Learning phonology from surface distributions, considering Dutch and English vowel duration

Daniel Swingley

University of Pennsylvania

Author Note

This work was supported by NIH grant R01-HD049681 to D. Swingley. Some of the same data, and the argument in quite preliminary form, were presented at the International Conference on Infant Studies in 2006. The author thanks Michael Brent and Jeff Siskind for making the English corpus available, and Paula Fikkert and Claartje Levelt for the Dutch corpus.
Correspondence should be directed to Daniel Swingley at 425 S. University Ave., Department of Psychology, University of Pennsylvania, Philadelphia PA 19104; phone, 215-898-0334; email, swingley@psych.upenn.edu.

Abstract

In learning language, children must discover how to interpret the linguistic significance of phonetic variation. On some accounts, receptive phonology is grounded in perceptual learning of phonetic categories from phonetic distributions drawn over the infant's sample of speech. On other accounts, receptive phonology is instead based on phonetic generalizations over the words in the lexicon. Tests of these hypotheses have been rare and indirect, usually making use of idealized estimates of phonetic variation. Here we evaluated these hypotheses, using as our test case English and Dutch toddlers' different interpretation of the lexical significance of vowel duration. Analysis of thousands of vowels of one Dutch and three English mothers' speech suggests that children's language-specific differences in interpretation of vowel duration are likely due to detection of lexically specific patterns, rather than bimodality in raw phonetic distributions.

Keywords: language acquisition, lexical development, word learning, phonetic categorization, speech perception

Learning phonology from surface distributions, considering Dutch and English vowel duration

As listeners interpreting speech, we must account for the phonetic signal we hear in terms of the linguistic options available to the talker. The listener's problem is to attribute the various components of the signal to their causes, which include, among other things, characteristics of the talker, such as the shape of his or her vocal tract, his or her emotional state, and so on; and characteristics of the message, such as its syntactic structure, its words, and its place in the discourse. The success of this attribution process among ordinary adults is evidence that we have an interpretive model that can quickly evaluate the likelihood of what is initially a huge range of possible explanations for the speech heard, and narrow these down to the most probable interpretation. Because all languages work somewhat differently, much of this interpretive model must be learned. A primary task of research on language acquisition is to explain how this learning takes place.

Here, we address the phonetic interpretation process, focusing on how children learn to interpret the phonetic feature of vowel duration. Vowel duration is an interesting case to examine because it is implicated in a wide range of linguistic functions. As a result, its interpretation demands sifting through a broad set of candidate explanatory causes. A vowel has the duration it has because of the speaking rate of the utterance, the vowel's position in the utterance, the degree of emphasis its word is being given, whether it is in a syllable receiving stress, the phonological features of the surrounding sounds (e.g., whether the following consonant is voiced), and the intrinsic duration of the vowel. In English, all of these matter. If a speaker wishes to emphasize a word, she may increase the duration of the stressed vowel in that word; to the listener, it will then seem longer than expected (according to an interpretive model that considers all of these factors) and he may attribute the extra duration to emphasis.

How can a child discover this model? Our exploration of this issue concerns whether children think that their language contrasts vowels (and therefore words) by vowel duration. Some languages, such as Japanese, have a set of phonetically short vowels and a contrasting set of otherwise similar-sounding long vowels (Fujisaki, Nakamura, & Imoto, 1975; Vance, 1987).

Others, such as Spanish, do not. And many languages, such as English and Dutch, fall somewhere in between, using duration as a cue to vowel identity, but usually weighted less strongly than phonetic quality. Children learning any language have to work out how strongly to weight word identity (or vowel category) in explaining vowel duration. For example, if a child learns the English word *plank* in the course of a conversation about pirates, and then the following day hears an instance of that word realized with a much longer vowel, should the child take seriously the possibility that *plaank* is a different word? Mature American English speakers know that this is unlikely, but how is this intuition learned?

As with many linguistic phenomena, the ready ease of our intuitions is misleading. Models of segment duration intended for the purpose of speech synthesis illustrate the complexity of the problem. In the early days of modern speech synthesis, researchers attempted to consider all of the linguistic contexts or functions that affect segment duration, establish the magnitude of each effect (how much longer is a stressed vowel? an utterance-final vowel?), and derive rules for managing the interaction of these these effects (how much longer is a stressed, utterance-final vowel?; Klatt, 1976, 1987). The rules turned out to interact in complex ways that seemed neither consistently additive nor multiplicative, and so such models had dozens of parameters to estimate (van Santen, 1992). More recent approaches forego discovering a descriptively accurate set of discrete rules and interactions, and instead train neural networks or other statistical models using huge training sets with labeled linguistic features (e.g., Gonzalvo, Tazari, Chan, Becker, Gutkin, & Silen, 2016; Henter et al., 2016; Ling et al., 2015). Still, inadequate prosodic modeling still contributes to a lack of naturalness in speech synthesis systems.

To help understand how children might begin to learn their language's duration model, we consider English and Dutch. These make an interesting pair to contrast because although the languages are similar, children seem to interpret vowel duration somewhat differently in each language. English and Dutch have many features in common: both allow complex syllable structures (e.g. in words like *strength* or *knecht*, "knight"); both commonly reduce vowels to schwa in unstressed syllables (though Dutch less so than English; Warner & Cutler, 2017); both

mark lexical stress with altered pitch and lengthened syllable duration; and both have similar numbers of consonants and vowels (English, about 24 and 15; Dutch, about 23 and 16). In addition, the vowels of both English and Dutch have characteristic intrinsic durations (e.g., English [æ] and Dutch [aː] tend to be long in citation form; English [ɛ] and Dutch [ɑ] tend to be short; Hillenbrand, Getty, Clark, & Wheeler, 1995; Adank, van Hout, & Smits, 2004). Changing the durations of citation-form vowels can lead native listeners of either language to categorize the vowels differently (e.g., Hillenbrand, Clark, & Houde, 2000; Chládková, Escudero, & Lipski, 2015; Tillman, Benders, Brown, & van Ravenzwaaij, 2017). Informally, English adults speak of "long" and "short" vowels when discussing English, and learn about this distinction in school; in Dutch the distinction is also familiar to adults, and it is consistently marked in the orthography (e.g., *man*, "man"; *maan*, "moon").

In spite of these broad similarities, native speakers of Dutch and English treat vowel duration somewhat differently, with Dutch listeners being more liable to change their categorization of vowels with manipulated durations than English listeners are (van der Feest & Swingley, 2011). This difference is manifested in very early childhood. Two studies of young children have indicated that Dutch toddlers are more strongly affected by vowel duration manipulations than English-learning toddlers. The first tested 18-month-olds' ability to learn two new words as labels for two unfamiliar objects, where the words differed only in their vowels' durations (Dietrich, Werker, & Swingley, 2007). In the first experient, children were habituated to one object being labeled as a [tɑm] (recorded by a Dutch speaker) and another object labeled as a [tɑːm]. Then, learning was tested by measuring children's reaction to the same stimulus pairings, or the reversed pairings (Stager & Werker, 1997). Dutch children, not English-learning children, responded to the change, by looking up at the screen longer on switched trials. English learners appeared indifferent to the switch. In a second experiment, children behaved similarly when tested on similar materials derived from English. Thus, given some training wherein vowel duration was used contrastively, Dutch 18-month-olds accepted the distinction, and English learners did not.

A second study confirmed this difference without training, using words children already

knew. Taking advantage of the fact that children find words harder to recognize when the words are pronounced with phonological deviations (Hallé & de Boysson-Bardies, 1994; Swingley & Aslin, 2000), Swingley and van der Feest (in press) presented 21-month-olds with pairs of pictures and named one of the pictures in a sentence, sometimes with the canonical pronunciation and sometimes with either an elongated or a shortened vowel. English learners looked at the named picture equally whether its label was pronounced normally or not; Dutch learners looked at the named picture significantly less when the vowel in its name had been shortened. Vowel lengthening, on the other hand, had no impact, a result that has also been found in prior studies of Dutch adults (e.g., Nooteboom & Doodeman, 1980; Van der Feest & Swingley, 2011).

Thus, by about one and a half years, children learning English and Dutch have already come to align with mature speakers, at least in a broad sense, about how duration should be interpreted lexically. This is the phenomenon we hope to explain, by examining corpora of speech directed to English or Dutch infants. Our starting point is the idea that children's different interpretations should be related to measurable statistical properties of the two languages. In particular, we might expect that vowel duration would tend to be more bimodally distributed in Dutch than in English, suggesting to infants that there are categories of duration values that are relevant to segmental phonology and therefore lexical differentiation (e.g., Bion et al., 2013). This expectation is grounded in theories of phonetic learning via distributional clustering (e.g., Cristià, McGuire, Seidl, & Francis, 2011; Goudbeek, Smits, & Swingley, 2009; Maye, Werker, & Gerken, 2002), which have the significant virtue of not requiring explicit teaching or labeling in the learning process. One reason why Dutch vowels might be more bimodal is that perhaps Dutch but not English makes a phonological distinction according to duration, and this has an impact on phonetic implementation. Another is that Dutch vowels do not exhibit variation due to the voicing of following consonants, because Dutch devoices codas. Thus, whereas in English the /v/ of *leave* causes the vowel to be longer than the vowel in *leaf*, no such process takes place in Dutch because voiced-coda words like *leave* are generally not permitted (e.g., Booij, 1999).

