Discriminability of sound contrasts in the face of speaker variation quantified

Christina Bergmann (chbergma@gmail.com)
Alejandrina Cristia (alecristia@gmail.com)
Emmanuel Dupoux (emmanuel.dupoux@gmail.com)
Laboratoire de Sciences Cognitives et Psycholinguistique, ENS/EHESS/CNRS; 29, rue d'Ulm
75005 Paris, France

Abstract

How does a naive language learner deal with speaker variation irrelevant to distinguishing word meanings? Experimental data is contradictory, and incompatible models have been proposed. Here, we examine basic assumptions regarding the acoustic signal the learner deals with: Is speaker variability a hurdle in discriminating sounds or can it easily be ignored? To this end, we summarize existing infant data. We then present machine-based discriminability scores of sound pairs obtained without any language knowledge. Our results show that speaker variability decreases sound contrast discriminability, and that some contrasts are affected more than others. However, chance performance is rare; most contrasts remain discriminable in the face of speaker variation. We take our results to mean that speaker variation is not a uniform hurdle to discriminating sound contrasts, and careful examination is necessary when planning and interpreting studies testing whether and to what extent infants (and adults) are sensitive to speaker differences.

Keywords: language acquisition; speech; acoustics; machine classification

Background

One important problem that infants are confronted with during language development is speaker variability. The same word might be implemented differently at the phonological level (as popularized in the song stating "I say tom[e]to, you say tom[ɑ]to"), acoustic targets can vary (Cristia, 2011), and speakers may differ in the shape of their vocal tract. As a result, the same word spoken by different people varies in its acoustic realization, even within the same accent. This problem is so potent that it requires drastic measures within speech recognition technologies to ensure proper handling of speaker variation: Typically, systems implement speaker normalization components which are trained with hundreds of speakers to attain a reasonable, but not perfect, performance on unknown speakers. Infants face the same problem: they have to recognize words as identical across speakers while maintaining the ability to distinguish different words (e.g., *take*$_1$ and *take*$_2$, versus *bake*$_1$). Alas, infants cannot use the same mechanisms as automatic speech recognition systems, while managing to learn their native language successfully and become adults who can generally deal with speaker variation (while being nonetheless affected by it). It is unclear which abilities infants bring to this task and which computational mechanisms they use.

Psycholinguistic models of speech and speaker processing range between two extremes. In the *abstractionist* view, infants are born with or rapidly acquire a system which separates speaker dependent and linguistically relevant phonetic information. For example, following a study with neonates, Dehaene-Lambertz and Pena (2001) state that "normalization is present from birth and is not the consequence of the establishment of phonetic prototypes following extensive exposure to speech." This means sounds and words are represented as abstract entities, invariant across speakers. The other end of the theoretical spectrum is the *episodic* view, where phonetic and speaker-specific information is not separated (Goldinger, 1998). In *hybrid* models, both invariant and speaker dependent formats of representation are combined to varying degrees (for a recent review see Kleinschmidt & Jaeger, 2015).

All models make implicit assumptions about the acoustic signal, which forms the basis for processing, representation, and learning. If a normalization and abstraction mechanism is innate, it seems necessary that, for a learner who starts from acoustic representations, linguistically contrastive information can be a priori separated from speaker-specific 'noise'. However, it is unlikely that such a separation is universally possible, since linguistic and speaker-specific variation has been stated to be intertwined and difficult to separate in several cases (Kleinschmidt & Jaeger, 2015). If, on the other hand, a normalization and abstraction mechanisms is learned, then previous to learning, abstractionist and exemplarist models are indistinguishable. In brief, before we can separate abstractionist from exemplarist models, it is important to establish which cases of speaker variability can be dealt with on the acoustic level and which require learning. As a preliminary of this, we need a method, which can, in a case by case fashion, quantify the amount of acoustic variation introduced by speaker variation.

Here, we propose a systematic measurement of the discriminability of phonetic categories of English (e.g., /a/ vs /o/, /a/, vs /i/, etc) as a function of speaker variability. We use a computerized ABX discrimibility score, whereby the acoustic distance between tokens that belong to the same category is compared to the distance between tokens from different categories. We systematically vary whether all tokens are spoken by the same or different speakers to quantify the impact of speaker variation[1].

