is short or long. The vowel is categorized as belonging to the category with higher probability.

To train the linear and logistic regressions, the data are divided into a training set (90% of the data) and a test set (10% of the data), keeping the relative proportion of short and long vowels constant across the sets. For normalization, the linear regression equation is estimated on the training set, but is used to normalize the acoustic cues in both the training and test set. The logistic regression is trained on the same training set as

Model	Accuracy	Short Accuracy	Long Accuracy	BIC
Model 1: Unnormalized baseline	91.0	98.8	13.4	28698
Model 2a: Normalized (no part-of-speech)	91.0	99.5	6.1	31552
Model 2b: Normalized (with part-of-speech)	91.0	99.5	5.8	31644
Model 3a: Top-down information (no part-of-speech)	95.1	99.0	60.0	17784
Model 3b: Top-down information (with part-of-speech)	95.3	99.0	61.6	16760

Table 1: Summary of results from Analysis 1 and 2. Analysis 1 compared the Unnormalized/Baseline model to the two Normalized models. Analysis 2 compared the Unnormalized/Baseline model to the two Top-down information models.

the linear regression (except that the acoustic cues are now normalized). In order to make sure that the models performed consistently, we randomly split the data into training and test sets 10 separate times, ran each model ten times, and averaged performance across these ten runs.

We report two types of evaluation metrics for each model we present. First, we report overall categorization accuracy on the unseen test set, which is the percentage of all vowels in the test set that the model categorized correctly, as well as the accuracy on just short vowels and just long vowels. Second, we report the Bayesian Information Criterion (BIC) for each model, which is a common metric used to select between different models. It has the property that it prefers simpler models, all else being equal, and lower values are better.

Results

A summary of the results is presented in Table 1.

Unnormalized Model (Model 1) The baseline model used unnormalized duration and formants as predictors of category membership, without regressing out the effects of context. This logistic regression model reached an overall accuracy of 91.0%. It got 98.8% of short vowels correct and 13.4% of long vowels correct. It had a BIC of 28698. While the overall accuracy and the accuracy on short vowels seem quite impressive, a model could get 90.9% accuracy just by guessing 'short' for every vowel token.

Normalized Models (Models 2a-2b) These two models both used normalized duration and formants as predictors of categorical membership, but we varied whether we regressed out part-of-speech. Without regressing out part-of-speech, the model reached an overall accuracy of 91.0%. It got 99.5% of the short vowels correct, and 6.1% of the long vowels correct. It had a BIC of 31552. With part-of-speech regressed out, the model reached an overall accuracy of 91.0%, getting 99.5% of the short vowels correct, and 5.8% of the long vowels correct. It had a BIC of 31644. Both models showed worse performance compared to the unnormalized model, with lower accuracy on long vowels and an increase in BIC. Normalizing out the effect of part-of-speech resulted in worse performance than ignoring its effect.

Discussion

Previous work suggests that factoring out systematic variability could be helpful in the acquisition and processing of the Japanese vowel length contrast; however, our current results did not support this hypothesis. Our main finding was that normalization did not increase separability between short and long vowels, as evidenced by no improvement in categorization. Given how prevalent this hypothesis has been in the literature, the results are surprisingly bad and suggest that normalization in the duration and formant cues may not be the solution to the Japanese vowel length contrast problem. We return to the question of why we observed these results in the General Discussion.

Analysis 2: Using Top-down Information

In this section, we test to what extent making use of top-down contextual information, in addition to bottom-up acoustics, could help in the acquisition and processing of Japanese vowel length. In this analysis, context is used to directly predict the identity of a vowel, rather than to predict its acoustics, as was done in Analysis 1.

Methods

As in the previous analysis, we use logistic regression models as our categorization models, but in this analysis, we do not run any linear regressions to regress out context from acoustics. The baseline model was identical to Model 1 in Analysis 1. This was compared against two logistic regression models that predicted vowel length from unnormalized acoustic cues, plus the contextual factors described in the Data section. As before, we varied whether part-of-speech was included. The models were again evaluated using overall, short vowel, and long vowel accuracies, as well as BIC.