If we find that duration distributions are more bimodal in Dutch than in English, this would

provide support for a theory that proposes that young Dutch children's apparently greater weighting of vowel duration in lexical tasks derives from this greater statistical separation. On the other hand, if the separation of long and short vowels is similar in Dutch and English, this would not be consistent with such a theory, and would call for an alternative explanation.

Methods

We examined four infant-directed speech corpora, one of Dutch and three American English. The Dutch corpus was made available by Paula Fikkert and Claartje Levelt (Fikkert, 1994; Levelt, 1994); in the present project we used 206 sentences directed to the child *Catootje*, age 1;10.10, and 239 sentences directed to *Robin*, age 1;9.11, each from a single recording of each child, selecting only sentences by a single speaker. The talker was a native Dutch-speaking mother (though not the mother of either of the two children) who, in general, used an "infant-directed" speech register. The speech was recorded in late 1989 and early 1990. Although it would have been better to use a corpus of speech directed to children in their first year, as was the case with the English comparison corpora, no such corpus was available. The use of a single corpus of Dutch is a limitation of the present study. This corpus was selected because an informal assessment of the sound quality of the recordings suggested that they were of adequate quality for this analysis.

The English samples were taken from the Brent and Siskind corpus (Brent & Siskind, 2001) of infant-directed speech, including 1088 sentences from three sessions of mother *f1* (ages 0;10.03, 0;10.13, and 0;11.06), 408 sentences from one session of mother *d1* (age 10;22) and 666 sentences from mother w1 (age 0;10.25). The speech was recorded in 1996. All mothers were reported as having college degrees (like the Dutch speaker). These English data overlap partially with the dataset whose vowel formant measurements were reported in Adriaans and Swingley (2017). These mothers and sessions were selected because the recording quality was deemed sufficiently clear. The number of utterances analyzed is somewhat fewer than the number of maternal utterances in each session recording because in some cases the timing of the utterance

was incorrect in the corpus (so the extracted sentence was cut off) or because sometimes extraneous noise or nonmaternal vocalizations rendered some or all of the maternal speech hard to analyze.

For both languages, the speech toolkit HTK (Young et al., 2006) was used to estimate word and phone boundaries using the *HVITE* forced-alignment tool. All boundaries were then hand-corrected by visualising each segment in context using Praat (Boersma & Weenink, 2001). The Dutch alignments were done by the author, a speaker of Dutch; the English alignments were done by phonetically trained research assistants supervised by the author or by Frans Adriaans.

In addition to correcting the word and phone alignments, annotators corrected the phonological labels in each phone interval. This was necessary because actual pronunciations of words are not always the same as the canonical pronunciations given in the dictionary that was used for the alignments. It is common for speakers to omit or change sounds. For example, Johnson's analysis of the Buckeye conversational English corpus (2007) found that more than 60% of word tokens deviated from the canonical (dictionary) form in at least one phone, and more than 20% had a phone deletion (Johnson, 2003). Mothers talking to their children do not speak just like the dictionary either. For example, in English-speaking Mom w1's sample, about 20% of word tokens were judged to have a pronunciation different from the dictionary (this is fewer than in Buckeye, but this may be due to our less detailed annotation scheme; for example, we did not differentiate glottalized and nonglottalized stop consonants). When our analyses required establishing a given pronunciation for a given word type, the most common pronunciation used by that mother was used, irrespective of whether this was the dictionary pronunciation or not.

The data from each corpus were assembled into a data structure with one vowel per row, tagged with its duration, its phonological transcription, the orthographic word it was part of, its syllable number, and whether it was the final syllable in the utterance. We focused on the specific pairs of monophthongal vowels in each language (a) for which listeners might be tempted to use duration to distinguish the vowels, based on their spectral similarity; and (b) which were sufficiently numerous for getting reasonable variability estimates. A premise of this approach is

that children evaluate the contrastiveness of vowel duration by evaluating whether similar-sounding vowels (such as Dutch [ɑ,aː] or English [ɛ,æ]) cluster into longer and shorter groups. The Dutch analysis compared the pairs /ɑ, aː/, /ɛ, eː/, and /ɔ, oː/. In some analyses we display the results for Dutch /ɪ, i/ for the purpose of comparison with the other Dutch cases. In Dutch this is a tense/lax pair that is not considered to be contrasted by duration. The English analysis compared the pairs /ɛ, æ/, /ʌ, a/, and /ɪ, i/.

With the dataset thus established, the first step was to evaluate whether Dutch children might rely on duration more than English learners simply because Dutch vowels are spectrally similar to one another, and therefore duration is more useful in distinguishing them. This hypothesis was not supported. Next, we tested whether distributions of Dutch long/short pairs' durations are more bimodal than English ones. This hypothesis was also not supported. We followed these analyses with tests in which lexical content was brought to bear on the grouping of vowel tokens for comparison of phonologically long and short vowels. In some of these analyses, Dutch revealed greater separation between long and short than English did. We will conclude by suggesting that the lexicon likely plays a role in children's determination of the phonological role of phonetic duration.

Results

Similarity in formant space

To characterize the spectral similarity of prospective vowel pairs, formant values for each English token were taken from the Adriaans & Swingley (2017) dataset, or, for Dutch, measured using Praat, with clear errors being excluded (e.g., f1 values 3 standard deviations from the mean, or other signs of frequency doubling in the measurement). Many of the formant measurements were hand-checked. The Dutch formant data were available for the subset of speech directed to one of the two children (55% of the dataset).

Spectral overlap among members of vowel pairs was evaluated by attempting to optimally separate the members of each pair using a quadratic discriminant analysis (QDA) on the z-scored

corpus	pair (short, long)	count	qda-overlap
Dutch	ɑ, aː	226, 243	23.5
Dutch	ɛ, eː	141, 167	14.3
Dutch	ɪ, i	221, 195	26.5
Dutch	ɔ, oː	87, 141	25.4
English (d1)	ɛ, æ	72, 89	28.7
English (f1)	ɛ, æ	303, 262	22.0
English (w1)	ɛ, æ	266, 224	32.2
English (d1)	ʌ, a	123, 107	35.2
English (f1)	ʌ, a	207, 220	30.1
English (w1)	ʌ, a	247, 148	27.2
English (d1)	ɪ, i	210, 163	28.0
English (f1)	ɪ, i	479, 339	17.6
English (w1)	ɪ, i	414, 399	17.8

Table 1

Vowel pairs from each corpus: descriptive statistics. Higher qda-overlap means greater classification error and therefore greater similarity between the vowels (see text).

first and second formant data. The classifier was trained on the labeled dataset and tested on the same data (i.e., supervised classification of tokens of each pair), and the proportion of errors was taken as the measure of overlap between the two categories. In principle, the existence of substantial overlap within a pair might motivate the child's use of additional dimensions of variation such as duration. Descriptive data on the tested pairs are given in Table 1.

As the table shows, the vowel pairs varied in how much they overlapped, with [ɛ, eː] relatively separate in Dutch and [ɪ, i] relatively separate in English. Overall, though, there was no indication of a large difference between the languages. For example, in this dataset Dutch [ɑ, aː] was not especially difficult to distinguish based on formant values relative to English [ɛ, æ]. Thus

this informal test does not support the hypothesis that children respond differently to durational manipulations in experimental settings as a result of a dramatically greater need among Dutch children to use a supplementary cue to vowel identity to distinguish members of these pairs (assuming, of course, that our Dutch speaker may be taken as representative).

These results suggest that there is no reason to suppose that Dutch toddlers cannot discriminate the long/short pairs on the basis of quality alone, or at least there is no reason to imagine that they would differ from English learners in this respect. To the best of our knowledge, this has not been tested in Dutch children, though there is evidence that Dutch 11-12 month olds and 14-15 month olds readily discriminate /ɪ, i/ (Liu & Kager, 2016).

Overall duration distributions

Though both English and Dutch conventions describe vowels as "long" and "short" (labels that overlap imperfectly with the tense/lax distinction), infant-directed speech might present Dutch infants with clear distinctions between short and long, and English infants with relatively overlapping distributions. Were this the case, it could explain the behavioral differences between Dutch and English learners in experimental tests.