In this paper, we first review some of the findings on the effect of speaker variability in infants' discrimination abilities. We then compute the acoustic discriminability scores of English vowels and consonants with a focus on the contrasts that were tested empirically.

[1]This method is not a model of infants or adults performing a particular discrimination paradigm. Instead, it is intended to evaluate how well the phonetic categories of a language are intrinsically separated in an acoustic/auditory representation.

Table 1: Studies on infants' sound discrimination abilities in the face of speaker variation. *Age* is reported in months; *Support* indicates how the results were interpreted; *Discrimination abilities* are divided into tasks where the speaker does not change (*within*) and where speaker variation was present (*across*). *Differences* are split into numerical ("$>$" signifies higher performance within-speaker, $\neq$ means there is a difference but a direction cannot be established) and statistical difference, based on tests of the interaction (sound type and task; significant: *, non-significant: ^{ns}.)

Contrast	Age	Reference	Support for	Discrimination results		
				Within	Across	Difference
/p/-/t/	0	Dehaene-Lambertz and Pena (2001)	Abstract	Yes	No	$>^{ns}$
/b/-/d/	2	Jusczyk, Pisoni, and Mullennix (1992)	Episodic	Yes	Yes/No	$>$
/a/-/i/	2,3,6	Marean, Werner, and Kuhl (1992)	Abstract	Yes	Yes	NA
/a/-/i/	5	Polka, Masapollo, and Ménard (2014)	Hybrid	NA	Yes	NA
/a/-/i/	6	Kuhl (1979)	Abstract	Yes	Yes	NA
/a/-/ɔ/	6	Kuhl (1983)	Abstract	Yes	Yes	NA
/b/-/p/ (/n/-/ŋ/)	7.5	Clough (2015)	Abstract	Yes	Yes	$\neq^{ns}$
/ɛ/-/ɪ/	12	Escudero, Bonn, Aslin, and Mulak (2015)	Episodic	Yes	NA	NA

Infants' Sound Discrimination Abilities in the Face of Speaker Variability

Infants' ability to deal with speaker variation has been investigated using a range of tasks, including word segmentation (Houston & Jusczyk, 2000), word learning (Rost & McMurray, 2009), learning of phonotactic rules (Seidl, Onishi, & Cristia, 2014), and sound discrimination. The last is most relevant for the present study, an overview can be found in Table 1. For reasons of space, we present in detail a representative selection of this line of work.

The logic of infant discrimination studies is as follows: infants first hear sequences of isolated syllables serving as background, followed by deviating stimuli. If a significant difference arises between (new tokens of) background and deviating stimuli, this is taken as evidence that infants can discriminate the two stimulus classes. Most studies analyze overt behavior, such as looking to an unrelated visual display, as indicator of infants' processing. We summarize each study in order of infant age. While infants' general discrimination ability matures over the course of the covered age range (Tsuji & Cristia, 2014), we cannot discern a clear developmental trend in the existing experimental literature pertaining to their ability to deal with speaker variation.

Dehaene-Lambertz and Pena (2001) measured electrophysiological responses to a deviating syllable compared to a background syllable, which was either spoken by one or multiple speakers, in a group of sleeping neonates. The authors took a main effect of condition (deviant versus background) and the lack of an interaction with speaker (within versus across) as evidence of infants' ability to ignore irrelevant information by normalizing over speakers. However, post-hoc tests within the infant group who heard multiple speakers revealed that discrimination in the "across" condition was not significant. Further, the analyzed electrodes (and thus presumably underlying regions) differed for the within- versus across-speaker condition. Despite these differences, the study by Dehaene-Lambertz and Pena (2001) is taken to support an abstractionist viewpoint, whereas the findings by Jusczyk et al. (1992), frequently cited to support an episodic view, are actually comparable, as we show next.