Results

The results are summarized in Table 1.

Baseline Model (Model 1) The baseline model was identical to the unnormalized model in Analysis 1 (see Table 1).

Top-Down Information Models (Models 3a-b) In these two models, in addition to using unnormalized duration and formants to predict vowel length, we also used contextual factors as direct predictors of vowel length. When we did not include part-of-speech, the model achieved 95.1% accuracy overall, getting 99% of short vowels right, and 60% of long vowels right. It had a BIC of 17784. With part-of-speech, the model had 95.3% overall accuracy, getting 99% of short vowels right and 61.6% of long vowels right. It had a BIC of 16760. Both of these models substantially outperformed the baseline model, while adding part-of-speech as a predictor

additionally improved performance, lowering BIC and increasing long vowel and overall accuracy.

Discussion

We investigated to what extent using contextual information as a direct cue to vowel length, in addition to acoustic cues, helped in categorization. We found that this strategy drastically improved accuracy and lowered BIC scores, suggesting that this method separates short and long vowels quite well. Given the relatively small set of factors we used, it is quite impressive that we achieved this level of performance, and it suggests that this is a promising strategy to pursue in future work.

General Discussion

This paper applied two ideas about how listeners might use context in overcoming the overlapping category problem to the test case of Japanese vowel length. Results showed that normalizing the effect of context out of the acoustic cues did not improve short and long vowel separability. On the other hand, using contextual information as a direct cue to category membership resulted in much better separability between categories, as evidenced by an improvement in categorization.

The fact that normalization was not helpful was surprising, given how well-established this idea is in the field. The models we tested were supervised, and were given information on what the vowel categories and relevant contextual factors were, yet were still unable to separate short and long vowels. The problem would be even greater in acquisition, where the learner would need to simultaneously learn what the categories are and how to factor out context from the acoustics.

Previous work using the same normalization techniques that we employ here found improvements in accuracy. Although it is difficult to directly compare improvement based on accuracy, two past studies reported increases of performance from 28.63% to 54% (Cole et al., 2010) and from 83.3% to 92.9% (McMurray & Jongman, 2011). There are a few reasons why we may have found different normalization performance.

First, in our case, there were many more short vowels than long vowels, whereas in previous work, the categories were more balanced. Binary classifiers can perform poorly on imbalanced data (He & Garcia, 2009), so it is possible that the bad performance we observe is not due to normalization, but rather to the categorization model. Although this is a legitimate concern, the categorization model in Analysis 2 achieves much better categorization, which makes this explanation unlikely. What is more likely is that imbalances in the proportion of long and short vowels differed across contexts, and this affected the results. Imbalances in a particular context - precisely the signal that top-down models use - can impede normalization by artificially shortening or lengthening the mean duration of vowels in a context.

Second, it is possible that we are not factoring out the right set of contextual factors for the Japanese vowel length case. We only considered a small number of contextual factors, and many of the factors we included were quite basic compared to the complicated processes they represent. For example, we reduced all of the complexity of pitch accent to a single binary variable. It is possible that representing the full complexity of all of the relevant factors would improve results. However, the fact that we were unable to include more sophisticated factors does not explain the lack of improvement from what was included. The model had access to many factors that have been shown to systematically affect vowel duration in Japanese, yet factoring them out did not improve results.

Third, it is possible that there was a problem with the particular implementation of normalization we used. While this implementation has been effective in other cases (Cole et al., 2010; McMurray & Jongman, 2011), it has not been applied to spontaneous speech before, and it may be unable to capture the structure of the contextual variability. We did not include interaction terms between factors in the linear regression as this was computationally difficult, yet research has revealed that these interactions exist. For example, vowels are lengthened more in slow speech than in fast speech. Linear regression models also assume that particular contextual factors add or subtract a fixed duration, but it is possible that these factors affect duration in a different way (e.g. in a multiplicative fashion by, for example, doubling the duration of the vowel). A more complex normalization model might be more successful, by better capturing the relationship between context and acoustics However, as before, the fact that a more complex model could improve performance fails to explain why normalization did not help here. Previous work has shown that individual factors such as vowel quality, speech rate, and prosodic position each have systematic effects on acoustics that are evident even when variability in other factors is not controlled for. Our model can normalize out these individual effects, but doing so did not improve performance.