Figure 1 characterizes the duration distributions using histograms and density plots (smoothed histograms). In every case, the phonologically long vowels tended to have greater durations than the phonologically short vowels (by Mann-Whitney-Wilcox test, all $p < .03$, with most $p < .00001$). The Mann-Whitney-Wilcox U statistic varies with sample size, so to properly compare the amount of non-overlap in each pair, the duration distributions were characterized using rank-biserial correlation (rbcc), a nonparametric variant of the point-biserial correlation (comparing the sum of ranks of one group against the sum of ranks of the other). Greater separation between the members of each pair is shown by higher correlations. These statistics are provided in the Figure and in Table 2. Remarkably, the Dutch cases are not special. In general it is not clear that particular Dutch pairs should be compared to particular English ones, except in the case of [ɪ,i] which are phonetically very similar in the two languages, and in this case the overlap

is actually greater in Dutch than in any of the three English examples. (This is consistent with standard phonetic accounts of Dutch; Booij, 1999). Of the 13 correlations, the 4 Dutch ones all fall somewhere in the middle; and the Dutch pair for which the developmental evidence for duration being interpreted contrastively is the strongest, namely [ɑ,aː], overlaps more than most of the other pairs of either language. The result is quite similar if we collapse over the pairs (omitting Dutch [ɪ,i]): the rbcc values are .31 for Dutch, and .26, .31, and .49 for the English samples. A plot detailing these results is given in the Supporting Online Materials (Fig. S1).

Table 2

Separation (rank-biserial correlation coefficients) of pairs, considering raw durations, and residuals of regression partialling out the lengthening effect of utterance-final position (see text)

corpus	analysis	ɑ, aː	ɛ, eː	ɪ, i	ɔ, oː	analysis	ɑ, aː	ɛ, eː	ɪ, i	ɔ, oː
Dutch	raw	0.19	0.36	0.20	0.46	resid.	0.24	0.42	0.20	0.54

corpus	analysis	ɛ, æ	ʌ, a	ɪ, i	analysis	ɛ, æ	ʌ, a	ɪ, i
Eng. d1	raw	0.31	0.15	0.39	resid.	0.36	0.16	0.34
Eng. f1	raw	0.44	0.47	0.53	resid.	0.50	0.56	0.50
Eng. w1	raw	0.14	0.23	0.37	resid.	0.19	0.28	0.34

Though this result leaves little confidence that a simple distributional difference can account for the crosslinguistic difference children manifest, it is possible that a somewhat more sophisticated model of children's duration interpretation would reveal a stronger separation between the languages. For example, utterance position exerts a strong influence on segmental duration (e.g., Cambier-Langeveld, 1997; Cho, 2015; Edwards, Beckman, & Fletcher, 1991). If a quirk of the vocabulary caused phonologically shorter vowels to appear in utterance-final position more often in Dutch than in English, this would bias the overall distribution and might conceal a phonological distinction. Of course, children could only uncover it if they took utterance-final

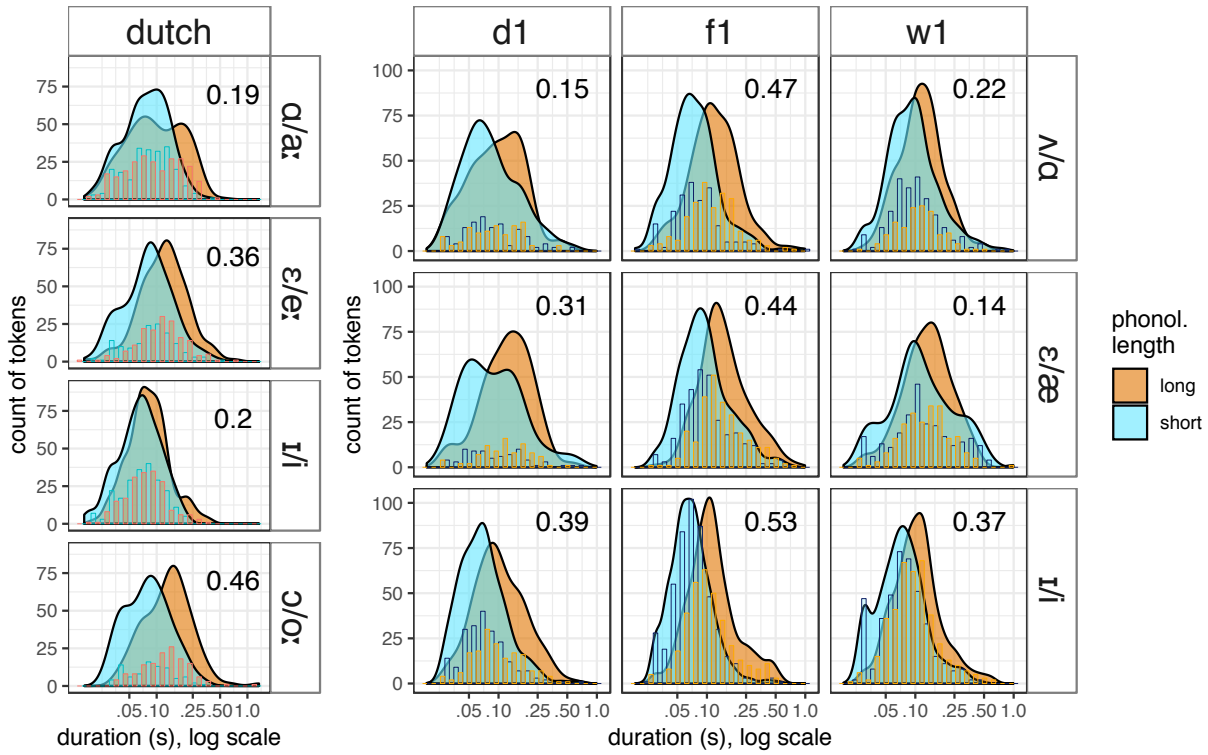

Figure 1. Density plots of vowel durations. Each plot shows one {long,short} pair, with a different pair in each row, and a different mother in each column. The Dutch data are shown on the left, English on the right. The plot for phonologically long vowels is filled in a dark color, short vowels a light color. Duration is given on the x-axis in logarithmically scaled space. The y-axis shows the counts in the histograms. The smoothed density plots are arbitrarily scaled on the y-axis, though each is scaled the same way. The number in the upper right corner or each plot is the rank-biserial correlation between duration and phonological length; higher numbers indicate greater separation.

lengthening into consideration, and it is not known whether they do. Here, we evaluate the consequences of modeling the utterance-final duration effect by examining the residuals of a simple linear regression partialing out the effects of utterance-final position. Because the logs of the durations are closer to normally distributed than the raw durations, the regression was done over logged durations.

First, we note that as expected, utterance-final vowels were substantially longer than

utterance-medial vowels. The median utterance-final vowel in Dutch was 153 ms long, compared with a median utterance-medial duration of 92 ms; for English the equivalent figures were 156, 142, and 152 ms (final) versus 71, 80, and 90 ms (medial) for each English corpus. In essence, the regression analysis served to place final and medial vowels on equal footing.

The results of this analysis are given in Table 2 (the density plots are given in the Supporting Online Materials). Once again, the degree of overlap in each pair is unexceptional in the Dutch case.

Having failed to discover a convincing bimodal separation of Dutch vowels into phonetically short and long categories (cf. Bion et al., 2013), it seemed reasonable to question whether some prosodically unusual words might be muddying the distributional pattern. For example, the word *ja* 'yes', which frequently occurs as the only word in an utterance, has a very broad range of durations (IQR 96–537 ms), in keeping with the many different sorts of meanings such an utterance might be employed to convey. Could it be that children set aside certain words, or certain types of words, in estimating their language's phonological properties?

Tokens from different word categories

The words in each corpus were exhaustively sorted into one of three categories: *content* words, *demonstratives*, and *function* words. The first intuition behind this division is that when learning phonology, children might set aside as exceptional those demonstrative words that are unusual in their discourse role by tending to appear alone and often with rather distinctive intonation contours, much as children might also exclude songs or animal sounds when learning prosody. Demonstratives were categorized as words like (English) *no, yay, whoops, yum, thanks, oops, meow, hi.* . . . These judgments were made type by type, considering the context in which the words were said in the corpora. The second intuition behind the division was that there might be something distinctive about *function* words, that might lead children to tend to consider them differently when learning about prosody (e.g., Shi, Werker, & Morgan, 1999). It is not really clear how such a category should be bounded. The approach taken here was to classify closed-class

