Nasal consonant discrimination in infant- and adult-directed speech

Bogdan Ludusan[1,2]*, Annett Jorschick*[1]*, Reiko Mazuka*[2]

[1] Faculty of Linguistics and Literary Studies, Bielefeld University, Germany
[2] Laboratory for Language Development, RIKEN Center for Brain Science, Japan
{bogdan.ludusan, annett.jorschick}@uni-bielefeld.de, reiko.mazuka@riken.jp

Abstract

Infant-directed speech (IDS) is thought to play a facilitating role in language acquisition, by simplifying the input infants receive. In particular, the hypothesis that the acoustic level is enhanced to make the input more clear for infants, has been extensively studied in the case of vowels, but less so in the case of consonants. An investigation into how nasal consonants can be discriminated in infant- compared to adult-directed speech (ADS) was performed, on a corpus of Japanese mother-infant spontaneous conversations, by examining all bilabial and alveolar nasals occurring in intervocalic position. The Pearson correlation between corresponding spectrum slices of nasal consonants, in identical vowel contexts, was employed as similarity measure and a statistical model was fit using this information. It revealed a decrease in similarity between the nasal classes, in IDS compared to ADS, although the effect was not statistically significant. We confirmed these results, using an unsupervised machine learning algorithm to discriminate between the two nasal classes, obtaining similar classification performance in IDS and ADS. We discuss our findings in the context of the current literature on infant-directed speech.

Index Terms: nasals, acoustic characteristics, infant-directed speech, adult-directed speech

1. Introduction

Infant-directed speech (IDS) is a speech register used in many cultures by adults to address infants. It stands apart from the speech register that adults use to speak with each other (adult-directed speech; ADS) on several aspects, including the use of shorter phrases, a more exaggerated intonation and a simpler vocabulary. It is believed that IDS might play different roles, including holding attention, expressing affect or facilitating language acquisition [1].

One of the research directions pursued in the field investigated how IDS and ADS differ from an acoustic point of view and how this may affect the language learning process (for an overview see [2]). In the case of vowels, Kuhl and colleagues [3] have shown that a vowel expansion phenomenon can be observed, by which the means of the vowel categories are further apart from each other in the acoustic space represented by the first two formants, in IDS, compared to ADS. These findings have been confirmed in several languages (e.g. [4, 5]), although more recent studies have also found more variable vowel categories in IDS than in ADS [6, 7, 8].

While IDS-ADS differences for vowel categories have been extensively studied, consonants have received less attention. Among consonantal classes, plosives (e.g. [9, 10]) and fricatives (e.g. [11, 12]) have been investigated. The vast majority of studies looking at differences in plosive production between the two registers focused on one dimension - voice onset time - with the results reporting both shorter [9] and longer voice onset times [10] in IDS, compared to ADS. In the case of fricatives, a longer duration was observed for /s/ in IDS than in ADS [11] as well as a higher spectral centroid for the same fricative, in IDS [12]. A large-scale computational study examining the discrimination of phonemes in ADS and IDS was performed by Martin and colleagues [8]. They employed a computational minimal-pair ABX task, comparing syllabic minimal pairs, by means of spectral representations used in speech technology. Their findings showed a small, but significant, advantage for ADS phoneme discrimination.

In this study, we investigated whether there are acoustic differences between registers in the realization of another class of consonants, nasals. Furthermore, we did not limit our analysis to establishing the existence of acoustic differences, but we also tested whether these changes have an effect on the discrimination of nasal categories, by means of a machine learning experiment. In particular, we looked at the discrimination of /n/ and /m/ in Japanese ADS-IDS. Similarly to the approach taken in [12], we focused here on perceptually relevant dimensions for discriminating the two nasal categories. According to the previous literature [13], formant transitions and nasal murmur are employed for identifying the place of articulation of nasal consonants, with equally important roles. Consequently, the amplitude of the lower part of the frequency spectrum (containing the previously mentioned characteristics) was considered for the computation of a similarity measure between the two nasal categories, as well as to train an Expectation Maximization based unsupervised classifier to discriminate between them. If any phonetic enhancement is present in IDS, a lower similarity between categories and a higher classification performance should be obtained in this register.

The paper is structured as follows: the following section introduces the dataset, while Section 3 presents the methods employed in the investigation. The results of the statistical analysis on the derived similarity values, as well as those of the classification experiments are illustrated in Section 4. The paper will conclude with a discussion of these findings and how they contribute to the current state of the field.