Fourth, it is possible that factoring out systematic variability is not a strategy that would work for the particular case of the Japanese vowel length contrast. For instance, it may be the case that top-down information is sufficient to distinguish most long/short minimal pairs without attending to the acoustic duration at all, so that in conversational speech, the durational contrast is mostly neutralized. Under this account, factoring out systematic variability would work for contrasts with high functional load, where speakers must produce a perceptible contrast in order to be understood, but might be ineffective for contrasts with low functional load, where speakers might not produce an acoustic contrast all the time. Further research is needed on what cues adult listeners use to distinguish Japanese long and short vowels, and in general, what sorts of problems can be solved by factoring out variability.

Finally, it is possible that factoring out systematic variability is not effective for spontaneously produced speech. Although a body of work has argued that listeners do factor out systematic variability from the acoustics (McMurray & Jongman, 2011; Mann & Repp, 1980), most of this work has studied carefully controlled laboratory speech or synthetic speech, instead of spontaneous speech, and typically manipulated the influence of one contextual factor at a time. It is possible

that in spontaneous speech, which has been shown to be quite different than laboratory speech (Wagner, Trouvain, & Zimmerer, 2015), there are so many individual factors involved and interacting with one another that normalization becomes ineffective. However, this idea is difficult to reconcile with the automatic speech recognition literature, where speaker and speech rate are taken into account in many systems, and are often used when dealing with real, messy speech. Pursuing this possibility would require revisiting some previous work that showed improvements from normalization.

Overall, we do not yet have enough evidence to make a strong claim about which of these (and other) possibilities is correct. In ongoing and future work, we will test the efficacy of normalization in other test cases of overlapping categories, on Japanese laboratory speech, with more and more sophisticated contextual factors, and we will use neural networks to implement more sophisticated normalization techniques. This will allow us to better understand the pattern of results observed here, as well as allow us better delineate when factoring out systematic variability is helpful and when it is not.

In Analysis 2, we showed that an alternative hypothesis where the factors were used as independent predictors of vowel length resulted in improved categorization performance. Given the small set of factors we included, it is quite impressive that we are able to correctly classify around 99% of the short vowels, as well as over 60% of the long vowels, given just how much they overlap along the duration dimension. At minimum, this should be taken as evidence that it is possible to improve categorization performance substantially, even when the base categorization rate is over 90% due to a high rate of short vowels. More strongly, this suggests a promising alternative to consider in future work. Although our use of supervised models means we cannot draw strong conclusions about acquisition, these results show that there is signal in the data that could be exploited. Ultimately, if we can show that short vowels have quite different distributions than long vowels when we include non-acoustic information, then we can start to study unsupervised versions of this model.

In this paper, we contrasted two ways of using context to overcome the overlapping categories problem: factoring out systematic variability arising from the context and using contextual information as a direct cue to category membership in a top-down fashion. Our intention is not to imply that the true solution is one or the other, but rather to study the relative contribution of each of these hypotheses separately. Our results call into question the idea of factoring out systematic variability on its own. It may still be useful when combined with other ideas, and future research should consider strategies, like adaptation, that integrate both ideas.

Acknowledgments This work was supported by NSF grants IIS-1421695, IIS-1422987, DGE-1449815, and NSF/JSPS EAPSI grant 1713974. We thank the reviewers, Adam Albright, Stephanie Antetomaso, Edward Flemming, Bill Idsardi, Chiyuki Ito, Kyoji Iwamoto, Jeff Lidz, Thomas Schatz, Kristine Yu, the RIKEN Lab for Language Development, the MIT Computational Psycholinguistics Lab, the MIT Phonology Circle, the MIT CompLang group, NECPhon 11, and Brown University LingLangLunch for their help and feedback.