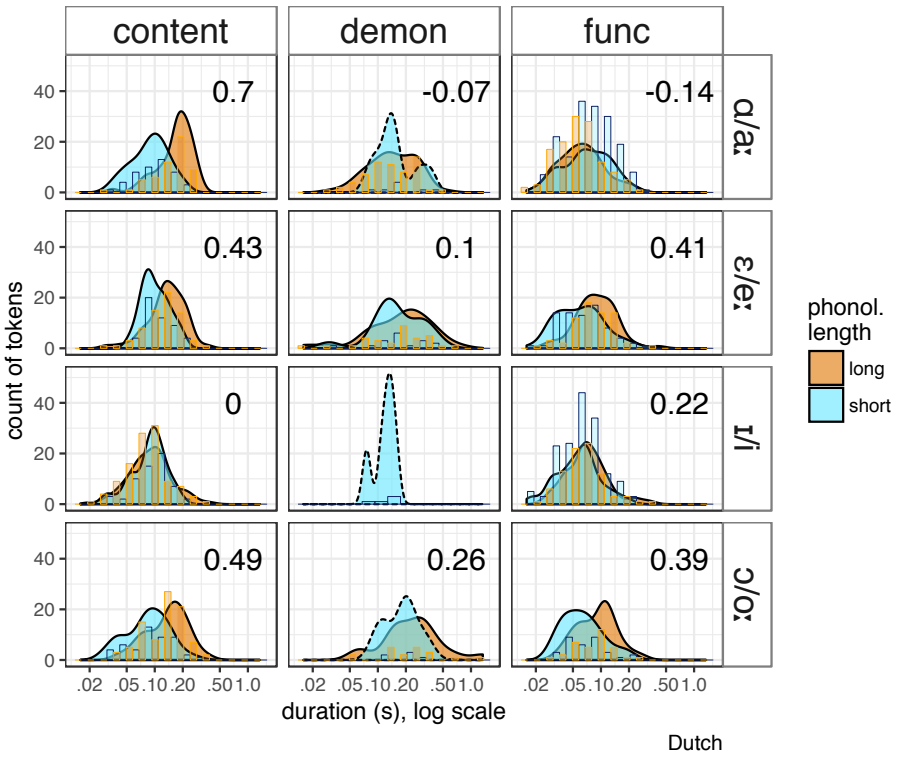

Figure 2. Dutch vowels' duration distributions, by pair, split according to word category (content word, demonstrative, function word), shown as histograms and density plots. Curves representing fewer than 10 tokens are bounded with dotted lines. The degree of separation between phonologically long and short vowels is operationalized using rank biserial correlation, indicated in the upper right of each panel, where larger numbers mean greater separation. Note that in Dutch, [ɪ-i] is not expected to show a duration distinction.

words as function words, including cases like *a, and, anymore, does, gonna, than*, most locative prepositions, and wh-words. Remaining as *content* words were those that children tend to learn when their vocabulary expands–words like *baby, carry, cow, pear, sad, summer*, and proper names.

Turning first to the Dutch dataset, content words did exemplify the phonological length distinction more clearly than the demonstrative words or function words. Density plots and the corresponding rank-biserial correlation coefficients are given in Figure 2.

The separation of the [ɑ--aː] vowels in Dutch content words (rbcc = 0.70) far exceeds the separation of the sample as a whole (rbcc = 0.19, as shown in Figure 1), suggesting that the phonological opposition conveyed by duration was being obscured by demonstratives (like *ja*) and function words, given that for these categories the nominally short vowels were slightly longer in duration than the nominally long ones (hence the negative correlation coefficients). In the cases of the pairs [ɛ--eː] and [ɔ--oː] separation was only slightly greater when considering content words alone. For the pair [ɪ--i], which in Dutch is not said to contrast by phonetic duration, the separation among content words was measured at zero.

How do these results compare to English? An analogous plot for Brent mother D1 is displayed in Figure 3; the other two English cases are presented in the Supporting Online Materials. Mother D1 showed reasonably strong separation in content words for all three pairs, with the greatest separation for [ɪ--i]. Separation between categories was again greater, overall, among content words than among functors, a pattern that held for the other two mothers as well, with a couple of exceptions. Comparison of Dutch and English suggests that with this analysis, Dutch begins to exhibit more separation between long and short categories than English does, overall. The rank-biserial coefficients for the content words in Dutch pairs (.70, .43, .49, excluding [ɪ-i]) are large compared with the English datasets overall, though mother F1's separation is almost as strong (.34, .46, .58). Mother W1 shows less separation in content words (.13, .14, .37) and mother D1 is in between (.23, .35, .40).

Thus, one possible explanation for Dutch children's greater reliance on vowel duration in categorizing words is that in their experience, the content words they are hearing (and that are being tested in experiments) have phonetically shorter or longer vowels depending on the vowels' phonological duration status. A weakness of the support for this hypothesis is that at least one English-speaking mother nearly matched the Dutch levels of phonetic distinction; the separation between the languages based on this analysis is not very consistent. Another weakness of the hypothesis is that it assumes that children divide vowel tokens according to a rough categorization of the words hosting the vowels, but then collapse over these words, by category, when judging

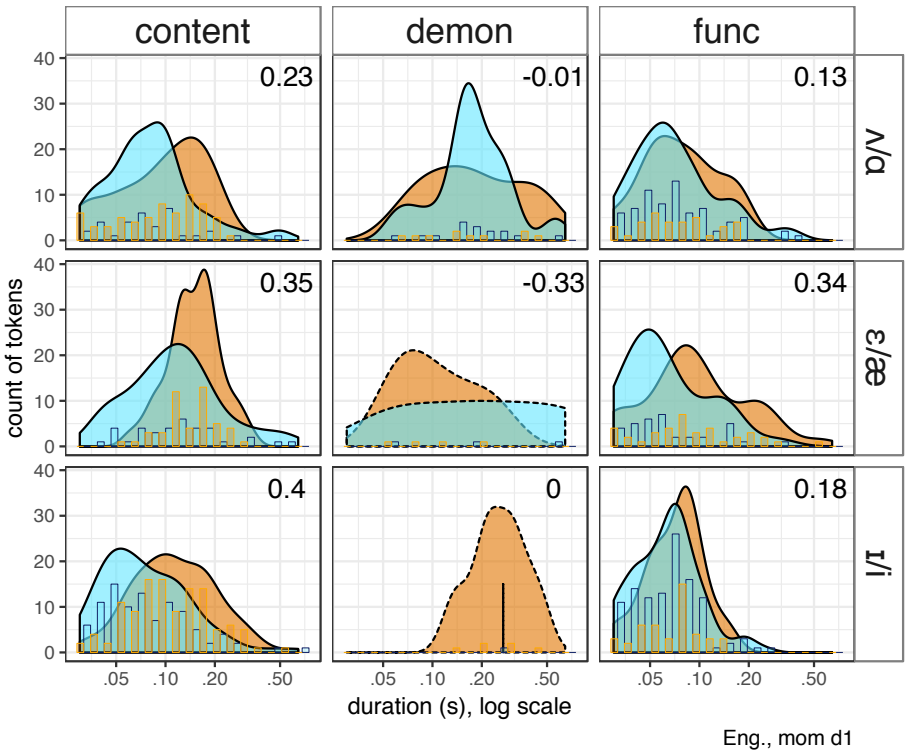

Figure 3. English mother D1's duration distributions, by pair, split according to word type (content word, demonstrative, function word). Plotting conventions are as in Figure 2

how the vowels work phonetically. Intuitively, it seems more plausible that if children take the kind of word into account, they analyze the phonology word by word, rather than mixing tokens of a class of words. Proposals of this nature have been suggested previously for explaining how infants learn phonetic categories (e.g., Feldman, Myers, White, Griffiths, & Morgan, 2013; Swingley, 2009; Thiessen, 2007). Analyses that do precisely this but for duration categorization are presented next. As we shall see, these reveal a more consistent distinction between Dutch and English.

Word types

Because we were interested in differentiating type statistics from token-level statistics, words were only included in this analysis if they occurred at least twice in a given corpus. The differentiation into three basic categories (content, demonstrative, and function) was retained.

Words were identified by their orthographic transcription, and the vowel(s) characteristic of a given word type were assigned according to the dominant transcription, as pronounced by a given mother. Once again our main outcome measure was the separation of the phonologically short and long categories, as characterized using rank-biserial correlation.

Figure 4 displays mean duration values for each of 49 Dutch [ɑ--aː] words. Correlation coefficients are given for content words and function words (demonstratives were too infrequent for correlations to be meaningful). Remarkably, the content words for this pair were almost perfectly separable by duration, with only the two words *appels* and *waait* standing in the way of a correlation of 1.0. This case is important because this particular Dutch pair is the one for which evidence is strongest, both in toddlers and in adults, that interpretation of native listeners can be affected by phonetic duration. Analogous plots for the other pairs are given in the Supporting Online Materials.

By contrast, Figure 5 displays an analogous plot for English mom W1's contrast of [ɛ--æ], chosen here because this pair shows durational effects on interpretation in judgment experiments (e.g., Hillenbrand, Clark, & Houde, 2000). Here, although there is some separation between the distribution (only at the short-duration end), there is also substantial overlap. Most of the content words this mother said with an [ɛ] vowel were as long as several content words with the [æ] vowel, and there were *no* [æ] words that outranked all of the [ɛ] words. The full set of plots of this sort is given in the Supporting Online Materials.

Figure 6 summarizes the word-type results for three pairs of Dutch vowels and three pairs of English vowels. (No difference is predicted for Dutch [ɪ--i], so that panel is omitted here.) Three patterns in the data are evident. First, in the case of Dutch, separation of the phonologically long and short vowels is quite strong in all cases except that of functors in the /a/ pair. In particular the /a/ pair's content words reach a rank-biserial correlation coefficient of 0.88, which is unmatched in the rest of the comparisons.