Jusczyk et al. (1992) tested two-month-olds in a high-amplitude sucking habituation-dishabituation paradigm. In this paradigm, infants are first exposed to the background stimulus contingently with their high-amplitude sucking. In this phase, sucking always results in repetition of the same stimulus list, so, usually, the sucking rate wanes over time. The measure of interest is infants' sucking rate when hearing test tokens, which are either the same as before for controls or new tokens for the experimental group. A difference in sucking rate in the latter compared to the former group indicates that infants dishabituated due to detecting a difference in the stimuli. In Jusczyk et al. (1992), infants in the single-talker condition were habituated with the word "bug" spoken by one voice, and tested with "dug" in the same voice; or they first heard "bug" spoken by 6 different talkers (3 male, 3 female), and were tested with "dug" in the same 6 voices. Infants detected changes regardless of condition. In a follow-up experiment, the authors introduced a 2-minute delay between habituation and test. In this setting, only infants in the single-talker condition detected the phoneme change, whereas infants in the multi-talker condition failed. This failure was replicated when the 6 habituation talkers were all drawn from one gender. Through additional experiments, Jusczyk and colleagues demonstrated that infants detected a phonemic change even in the face of within-talker variation (i.e., using multiple different tokens from the same talker), leading them to conclude that talker variability can be disruptive. However, considering only the first experiment, infants succeeded in the multi-talker condition, and additionally there were no direct statistical comparisons of within- versus across talker-conditions.

Kuhl (1979) trained six-month-olds over several days to react to a change in stimuli with a head-turn, by initially presenting a salient deviating stimulus and letting a toy appear on one side at the same time. Thus, turning the head to the side indicates infants' detection of phonemic changes, which the authors then used to test infants with vowels in the face

of variation of vowel pitch, talker identity, and both, in that order. Only when infants succeeded in detecting changes at a given performance level and within a set number of trials were they exposed to subsequent stages with more variability. All tested infants completed the experiment, including the most variable stage, which included both talker and pitch variation. This finding is frequently taken as evidence for abstraction. A follow-up study by Kuhl (1983) extended these results to different vowels, and Marean et al. (1992) obtained similar results with younger babies. Due to the study design it is not possible to compare within- and across-talker performance directly.

The most recent study on the topic measured dishabituation to a psycho-acoustically salient contrast versus a more subtle one (/p/-/b/ versus /n/-/ŋ/), either in the presence of, or without, talker variability (Clough, 2015). Infants showed a significant difference between habituation and novel stimuli, although in opposite directions across the two phonetic contrasts. As is typical in this paradigm, infants looked longer when hearing the novel stimulus in the /p/-/b/ condition, for the more subtle contrast they looked longer when hearing the habituation stimulus. The second pattern, a so-called familiarity preference, is difficult to interpret, but might indicate greater processing demands (Hunter & Ames, 1988). A direct test of discrimination in within- and across-speaker conditions revealed no difference in performance.

In summary, the evidence of infants' ability to discriminate sound contrasts in the face of speaker variation and change is scarce and has been used to support incompatible standpoints: Either infants abstract over speaker-specific characteristics or not. It becomes crucial to establish how speaker differences impact the acoustic signal to determine how infants will be affected by them in an episodic framework or to better understand which problem they have to solve to achieve abstraction. In this paper, we assess the impact of speaker variation on the acoustic signal, both overall and focusing on those contrasts previously tested on infants.

Experiment

We test whether speaker variation impacts the discriminability of speech sounds in the absence of lexical or phonological knowledge (Schatz et al., 2013, 2014; Martin et al., 2015). Importantly, we do not implement automatized speaker normalization procedures and provide no language-specific information on any level.

Supplementary information, including lists of experimental tokens, raw data, figures, scripts, and results is available at `https://osf.io/mvnjy/`.

Speech Material

The Articulation Index Corpus (Schatz et al., 2015) contains noise-free recordings of all American English phones in diphone pairs (e.g., /ba/, /la/, ...), pronounced by 20 speakers (8 women). The contrast /n/-/ŋ/ cannot be tested, as /ŋ/ does not naturally appear at the onset of syllables in English (Clough, 2015); it thus was not included in our study. Otherwise, our corpus choice is similar to the stimuli typically used in infant studies, as both are recorded under noise-free conditions and based on prompted and isolated instead of spontaneous, connected productions.

Acoustic Representations

We use two common acoustic representations of the speech signal: Mel filterbanks and Mel Frequency Cepstral Coefficients (MFCCs). These representations encode the spectral properties of thin slices of the speech signal, and have been argued to be similar to the first stages of human auditory processing (Gold & Morgan, 2000). In the subsequent reports we focus on Mel filterbanks since the results do not differ substantially for MFCCs.