References

Antetomaso, S., Miyazawa, K., Feldman, N., Elsner, M., Hitczenko, K., & Mazuka, R. (2017). Modeling phonetic category learning from natural acoustic data. In *BUCLD 41: Proceedings of the 41st Annual Boston University Conference on Language Development.*

Bion, R. A., Miyazawa, K., Kikuchi, H., & Mazuka, R. (2013). Learning phonemic vowel length from naturalistic recordings of Japanese infant-directed speech. *PloS ONE*, *8*(2), e51594.

Boersma, P. (2001). Praat: A system for doing phonetics by computer. *Glot International*, *5*(9/10), 341-345.

Brown, R. W., & Hildum, D. C. (1956). Expectancy and the perception of syllables. *Language*, *32*(3), 411-419.

Cole, J., Linebaugh, G., Munson, C., & McMurray, B. (2010). Unmasking the acoustic effects of vowel-to-vowel coarticulation: A statistical modeling approach. *Journal of Phonetics*, *38*(2), 167-184.

Dillon, B., Dunbar, E., & Idsardi, W. (2013). A single-stage approach to learning phonological categories: Insights from Inuktitut. *Cognitive Science*, *37*(2), 344-377.

Feldman, N. H., Griffiths, T. L., Goldwater, S., & Morgan, J. L. (2013). A role for the developing lexicon in phonetic category acquisition. *Psychological Review*, *120*(4), 751-778.

Fujisaki, H., Nakamura, K., & Imoto, T. (1975). Auditory perception of duration of speech and non-speech stimuli. *Auditory Analysis and Perception of Speech*, 197-219.

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, *6*(1), 110-125.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263-1284.

Hirata, Y. (2004). Effects of speaking rate on the vowel length distinction in Japanese. *Journal of Phonetics*, *32*(4), 565-589.

Hirata, Y., & Whiton, J. (2005). Effects of speaking rate on the single/geminate stop distinction in Japanese. *The Journal of the Acoustical Society of America*, *118*(3), 1647-1660.

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, *122*(2), 148-203.

Mann, V. A., & Repp, B. H. (1980). Influence of vocalic context on perception of the [ʃ]-[s] distinction. *Attention, Perception, & Psychophysics*, *28*(3), 213–228.

Martin, A., Igarashi, Y., Jincho, N., & Mazuka, R. (2016). Utterances in infant-directed speech are shorter, not slower. *Cognition*, *156*, 52-59.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*(3), B101-B111.

Mazuka, R., Igarashi, Y., & Nishikawa, K. (2006). Input for learning Japanese: RIKEN Japanese mother-infant conversation corpus. *The Technical Report of the Proceedings of the Institute of Electronics, Information and Communication Engineers*, *106*(165), 11-15.

McMurray, B., & Jongman, A. (2011). What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological Review*, *118*(2), 219-246.

Moreton, E., & Amano, S. (1999). Phonotactics in the perception of Japanese vowel length: Evidence for long-distance dependencies. In *Eurospeech.*

Nearey, T. M. (1978). Vowel space normalization in synthetic stimuli. *The Journal of the Acoustical Society of America*, *63*(1).

Nearey, T. M. (1990). The segment as a unit of speech perception. *Journal of Phonetics*, *18*(3), 347-373.

Port, R. F., & Dalby, J. (1982). Consonant/vowel ratio as a cue for voicing in English. *Attention, Perception, & Psychophysics*, *32*(2), 141-152.

Toscano, J. C., & McMurray, B. (2012). Cue-integration and context effects in speech: Evidence against speaking-rate normalization. *Attention, Perception, & Psychophysics*, *74*(6), 1284-1301.

Wagner, P., Trouvain, J., & Zimmerer, F. (2015). In defense of stylistic diversity in speech research. *Journal of Phonetics*, *48*, 1-12.