2. Materials

The investigation was conducted on the RIKEN Mother-Infant Conversation corpus [14]. It consists of spontaneous interactions between 22 Japanese mothers and their 18-24 month-old infants, while reading a book or playing with toys. The same mothers were, subsequently, recorded discussing child-rearing topics with an interviewer. The entire corpus contains more than 11 hours of infant-directed speech and around 3 hours of adult-directed speech. It was manually transcribed and annotated at both segmental and prosodic levels.

We considered in this study all the bilabial (/m/) and alveolar (/n/) nasals in the two subparts of the corpus. We did not

Table 1: *Statistics regarding the analyzed nasal instances.*

Context	ADS		IDS	
	n	m	n	m
a_a	84	14	164	27
a_o	70	15	59	18
e_a	97	32	79	27
e_o	35	22	5	51
i_a	104	69	337	43
i_o	62	19	62	21
o_e	36	5	19	23
o_o	38	22	33	46
u_a	43	12	27	24
u_o	101	10	90	9
Total	670	220	875	289

include the Japanese moraic nasal (/N/) in our investigation, as its phonetic realization varies with the context it is produced in. We, then, restricted the analysis to the sequences composed of nasals preceded and followed by a vowel (we did not differentiate between short and long vowels here), not found at the beginning or at the end of a prosodic phrase. For each nasal category and each vowel context we calculated the frequency of occurrence of the given vowel-nasal-vowel sequence in the two speech registers. We kept only those sequences which had at least five occurrences (instances) in both registers. Statistics on the included contexts are presented in Table 1.

3. Methods

Each nasal instance of the analyzed vowel-nasal-vowel contexts had its spectrum extracted at nine, equally distant, time instants within the consonants, starting and ending at its boundaries (see Figure 1 for an illustration of the sampled times). The Praat software [15] was used for extracting the spectral slices at the considered time instants and the amplitude of the spectrum was employed in all subsequent analysis. Since place of articulation for nasal consonants is discriminated based on the trajectory of the lower formants and the spectral characteristics of the nasal murmur [13], only the frequencies which contain this information (between 0 and 2000 Hz) were examined. Thus, each nasal was represented by nine spectral slices (feature vectors). Each vector contained 33 values, corresponding to the amplitude of the spectrum in the $[0, 2000]$ Hz interval, at the sampled time instants.

As a measure of similarity between nasal classes we employed the Pearson correlation. Pearson correlation is related

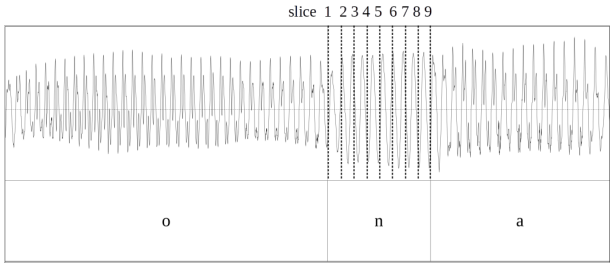

Figure 1: *Illustration of the time slices for which the amplitude of the speech spectrum is computed.*

to the cosine distance, used in many current speech-based systems for computing similarity scores. The former measure has the advantage of being more stable to variation present in the data, being equivalent to the latter when the input data is mean-normalized. We computed the correlation between the amplitudes of the corresponding spectrum slices of all the /n/s in a register and context and all the /m/s in the same register and context (e.g. slice 1 of an ADS /n/ produced in /a_a/ context with slice 1 of an ADS /m/ produced in /a_a/ context, etc). We then aggregated across each context, in each register, and we considered the average correlation within each context as the dependent variable in the ensuing statistical analysis. Besides the /n/-/m/ similarity we also computed, as control, the /n/-/n/ and /m/-/m/ similarities, respectively, within register and context, as well. We expect the within-class similarity to be higher than between-class similarity. Also, any spectral enhancement in IDS should result in a lower between-class similarity in IDS than in ADS.

Two complementary analyses were performed: First, a statistical model of our similarity measure was built, in order to ascertain the presence of any acoustic enhancement in the spectrum of the two nasal categories, across the two registers. Since we do not know whether the presumed spectral differences would have an effect on the learning process, we performed, in a second step, machine learning experiments using an unsupervised learning paradigm to test the /n/-/m/ discrimination.