Second, there are no cases in any of the English comparisons where a phonetic duration threshold would perform well in separating the phonologically short and long vowels. Indeed,

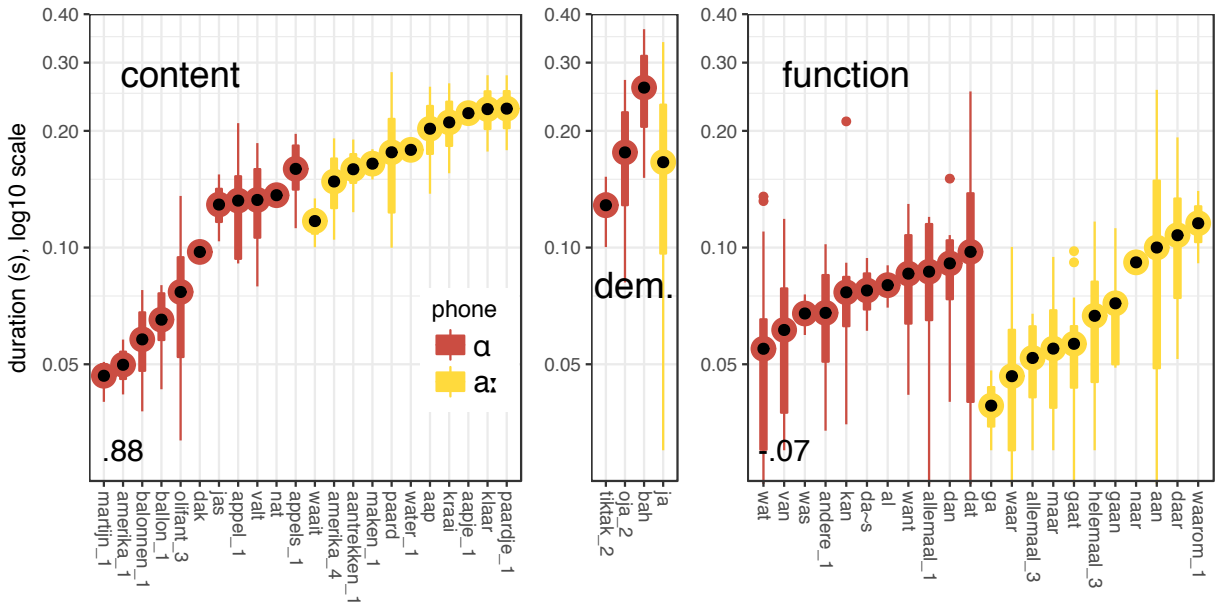

Figure 4. Dutch duration distributions for individual word types containing the vowels [ɑ] or [aː]. Each column presents the mean duration (black point) and boxplot (IQR and range) for each word type, with the relevant syllable number indicated after the word along the x axis. The phonologically short vowel is shown with the darker symbols (red online), the long vowel with lighter symbols (yellow online). Content words, demonstratives, and function words are shown in separate panels. Numbers at lower left in the first and third panels are rank-biserial correlation coefficients.

except for the Dutch [ɑ-aː], this is true for Dutch as well: in general, there are too many words with nominally long vowels and phonetically short average durations, and vice versa. This is itself not surprising, given the mixed nature of the duration distributions over tokens.

Third, when the word-type distributions did show visible separation (and the correlations were high), in English, this was often because the phonologically shorter vowel distributions were anchored by many phonetically short words, whereas in Dutch, it was often the phonetically long words that were distinct. We can quantify this by looking at the phonetically longest words in each of the comparisons. If for each of the 24 long/short pairings displayed on the plot (leaving out the demonstrative words: 4 corpora x 3 vowel pairs x 2 word categories) we take the longest

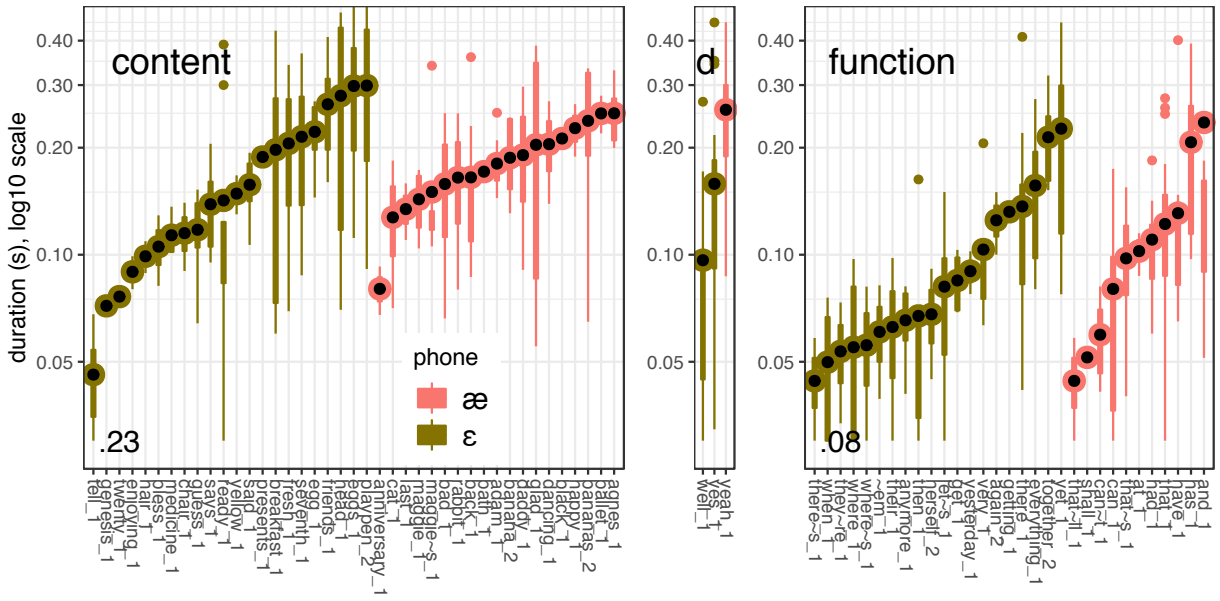

Figure 5. English duration distributions for individual word types containing the vowels [ɛ] or [æ] for mother W1. Conventions are as in Figure 4. The right panel omits one token of *and* that was 990 ms long.

1/3 of the words, in terms of mean phonetic duration, and examine the proportion of phonologically long-vowelled words in that set, in Dutch content words this yields 100% [aː] words, 100% [eː] words, and 80% [oː] words. Thus, phonologically short words rarely emerged as consistently phonetically long in Dutch. In English there were more phonologically short words surfacing with phonetically long vowels, reducing the analogous proportions to 54.8% ([ɛ,æ]), 68.5% ([ʌ,a]), and 68.6% ([ɪ,i]), averaging over the 3 English datasets. (Considering the function words, these figures are either similar between English and Dutch, or yield purer sets of phonologically long words in the Dutch case.) The same pattern holds if (for example) we consider the top half of the words rather than the top third.

The preceding analyses over types consider the case in which children might give special weight to content words. How separable are the pairs considering all words together, whether content words, function words, or demonstratives? In this analysis Dutch is less distinctive. The RBCC coefficients are given in Table 3. This suggests that demonstratives and function words