Discriminability Scores

To quantify how discriminable a contrast is within and across speakers, we compute an ABX discriminability score (Schatz et al., 2013, 2014). This score quantifies how often a test sound X, e.g., ba_1, is correctly identified as member of the same category as A (ba_2) and not B (da_1).

The machine-based discriminability score was used previously by Martin et al. (2015) to assess whether mothers speak more or less clearly to their infants compared to adults, systematically testing the so-called hyperarticulation hypothesis. The scores can be computed automatically over large data sets and allow us to quantify discriminability in a non-parametric way over various sound contrasts using a single metric.

We use the ABXpy package (Schatz, Thiolliere, Synnaeve, & Dupoux, 2016) and follow essentially the same procedure as Martin et al. (2015).[2] To calculate the score the following steps are taken: (1) Encode all available tokens in terms of their acoustic properties (here: Mel filterbanks or MFCCs); (2) Align pairs (either from the same category, e.g., ba_1 and ba_2 or crossing categories, e.g., ba_1 and da_1) via dynamic time warping (using the implementation by Synnaeve, 2016) and compute their distance; here, we use the mean of the inverse cosine deviation from the diagonal (signifying identity); (3) Identify all possible combinations of tokens which can function as A, B, and X, respectively; (4) Compare distances of $A-X$ and $B-X$, where A, X versus B represent a given contrast (e.g., /b/-/d/), and count a success when X was attributed to the correct category because the pairwise distance was smaller for the pair from the same category; (5) Count the successfully categorized triplets and divide by the total number of triplets to get the normalized final score, which ranges between 0 and 1; chance level is .5.

We compute ABX scores in two conditions: Either all three tokens in a triplet stem from the same speaker, or the test token X is sampled from a different speaker than both A and B. To directly tap into the difference of within- versus across-speaker discriminability, we compare absolute

[2] We use mean and not sum difference, which leads to overall more robust performance. Details can be found at `https://osf.io/mvnjy/`.

scores for the two conditions. We further compute a difference score by subtracting across-speaker scores from within-speaker scores. The absolute scores take into account the ease of discriminating each contrast while the difference scores show how much each contrast is affected by speaker variability. To obtain distributional information, we randomly subsample speakers for each contrast 1000 times.

Results

The mean discriminability scores, both within- and across-speaker, for every available contrast in the corpus (a total of 358) are shown in Figure 1. Figure 2 shows the difference scores for the subset of contrasts that we selected on the basis of infant research on this topic (results using both male and female speakers). Table 2 contains the discriminability scores that form the basis for the depicted difference scores. Together, the two figures and the table illustrate the following:

1. There is always an advantage (a higher score,) for within- compared to between- speaker discriminability. The mean difference score for all contrasts is .115 (95% CI: [.087, .142]; score range: .003-.278).
2. A ceiling effect leads to the smallest observed differences (points closest to the diagonal), as exemplified by /a/-/i/.
3. Only very few contrasts drop to chance level (see Figure 1).
4. The speaker variability effect seems to be lower when only including women (see the subsets in Figure 1), but the difference between these groups is reliable (mean across-speaker score .827; mean across-women score .853; mean within-speaker score for both speaker sets .941; 95% CI for paired difference scores on all contrasts: [-.020, .083]).

These general observations hold when changing the acoustic representation to MFCCs, which leads to overall lower absolute scores, leaving the difference score largely unaffected (mean scores within-speaker .876 (vs .941); across-speaker .760 (vs .827); difference score .116 (vs .115)).

Table 2: Summary of results for contrasts tested in infant studies, presented as means and 95% CIs.

	Within-Score	Across-Score	Difference
a-ɔ	.734 [.682, .788]	.553 [.523, .593]	.181 [.143, .218]
p-t	.858 [.828, .888]	.686 [.658, .717]	.173 [.140, .207]
b-p	.813 [.777, .848]	.642 [.623, .664]	.171 [.142, .202]
ɛ-ɪ	.871 [.837, .906]	.710 [.675, 746]	.161 [.125, .194]
b-d	.813 [.781, .848]	.654 [.629, .682]	.159 [.124, .194]
a-i	.999 [.998, 1]	.990 [.984, .996]	.009 [.004, .014]

Discussion

We set out to quantify the impact of speaker variability on sound discriminability. Our results show that there is an overall negative impact of speaker variation on sound discriminability. Considering psycholinguistic models, our findings have implications for both abstractionist and episodic models. For the former, this study specify the set of contrasts that are especially non-invariant, and for which a putative innate