3.1. Statistical modelling

We employed linear mixed-effects models [16] to estimate whether the Pearson correlation within and between nasal classes (nasal conditions: /n/-/n/, /m/-/m/, /n/-/m/) differed between speech registers (ADS and IDS). Crossed random effects were modeled for vocalic context and spectrum slice with varying slopes for speech register and nasal condition. The random effects structure was then reduced following the procedure outlined by Bates and colleagues [17] to avoid over-fitting and to enhance power [18]: Starting with the maximal model, we used likelihood ratio tests to compare models with a more complex random effects structure to nested models with a simpler random effects structure. Terms were kept in the model if their removal significantly reduced the fit of the model.

The two predictors (speech register and nasal condition) and their interaction were modeled using successive difference contrasts [19]. Likelihood ratio tests between hierarchically nested models with a reduced fixed effects structure served to estimate χ^2 and p-values for main effects and the interaction. The differences between speech registers within the nasal condition were tested using paired t-tests, since the correlation coefficients within nasal condition (dependent variable) were normally distributed.

3.2. Machine learning

Besides the statistical analysis, we also performed machine learning experiments, to discriminate between the nasal classes, in the two speech registers. The experiments employed an unsupervised learning paradigm, meaning that it did not make use of class information at training time. The chosen machine learning algorithm assumes that each input feature is independent from one another and follows a different Gaussian distribution given the category. It tries to fit N Gaussian distributions to the training data, by means of the Expectation Maximization algorithm, where N is the expected number of categories. At test time, the algorithm returns, for each instance, the probability of be-

longing to one of the obtained clusters, with the instance being assigned to the class having the highest probability. The implementation employed here is the one given by the python sklearn package [20].

The system was given the number of nasal classes, two, and it was run for a maximum of 100 iterations, with a convergence threshold of $1E{-}3$ and using full covariance matrices. We ran it separately, for each slice, within each context and each register, and then we aggregated the results across context and register, obtaining one discrimination score per context (10 scores per register). This process was run for 1000 times and the average performance across the runs was reported. The goodness of the clustering process was evaluated using the F1-score, defined as the harmonic mean between precision (the proportion of correct assignments among the total number of instances assigned to a class) and recall (the proportion of instances of a class found, among the total number of instances of that class). A higher F1-score represents a better classification performance. We used here the micro-averaged F1-score, such that the results were not biased by the unbalanced distribution of classes in our dataset.

4. Results

We first demonstrate the soundness of the proposed approach, by validating it on the ADS data. For our similarity measure to perform as expected, it would have to return significantly higher correlations in the /n/-/n/ and /m/-/m/ cases than in the /n/-/m/ case. We tested this by means of paired two-tailed t-tests between the average /n/-/n/ correlation and the average /n/-/m/ correlation, as well as between /m/-/m/ and /n/-/m/. Both tests were highly significant ($p < .001$ in both cases), suggesting that our measure successfully captures the desired characteristic.

The results of the linear mixed effects model (see Table 2 for the estimates of the fixed effects) revealed a significant interaction between speech register and nasal condition ($\chi^2(2) = 32.46, p < .001$) and a main effect for nasal condition ($\chi^2(2) = 20.35, p < .001$) but none for speech register ($\chi^2(1) < 1$). The difference between speech registers became significant in the /n/-/n/ condition: correlations in IDS were slightly higher ($mean = .84$) than in ADS ($mean = .83$); $t(89) = -2.55, p < .05$. The difference between speech registers was significant also in the /m/-/m/ condition, though in the opposite direction: correlations in IDS were lower ($mean = .83$) than in ADS ($mean = 0.85$); $t(89) = 4.00, p < .001$. Critically, we found no difference between speech registers in the /n/-/m/ condition (IDS: $mean = .81$, ADS: $mean = .81$, $t(89) = 1.00, p = .32$). Overall, the correlation coefficients were significantly lower in the /n/-/m/ condition than in the /n/-/n/ condition ($\beta = -.023, t = -3.90, p < .001$) or the /m/-/m/ condition ($\beta = -.029, t = -4.13, p < .001$). Figure 2 provides a graphical description of the obtained results.

The findings of the machine learning experiments are illustrated in Figure 3. A slightly better discrimination between the two nasal classes was obtained in IDS, compared to ADS, but a paired two-tailed t-test showed that the difference is not significant ($t = -2.113, df = 9, p = 0.064$).