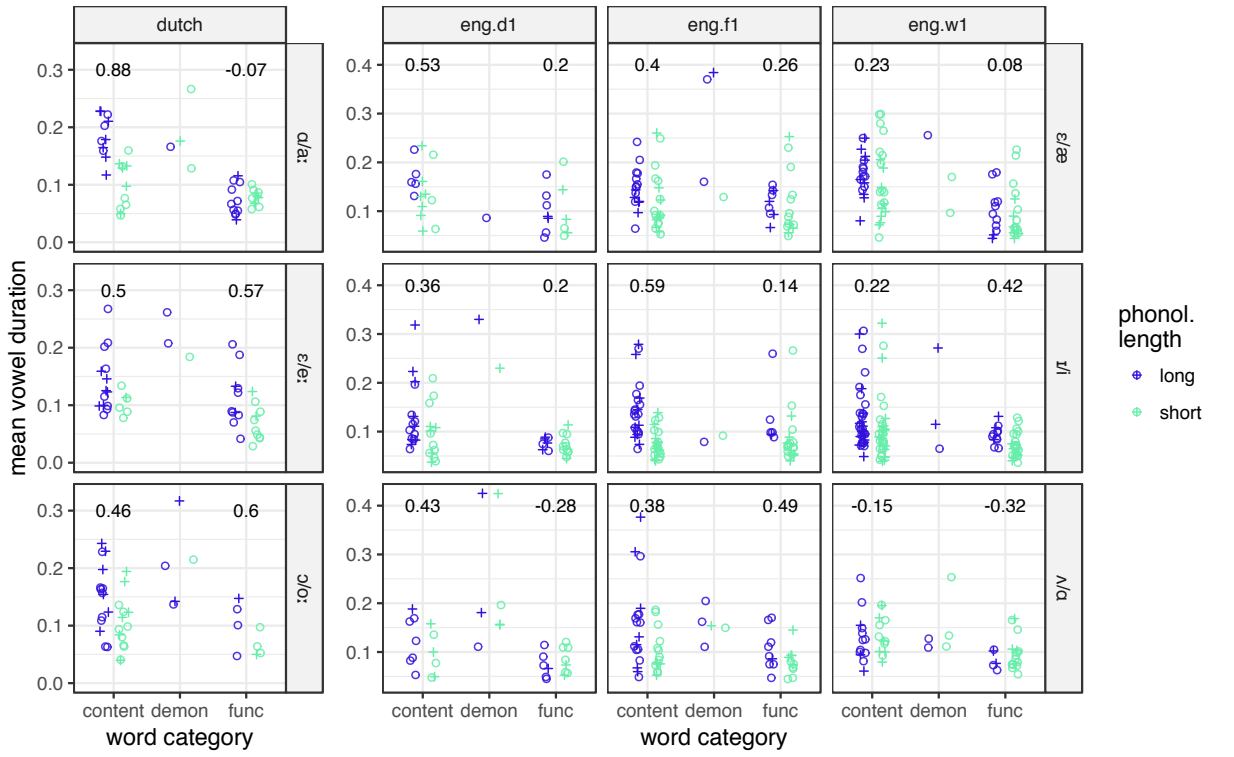

Figure 6. Duration distributions for the phonologically long and short vowels in individual words, for the 3 pairs in each language where phonetic differences would be expected. Columns of panels present data from different corpora, with Dutch on the left. Rows of panels correspond to vowel pairs. Within each panel, words are divided by category: content words, demonstratives, and function words. Darker points, on the left, are phonologically long vowels; lighter points are phonologically short vowels. Words occurring twice in each corpus are plotted with plus signs; words occurring more than twice are plotted with circles. Numbers above each pair of sets of points are rank-biserial correlation coefficients, which were computed when the number of word types exceeded ten. The data in the upper left plot correspond to the data in Figure 4.

display phonological duration less clearly phonetically than content words do.

In most of the preceding analyses we have assumed that children's representations of word duration are based on an average of their experiences with words (estimated using the sample we have). It is possible that in fact children do not interpret durations "as given" in the signal, but

corpus	pair (short, long)	rbcc
Dutch	ɑ, aː	0.27
Dutch	ɛ, eː	0.54
Dutch	ɔ, oː	0.52
English (d1, f1, w1)	ɛ, æ	0.11, 0.37, 0.26
English (d1, f1, w1)	ʌ, a	0.07, 0.39, -0.06
English (d1, f1, w1)	ɪ, i	0.40, 0.42, 0.35

Table 3

By-types analysis of separation between long and short vowels, quantified as rank-biserial correlation coefficient, collapsing over different word categories (considering content words, demonstratives, and function words together).

compute lexical representations based on durations that have been adjusted in various ways, taking into account aspects of context. For example, as described earlier, children might downweight long durations that come at the ends of utterances, given that such durations might be attributable in part to the context and not the vowel category. Presently there is no direct evidence that toddlers do anything of the sort. Still, to explore this issue in a types-based analysis we considered the results of the linear model (estimated separately for each child) regressing (log) duration on the binary variable of whether a vowel was the final vowel in an utterance or not, and using the average residual as the word's duration representation. The results of this manipulation were mixed. RBCC values for the three Dutch pairs (considering the content words) were still fairly high (.57, .45, and .55 for the /a/, /e/, and /o/ pairs), but the English samples were variable, with mom f1 competitive with the Dutch values (.33, .51, and .51 for the /a/, /æ/, and /i/ pairs) and the others rather lower (d1: .20, .40, .36; w1: .22, .31, .33). Thus, it appears that if children take utterance position into account in a way that resembles our regression analysis, this would not seem to clarify the durational opposition strongly in Dutch, nor strongly dismiss it in English.

Figure 6 shows variability in the RBCC values for a given pair among the English mothers.

We cannot say whether this represents true individual variation, or variability due to the samples (which are small relative to infants' experience, of course). It is possible that the quantitative amount of overlap in a pair is interpreted by children in a monotonic way as a signal of the probability that the vowels are not differentiated by phonetic duration. If so, then we would expect the child of mother d1, for example, to be more affected by duration manipulations in word recognition than the child of mother w1. Alternatively, perhaps there is some threshold above which children simply draw the conclusion that durational differences are part of the segmental phonology of the language. In principle, sensitivity experiments together with measurements of the language environment could settle this issue.

For all four corpora, there was a substantial number of words whose vowels did not appear to exhibit phonological length phonetically, at least not if one simply measures raw durations. Even when examining mean durations over word types, the predominant pattern was one of massive overlap in duration distributions, with separation being the exception. That being said, it is in looking at words that Dutch showed greater distinctiveness than English in comparison of phonological length. This result is consistent with the hypothesis that children learn about contrastiveness in interpreting readily perceptible phonetic variation by paying attention to the phonetic characteristics of words in the lexicon.

Conclusions

It is widely assumed that the development of language-appropriate attention to a given phonetic feature for determining phonological category membership is primarily dependent on the detection of a bimodal (or multimodal) distribution of heard instances along the relevant phonetic dimension, indicating to the child that two (or more) categories are present. Laboratory studies have demonstrated that this sort of learning is feasible for infants, at least for some speech dimensions (e.g., Cristià et al., 2011; Maye et al., 2002). However, as described in the Introduction, there are good reasons to expect that for a dimension like duration, the distributions might not be transparently available on the surface, because duration is put to so many uses in

language.

What we have observed in the present study is that duration distributions in child-directed English and Dutch are, to some degree, bimodal, when comparing instances of spectrally similar vowel categories, but that in the two languages the degree of overlap among long and short vowels was considerable, and overlap was not markedly less for Dutch than for English. This result suggests that surface distributions of vowel durations might not provide the experience that teaches Dutch toddlers, but not English ones, to use duration to distinguish words such as [tam] and [taːm] (Dietrich et al., 2007). Bion et al. (2013) came to a similar conclusion in considering the nominally long and short vowels of child-directed Japanese speech.

However, once we considered the duration distributions of word *types*, rather than vowel *tokens*, the Dutch distributions were differentiated to a greater degree (and in one case, a much greater degree) than the English ones. For the Dutch pair [ɑ,aː], this required considering just the content words, and leaving off the function words and demonstratives. It might be that toddlers' intuitions about phonology rely on the words that they are most likely to learn, and if so, the remarkable durational separation between the [ɑ] and [aː] of Dutch content words could explain why Dutch toddlers treat these as distinct. Admittedly, there is no evidence at present that children take some words to be more important than others during phonetic learning (a question that does not seem to have been tested experimentally).

The observations presented here offer the suggestion that Dutch and English learners come to differ in their interpretations of vowel duration by paying attention to words. Of course, this evidence is far from being a proof; the argument is an inferential one based on what learning options the data support, not based on detailed measurements of children's learning process itself. Also, as a study of Dutch and English learning environments, the most important limitation is that the Dutch sample came from only one speaker. It is possible that other speakers would not show the same pattern. Individual speakers of a given dialect can vary in diverse ways; for example, some seem to hyperarticulate more than others. That being said, it is not clear how this particular sort of variation would lead talkers to vary in the degree to which their type, but not token,

durational distributions signal a length distinction. At present we must consider this question an avenue for further research.

Other results in the literature accord well with the notion that type-level statistics are relevant in phonetic learning. Such effects have been shown in several experimental studies of infants and toddlers (Feldman et al., 2013; Thiessen, 2007; Thiessen & Yee, 2010; Yeung & Werker, 2009) and supported in corpus modeling work on the acquisition of vowel categories (Swingley & Alarcon, 2018). In adults, some phonological generalizations appear to track type frequency rather than token frequency (Hay, Pierrehumbert, & Beckman, 2004), and in children, phonological representations have been argued to gain fidelity through participation in multiple different words (Edwards, Beckman, & Munson, 2004).

How might children use type-level statistics to learn about vowel duration? Viewpoints differ on the degree to which early lexical representations are composed of sequences of "digital" labels, i.e. strings of identifiable consonants and vowels, as opposed to consisting of more holistic phonetic objects (e.g., Beckman & Edwards, 2000; Ramon-Casas, Swingley, Sebastián-Gallés, & Bosch, 2009; Werker & Curtin, 2005), but we assume that by 18 months, children's representations of familiar words contain a specification of the words' typical constituent consonants and vowels. We also consider it likely that in languages like English and Dutch, children come to appreciate the spectral (formant) characteristics of their vowel categories earlier than they come to grips with the characteristics of vowels that are strongly implicated in prosodic regularities, including duration, because of the diverse and complex range of influences on these characteristics. (We will explore weakening this assumption in a moment.)

Given these assumptions, Dutch children could learn that (e.g.) the [ɑ] and [a] categories differ in duration as well as quality by noting that content words with [ɑ] (defined by vowel quality) typically have a phonetically short vowel, whereas content words with [a] (also defined by vowel quality) typically have a phonetically long vowel, where "short" and "long" could be defined with reference to the sets of relevant words themselves. Thus, for example, the prototypical [a] in *paard* is longer than the [ɑ] typical of essentially all of the familiar words with

[ɑ]. English learners do not have a strong basis for drawing such a conclusion. When the mental processes responsible for phonological generalizations consider the English [ɛ] and [æ] words, for example, these words' vowels overlap almost entirely in their durations.