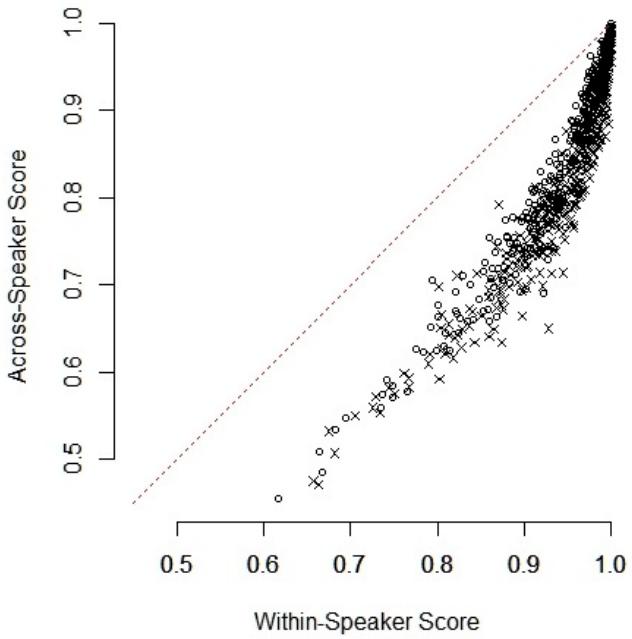

Figure 1: Mean discriminability scores for all possible contrasts in the corpus, showing within- versus across-speaker scores. The **x** mark comparisons across all available speakers, the **o** the results for a subset containing only female speakers. The diagonal (no difference) is indicated by a dotted line.

normalization module has to be specified. For episodic and hybrid models, we make quantitative predictions on the deterioration expected when speaker variability is introduced.

How do our data compare to studies testing whether infants are able to normalize across speakers? The adverse effect of speaker variation we observe is not catastrophic, and can even be seen as modest, as the vast majority of sound contrasts remain discriminable acoustically (better than chance) in the face of talker variation. Furthermore, the impact varies greatly across contrasts, and we observe a ceiling effect. It is important to reflect that this could potentially undermine the possibility of empirically testing episodic models, as follows. A frequent prediction, at least in infant literature, seems to be that episodic models should yield a complete absence of discrimination abilities as soon as talker variation is introduced (Dehaene-Lambertz & Pena, 2001). However, our results show an overall adverse effect of speaker variation. It should be noted that infants' abilities might not be reflected in their performance in the indirect measures described in detail above, although the two are frequently equated (e.g., Apfelbaum & McMurray, 2011; Bergmann, ten Bosch, Fikkert, & Boves, 2013).

In addition, as the size of the adverse effect depends strongly on the chosen contrast, it is possible to observe no difference at all. In fact, the first study supporting early normalization in infants (Kuhl, 1979) chose /a/-/i/, a contrast that is, put simply, always easy. The follow-up study with

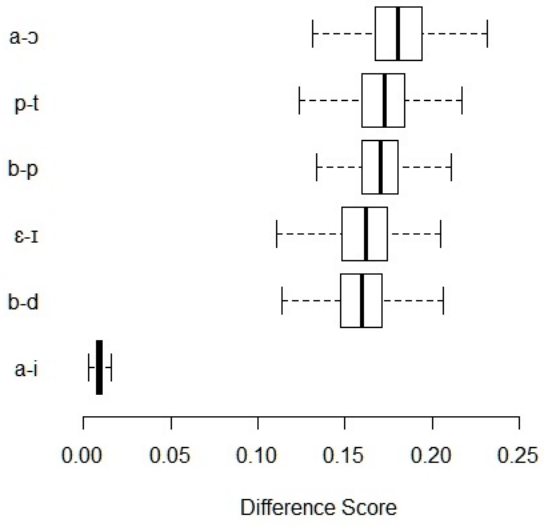

Figure 2: Differences of mean discriminability scores within and across speaker, for the six contrasts tested on infants. A positive difference score indicates that the performance is better within than across speaker.

/a/-/ɔ/ (Kuhl, 1983), the only contrast in the infant literature that approaches (but is above) chance level, trained infants over several days in increasingly variable and difficult conditions. This training might explain continued success with added speaker variability, an advantage our evaluation did not implement.