Do children really learn phonetic attributes like duration later than they learn other phonetic attributes of phonological categories? Not much data speaks to this point. Infants are certainly attentive to duration; for example, geminate consonants stand out to them more than singleton consonants (Vihman & Majorano, 2017). Two developmental experiments considered Japanese and English learners' discrimination of vowel duration contrasts, providing bit of additional purchase on this question. Sato, Sogabe, and Mazuka (2010) found that 4- and 7.5-month-old Japanese-learning infants did not discriminate a durational difference with a 2:1 ratio, though 9.5-month-olds did succeed in telling these vowels apart, like the Japanese 10-month-olds tested by Mugitani et al. (2009). This might indicate learning occurring only a few months later than learning has been first reported for vowel quality distinctions (e.g., Polka & Werker, 1994). However, Mugitani et al. also found in a comparison of English and Japanese learners that when their Japanese and English groups differed from one another, at 18 months, the English-learning toddlers discriminated the length contrast, while the Japanese toddlers only discriminated a change in one direction, from a long baseline to a short test vowel. This would not appear to be an adaptive development relative to the English toddlers, given that Japanese has a length contrast and English does not.

A third study tested whether English learners would react with longer listening to syllables violating the typical pattern of vocalic lengthening before a voiced coda (Ko, Soderstrom, & Morgan, 2009). Eight-month-olds showed no listening preference, whereas 14-month-olds listened longer to syllables containing short vowels with voiced codas than syllables containing short vowels with unvoiced codas. Ko et al. argue that English learners begin to detect the phonological patterning of vowel duration and coda voicing between 8 and 14 months. This developmental range overlaps with the time period in which infants are typically viewed as

discovering their language's consonant categories.[1]

Given the prevailing uncertainty over whether children learn vowel duration features later than they learn the primary features of their language's speech sounds, it is sensible to consider how children might learn quantity and quality features at the same time. This takes us back to the proposal that words are used in the discovery of phonetic categories (Feldman et al., 2013; Swingley, 2009). Infants' first words may be learned when their phonological categories are only minimally adapted to the native language. Perhaps they achieve this by noting repetitions of similar-sounding sequences of speech, probably with some correlated semantic information, that coalesce into discrete proto-lexical items (Feldman et al., 2013; Kamper, Jansen, & Goldwater, 2016; Park & Glass, 2008). Once infants are familiar with some words, they might detect that portions of those words are similar across instances of the same word; for example, that *mommy* always starts the same way, or that *dog*'s vowel consistently occupies a restricted range of the F1/F2 space. Infants may also recognize that the speech-sound categories that link instances of the same word (*mommy* always starts with /m/) also link similarities in lexical representations *across* words, noting the similarity of the initial consonant of *mommy* and *milk*, or the vowel of *dog* and *ball*. If infants learn their language's speech-sound categories by detecting overlap in their representations of portions of familiar words, such a process could also lead Dutch learners to emphasize vowel duration more than English learners do, because the Dutch words they are learning exemplify correlations between duration and regions of vowel quality space. As we have seen, these correlations are weaker when considering the mass of Dutch vowels than when considering the durations typical of individual word types.

The introduction to this paper described the complexity of arriving at a language-appropriate *model* of the sound pattern of utterances. To identify the words in a spoken sentence, we do not simply identify a list of consonants and vowels, and then work out the most

[1]The regularity studied by Ko et al. (2009) concerns an allophonic rule, and not just phonetic categories as acoustic objects. We acknowledge, of course, that learning phonology involves much more than learning phonetic categories; e.g., Peperkamp, Le Calvez, Nadal, & Dupoux, 2006.

likely lexical sequence. We try to account for the phonetic signal as it comes to us, where the explanatory features include quite a broad range, including talker characteristics, discourse features, prosodic groupings, idiosyncrasies of the lexicon, phonetic context, and others, as well as the quantity and quality features that specify particular consonants and vowels. Even if talkers actually articulated the canonical speech-sound sequence that a pronouncing dictionary might supply (and they don't; Johnson, 2003; Hawkins, 2003), learners would still need to uncover what seems to be quite a complex set of quantitative sources of influence on the signal. These sources of influence readily trespass linguistic and other levels of description; for example, duration can be affected by just about anything, including speaking rate, vowel identity, lexical stress, novelty of a word in the discourse, or loudness of background noise.

It is unlikely that infants work out the full model all at once; it is more plausible that they start from the most readily discoverable islands of continuity and proceed from there. These islands may often be words. What we found here is that teasing apart a fairly subtle crosslinguistic difference did not require a full-on model of prosodic structure; it only required categorization of vowels and words. This does not mean that toddlers are limited to lexical and segmental models of duration; it only suggests that the empirically observed difference between Dutch and English one-year-olds could be accounted for without presupposing more sophisticated knowledge of each language's duration model. If we are wrong about this, and the greater distinction for Dutch (in content words) turns out not to characterize larger and more diverse samples, we return to the puzzle we started with. What other sources of information might Dutch children have access to? One possibility is that they have already developed a model of prosodic interpretation under which the segmental duration distinctions are clearer, because other sources of variance have been accounted for. Another is that there are distinctive phonetic features of Dutch long and short vowels that separate the categories (features beyond the formant measurements we examined already).

For the most part, research on young children's interpretation of speech has focused on demonstrating children's achievement of developmental milestones and describing individual

differences and their correlates. Attempts to *explain* these developments have often come in the form of artificial-language experiments that exemplify a linguistic regularity very clearly in a tiny language sample, and demonstrate children's ability to detect and generalize that regularity. Such studies by their nature present analogies to real language learning, but the fidelity of the analogy is open to question. Crosslinguistic studies, in which the experimental conditions are actual languages being learned by children, provide an essential counterpart to laboratory learning studies. But to interpret the learning that crosslinguistic experiments on children's speech interpretation imply, we need quantitative descriptions of the languages children are learning. We cannot understand how children learn something without knowing about the data they are exposed to. Such descriptions often have characteristic flaws, including sacrificing breadth for depth or vice versa. The present work is no exception, and so the conclusions admit reasonable questions about generality (just as laboratory learning studies do). Within these limitations, though, our examination of child-directed speech suggests that children could learn to treat vowel duration in Dutch contrastively without having access to a full, adult-like model of duration, if children were to consider the distinctiveness of durational patterns over word types rather than over tokens.

References

Adank, P., van Hout, R., & Smits, R. (2004). An acoustic description of the vowels of Northern and Southern Standard Dutch. *Journal of the Acoustical Society of America*, *116*, 1729-1738.

Adriaans, F., & Swingley, D. (2017). Prosodic exaggeration within infant-directed speech: Consequences for vowel learnability. *Journal of the Acoustical Society of America*, *141*, 3070-3078. doi: 10.1121/1.4982246

Beckman, M. E., & Edwards, J. (2000). The ontogeny of phonological categories and the primacy of lexical learning in linguistic development. *Child Development*, *71*, 240-249. Retrieved from `http://dx.doi.org/10.1111/1467-8624.00139` doi: 10.1111/1467-8624.00139

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. doi: 10.1371/journal.pone.0051594

Boersma, P., & Weenink, D. (2001). Praat, a system for doing phonetics by computer. *Glot International*, *5*, 341-345.

Booij, G. (1999). *The phonology of Dutch*. Oxford University Press.

Brent, M. R., & Siskind, J. M. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition*, *81*, B33-B44. doi: /10.1016/s0010-0277(01)00122-6

Cambier-Langeveld, T. (1997). The domain of final lengthening in the production of Dutch. In J. A. Coerts & H. de Hoop (Eds.), *Linguistics in the netherlands* (Vol. 14, p. 13-24). Benjamins. doi: 10.1075/avt.14.04cam

Chládková, K., Escudero, P., & Lipski, S. C. (2015). When "aa" is long but "a" is not short: speakers who distinguish short and long vowels in production do not necessarily encode a short–long contrast in their phonological lexicon. *Frontiers in Psychology*, *6*, 438. doi: 10.3389/fpsyg.2015.00438

Cho, T. (2015). Language effects on timing at the segmental and suprasegmental levels. In

M. Redford (Ed.), *The handbook of speech production* (pp. 505–529). John Wiley and Sons, Oxford.

Cristià, A., McGuire, G. L., Seidl, A., & Francis, A. L. (2011). Effects of the distribution of acoustic cues on infants' perception of sibilants. *Journal of phonetics*, *39*(3), 388–402.

Dietrich, C., Swingley, D., & Werker, J. F. (2007). Native language governs interpretation of salient speech sound differences at 18 months. *Proceedings of the National Academy of Sciences of the USA*, *104*, 16027-16031. doi: /10.1073/pnas.0705270104

Edwards, J., Beckman, M. E., & Fletcher, J. (1991). The articulatory kinematics of final lengthening. *Journal of the Acoustical Society of America*, *89*(1), 369–382.

Edwards, J., Beckman, M. E., & Munson, B. (2004). The interaction between vocabulary size and phonotactic probability effects on children's production accuracy and fluency in nonword repetition. *Journal of Speech, Language, and Hearing Research*, *47*, 421-436. doi: 10.1044/10924388.2004.034

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*, 751-778. doi: 10.1037/a0034245

Feldman, N. H., Myers, E. B., White, K. S., Griffiths, T. L., & Morgan, J. L. (2013). Word-level information influences phonetic learning in adults and infants. *Cognition*, *127*(3), 427–438.

Fikkert, P. (1994). *On the acquisition of prosodic structure*. Leiden, The Netherlands: Holland Institute of Generative Linguistics.

Fujisaki, H., Nakamura, K., & Imoto, T. (1975). Auditory perception of duration of speech and non-speech stimuli. In G. Fant & M. Tatham (Eds.), *Auditory analysis and perception of speech* (p. 197-219). London: Academic Press.

Gonzalvo, X., Tazari, S., Chan, C.-a., Becker, M., Gutkin, A., & Silen, H. (2016). Recent advances in Google real-time HMM-driven unit selection synthesizer. In *Interspeech* (pp. 2238–2242).

Goudbeek, M., Smits, R., & Swingley, D. (2009). Supervised and unsupervised learning of

multidimensional auditory categories. *Journal of Experimental Psychology: Human Perception and Performance*, *35*, 1913-1933.

Hallé, P. A., & de Boysson-Bardies, B. (1994). Emergence of an early receptive lexicon: infants' recognition of words. *Infant Behavior and Development*, *17*, 119-129.

Hawkins, S. (2003). Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of Phonetics*, *31*(3-4), 373–405.

Hay, J., Pierrehumbert, J., & Beckman, M. (2004). Speech perception, well-formedness, and the statistics of the lexicon. *Papers in laboratory phonology VI*, 58–74.

Henter, G. E., Ronanki, S., Watts, O., Wester, M., Wu, Z., & King, S. (2016). Robust TTS duration modelling using DNNs. In *Acoustics, speech and signal processing (icassp), 2016 ieee international conference on* (pp. 5130–5134). doi: 10.1109/ICASSP.2016.7472655

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America*, *97*, 3099-3111.

Hillenbrand, J. M., Clark, M. J., & Houde, R. A. (2000). Some effects of duration on vowel recognition. *Journal of the Acoustical Society of America*, *108*, 3013-3022.

Johnson, E. K., Jusczyk, P. W., Cutler, A., & Norris, D. (2003). Lexical viability constraints on speech segmentation by infants. *Cognitive Psychology*, *46*, 65-97.

Kamper, H., Jansen, A., & Goldwater, S. (2016). Unsupervised word segmentation and lexicon discovery using acoustic word embeddings. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, *24*(4), 669–679.

Klatt, D. H. (1976). Linguistic uses of segmental duration in English: acoustic and perceptual evidence. *Journal of the Acoustical Society of America*, *59*, 1208-1221.

Klatt, D. H. (1987). Review of text-to-speech conversion for English. *Journal of the Acoustical Society of America*, *82*, 737-793.

Ko, E.-S., Soderstrom, M., & Morgan, J. (2009). Development of perceptual sensitivity to extrinsic vowel duration in infants learning american english. *The Journal of the Acoustical Society of America*, *126*(5), EL134–EL139. doi: 10.1121/1.3239465

Levelt, C. (1994). *On the acquisition of place*. Leiden, The Netherlands: Holland Institute of Generative Linguistics.

Ling, Z.-H., Kang, S.-Y., Zen, H., Senior, A., Schuster, M., Qian, X.-J., … Deng, L. (2015). Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends. *IEEE Signal Processing Magazine*, *32*(3), 35–52. doi: 10.1109/MSP.2014.2359987

Liu, L., & Kager, R. (2016). Perception of a native vowel contrast by dutch monolingual and bilingual infants: A bilingual perceptual lead. *International Journal of Bilingualism*, *20*(3), 335–345.

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

Mugitani, R., Pons, F., Dietrich, C., Werker, J. F., & Amano, S. (2009). Perception of vowel length by Japanese- and English-learning infants. *Developmental Psychology*, *45*, 236-247. doi: 10.1037/a0014043

Nooteboom, S. G., & Doodeman, G. J. N. (1980). Production and perception of vowel length in spoken sentences. *Journal of the Acoustical Society of America*, *67*, 276-287.

Park, A. S., & Glass, J. R. (2008). Unsupervised pattern discovery in speech. *IEEE Transactions on Audio, Speech, and Language Processing*, *16*, 186-197.

Peperkamp, S., Le Calvez, R., Nadal, J.-P., & Dupoux, E. (2006). The acquisition of allophonic rules: Statistical learning with linguistic constraints. *Cognition*, *101*, B31–B41.

Pitt, M. A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., & Fosler-Lussier, E. (2007). *Buckeye corpus of conversational speech (2nd release)*. Columbus, OH: Department of Psychology, Ohio State University.

Polka, L., & Werker, J. F. (1994). Developmental changes in perception of nonnative vowel contrasts. *Journal of Experimental Psychology: Human Perception and Performance*, *20*, 421-435.

Ramon-Casas, M., Swingley, D., Bosch, L., & Sebastián-Gallés, N. (2009). Vowel categorization

during word recognition in bilingual toddlers. *Cognitive Psychology*, *59*, 96-121. doi: /10.1016/j.cogpsych.2009.02.002

Sato, Y., Sogabe, Y., & Mazuka, R. (2010). Discrimination of phonemic vowel length by japanese infants. *Developmental Psychology*, *46*, 106.

Shi, R., Werker, J. F., & Morgan, J. L. (1999). Newborn infantsIJôòù sensitivity to perceptual cues to lexical and grammatical words. *Cognition*, *72*(2), B11–B21.

Stager, C. L., & Werker, J. F. (1997). Infants listen for more phonetic detail in speech perception than in word-learning tasks. *Nature*, *388*, 381-382.

Swingley, D. (2009). Contributions of infant word learning to language development. *Philosophical Transactions of the Royal Society B*, *364*, 3617-3622. Retrieved from http://dx.doi.org/10.1098/rstb.2009.0107 doi: 10.1098/rstb.2009.0107

Swingley, D., & Alarcon, C. (2018). Lexical learning may contribute to phonetic learning in infants: a corpus analysis of maternal Spanish. *Cognitive Science*, *42*, 1618-1641. doi: 10.1111/cogs.12620

Swingley, D., & Aslin, R. N. (2000). Spoken word recognition and lexical representation in very young children. *Cognition*, *76*, 147-166.

Thiessen, E. D. (2007). The effect of distributional information on children's use of phonemic contrasts. *Journal of Memory and Language*, *56*, 16-34.

Thiessen, E. D., & Yee, M. N. (2010). Dogs, bogs, labs, and lads: What phonemic generalizations indicate about the nature of childrenIJôòùs early word-form representations. *Child Development*, *81*(4), 1287–1303. doi: 10.1111/j.1467-8624.2010.01468.x

Tillman, G., Benders, T., Brown, S. D., & van Ravenzwaaij, D. (2017). An evidence accumulation model of acoustic cue weighting in vowel perception. *Journal of Phonetics*, *61*, 1–12. doi: 10.1016/j.wocn.2016.12.001

van Santen, J. P. H. (1992). Contextual effects on vowel duration. *Speech Communication*, *11*, 513-546.

Vance, T. J. (1987). *An introduction to Japanese phonology*. Albany: State University of New

York Press.

Van der Feest, S. V. H., & Swingley, D. (2011). Dutch and English listeners' interpretation of vowel duration. *The Journal of the Acoustical Society of America*, *129*(3), EL57–EL63. doi: 10.1121/1.3532050

Van der Feest, S. V. H., & Swingley, D. (in press). A crosslinguistic examination of toddlers' interpretation of vowel duration. *Infancy*.

Vihman, M., & Majorano, M. (2017). The role of geminates in infants' early word production and word-form recognition. *Journal of child language*, *44*, 158–184.

Warner, N., & Cutler, A. (2017). Stress effects in vowel perception as a function of language-specific vocabulary patterns. *Phonetica*, *74*, 81–106.

Werker, J. F., & Curtin, S. (2005). Primir: A developmental model of speech processing. *Language Learning and Development*, *1*, 197-234.

Yeung, H. H., & Werker, J. F. (2009). Learning words' sounds before learning how words sound: 9-month-old infants use distinct objects as cues to categorize speech information. *Cognition*, *113*, 234-243. doi: 10.1016/j.cognition.2009.08.010

Young, S. J., Evermann, G., Gales, M. J. F., Hain, T., Kershaw, D., Moore, G., ... Woodland, P. C. (2006). *The HTK book, version 3.4*. Cambridge, UK: Cambridge University Engineering Department.