Clough (2015) tested /b/-/p/ and found a numerical difference of within- versus across-speaker performance, which can be taken as consistent with our finding that this contrast is strongly affected by speaker variation. There was, however, no statistically significant effect, which the author herself attributes to insufficient statistical power. We believe that the same low power may also explain the lack of a significant interaction in the study by Dehaene-Lambertz and Pena (2001). Although this is difficult to determine (Gelman & Stern, 2006), we repeat that the discrimination response was significant in the within-speaker condition, and not in the across-speaker condition.

The observation that non-significant results might be grounded in a lack of power, especially when testing interactions, makes it necessary to discuss the sensitivity of infant measures. As a recent meta-analysis showed, there is substantial variability (I^2=76.87%) in infants' native vowel discrimination performance and only a medium-sized overall effect (Cohen's d=.6, SE=.05; data retrieved from `metalab.stanford.edu`, consulting the dataset of Tsuji & Cristia, 2014). This leads to frequently underpowered studies on the topic, a problem only worsened when testing an interaction with a small effect (within- or across-speaker discrimination). In other words, since speaker variation has a consistent, but moderate effect (we cannot expect a complete breakdown of discrimination abilities across speakers), the impact on infant behavior might be difficult to measure and/or require a(n unfeasibly) high number of participants.

We suggest, nonetheless, to test infants' developing discrimination abilities in the face of speaker variability by comparing two contrasts that are matched on within-speaker discriminability, but that differ maximally in the across-speaker task, as measured by the difference score. The difference between these two scores would provide a measure of how much infants are sensitive to specific difficulties introduced by speaker change, and would therefore provide us with the required quantitative evaluation of episodic versus abstractionist models.

Extrapolating from our data to language acquisition outside the laboratory, our results suggest that infants' input becomes more difficult to learn from when talker variability is present. This holds from both theoretical viewpoints, as long as abstraction is not innate, and has to be (at least partially) acquired. Given that infants tune into their native language based on the acoustic signal, being exposed to higher input variability in the form of more speakers leads to a more difficult learning problem.

The present results were based on a corpus that was both maximally exhaustive and recorded under ideal conditions, much like the stimuli infants are typically confronted with in the lab. Follow-up experiments will address how our results generalize to corpora of infant-directed speech (IDS), which are often not available in sufficiently high quality. It remains an open question, and one orthogonal to the issue of hypo- or hyperarticulation in IDS (Martin et al., 2015), whether the acoustic markers, such as higher, but also more variable, pitch, emphasize or lessen speaker differences. We may thus in an extension of the present work quantify the learning problem infants face in real life when confronted with multiple talkers on a daily basis.

The present work has implications for adult models of speech processing, as it maps out the extent of the problem of introducing speaker variation, a task adults need to solve as well, albeit with more knowledge and experience. Here, too, predictions from abstractionist models have to be carefully examined. To better test whether or not listeners generate a speaker invariant representation it is not correct to expect chance performance when introducing multiple speakers; for some contrasts we predict no adverse effects for a system without any normalization and language knowledge. Thus, paradigms sensitive to decreased performances and a contrast chosen to maximize the predicted difference can better tap into the question of how adults process speaker variance.

In sum, we have shown that speaker variation poses a hurdle for sound discrimination but does not necessarily lead to confusion, even for a naive learner. In quantifying the effect, we can derive more precise predictions when testing competing psycholinguistic models.

Acknowledgements

We thank T. Schatz, X.-N. Cao, R. Thiolliere, M. Bernard, and G. Synnaeve for their help. The present work was supported by the European Horizon 2020 programme (Marie Skłodowska-Curie grant No 660911), the European Research Council (E-2011-AdG 295810 BOOTPHON), the Agence Nationale pour la Recherche (ANR-2010-BLAN-1901-1 BOOTLANG, ANR-14-CE30-0003 MechELex, ANR-10-IDEX-0001-02 PSL*, ANR-10-LABX-0087 IEC), and the Fondation de France.

References

Apfelbaum, K. S., & McMurray, B. (2011). Using variability to guide dimensional weighting: Associative mechanisms in early word learning. *Cognitive Science*, *35*(6), 1105–1138.

Bergmann, C., ten Bosch, L., Fikkert, P., & Boves, L. (2013). A computational model to investigate assumptions in the headturn preference procedure. *Frontiers in Psychology*, *4*(676).

Clough, L. T. (2015). *Does variability impact infants' sound discrimination?* (Unpublished Thesis, accessed via `http://hdl.handle.net/10150/578957`)

Cristia, A. (2011). Fine-grained variation in caregivers' /s/ predicts their infants' /s/ category. *The Journal of the Acoustical Society of America*, *129*(5), 3271–3280.

Dehaene-Lambertz, G., & Pena, M. (2001). Electrophysiological evidence for automatic phonetic processing in neonates. *Neuroreport*, *12*(14), 3155–3158.

Escudero, P., Bonn, C. D., Aslin, R. N., & Mulak, K. E. (2015). Indexical and linguistic processing in infancy: Discrimination of speaker, accent and vowel differences. In *Proceedings of the International Congress of Phonetic Sciences.*

Gelman, A., & Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician*, *60*(4), 328–331.

Gold, B., & Morgan, N. (2000). Chapter 14: Ear Physiology. In *Speech and Audio Signal Processing: Processing and Perception of Speech and Music* (pp. 189–203). New York: J Wiley.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*(2), 251–279.

Houston, D. M., & Jusczyk, P. W. (2000). The role of talker-specific information in word segmentation by infants. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(5), 1570–1582.

Hunter, M. A., & Ames, E. W. (1988). A multifactor model of infant preferences for novel and familiar stimuli. *Advances in Infancy Research*, 69–95.

Jusczyk, P. W., Pisoni, D. B., & Mullennix, J. (1992). Some consequences of stimulus variability on speech processing by 2-month-old infants. *Cognition*, *43*(3), 253–291.

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.

Kuhl, P. K. (1979). Speech perception in early infancy: Perceptual constancy for spectrally dissimilar vowel categories. *The Journal of the Acoustical Society of America*, *66*(6), 1668–1679.

Kuhl, P. K. (1983). Perception of auditory equivalence classes for speech in early infancy. *Infant Behavior and Development*, *6*(2), 263–285.

Marean, G. C., Werner, L. A., & Kuhl, P. K. (1992). Vowel categorization by very young infants. *Developmental Psychology*, *28*(3), 396–405.

Martin, A., Schatz, T., Versteegh, M., Miyazawa, K., Mazuka, R., Dupoux, E., & Cristia, A. (2015). Mothers speak less clearly to infants than to adults: A comprehensive test of the hyperarticulation hypothesis. *Psychological Science*, *26*(3), 341–347.

Polka, L., Masapollo, M., & Ménard, L. (2014). Whos talking now? infants perception of vowels with infant vocal properties. *Psychological Science*, *25*(7), 1448–1456.

Rost, G. C., & McMurray, B. (2009). Speaker variability augments phonological processing in early word learning. *Developmental Science*, *12*(2), 339–349.

Schatz, T., Cao, X.-N., Kolesnikova, A., Bergvelt, T., Wright, J., & Dupoux, E. (2015). *Articulation Index lscp* (Speech Corpus No. LDC2015S12). Linguistic Data Consortium.

Schatz, T., Peddinti, V., Bach, F., Jansen, A., Hermansky, H., & Dupoux, E. (2013). Evaluating speech features with the minimal-pair ABX task: Analysis of the classical MFC/PLP pipeline. In *14th Annual Conference of the International Speech Communication Association.*

Schatz, T., Peddinti, V., Cao, X.-N., Bach, F., Hermansky, H., & Dupoux, E. (2014). Evaluating speech features with the minimal-pair ABX task (II): Resistance to noise. In *15th Annual Conference of the International Speech Communication Association.*

Schatz, T., Thiolliere, R., Synnaeve, G., & Dupoux, E. (2016). *ABXpy v0.2.* Retrieved from `http://dx.doi.org/10.5281/zenodo.45268`

Seidl, A., Onishi, K. H., & Cristia, A. (2014). Talker variation aids young infants phonotactic learning. *Language Learning and Development*, *10*(4), 297–307.

Synnaeve, G. (2016). *DTW_Cython.* Retrieved from `http://dx.doi.org/10.5281/zenodo.45257`

Tsuji, S., & Cristia, A. (2014). Perceptual attunement in vowels: A meta-analysis. *Developmental Psychobiology*, *56*(2), 179–191.