Table 2: *Fixed effects of the final mixed model*

Fixed effect	β	**SE**	**t-value**
Intercept	0.827	0.010	85.51
Speech register	−0.004	0.008	−0.54
n-m / n-n	−0.023	0.006	−3.99
m-m / n-m	0.029	0.007	4.13
Speech register : n-m / n-n	−0.013	0.005	−2.80
Speech register : m-m / n-m	−0.014	0.005	−3.00

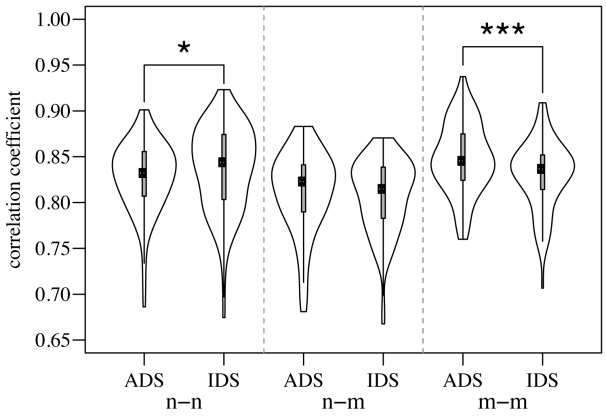

Figure 2: *Pearson correlation coefficients, characterizing the within- and between-class similarity, broken down by nasal condition and speech register. Asterisks above the plots represent the degree of significance of the difference between speech registers (* $p < .05$; *** $p < .001$).*

5. Discussion and conclusions

We have presented here a study exploring the acoustic differences between ADS and IDS in the realization of nasal categories /n/ and /m/, in Japanese, and how they might affect the discrimination of these categories in the aforementioned registers. We employed two approaches: First, we computed the similarity between categories based on features characterizing perceptually relevant dimensions for discriminating the two nasal categories (the amplitude of the lower part of the spectrum, including the nasal murmur as well as the F1 and F2 transitions). Then, we attempted to classify, in unsupervised fashion, the two categories based on the extracted features.

A lower similarity between /n/ and /m/ was observed in IDS compared to ADS, along with a higher classification per-

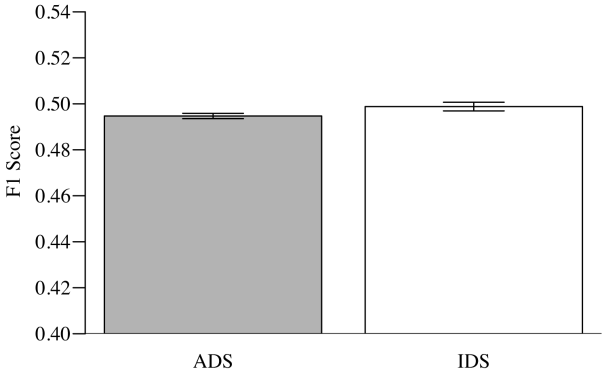

Figure 3: *F1-score obtained for the classification of /n/ and /m/, employing an unsupervised learning paradigm.*

formance in the former register, although none of the differences were statistically significant. The results obtained point towards the same conclusion - there seems to be no phonetic enhancement that could help discriminate better the two nasal categories in IDS than in ADS. Furthermore, the validity of the proposed metric for computing the similarity between categories is supported by the fact that the same results were obtained with two different methods, as well as by comparing the between-categories with the within-category conditions, in which a higher similarity was obtained for the latter conditions.

Our findings, employing two different methodologies and perceptually relevant features, are in line with the results of a previous study [8], on the same dataset. Even with a more controlled analysis, no significant difference was observed in the discrimination of /n/-/m/, between the two registers.

Some interesting effects were observed between registers: an increase in similarity for /n/ (suggesting some sort of enhancement) and a decrease in similarity for /m/ (implying the inverse process), in IDS compared to ADS (both differences significant). The opposing nature of the two resulted in a small overall effect for /n/-/m/ similarity. In order to try to understand the differing directions for the within-category conditions, we looked at the distribution of the two categories in our data. We observed that the proportion of word-medial /m/s (out of the total number of instances) increased significantly in IDS, while the proportion of word-medial /n/s was relatively stable in the two registers. It might be that word edges in IDS are phonetically enhanced (similarly to the acoustic changes occurring in the case of initial strengthening in ADS [21]), and that the resulting vocal effort has an opposite effect on word-medial syllables. Comparable acoustic enhancement phenomena have been observed in IDS, e.g. for vowels produced in focus syllables [22]. Further analyses, run on datasets containing more balanced distributions of the nasal categories with respect to their position inside the word, would be needed in order to untangle these effects.

The current study gives further insights into the complex issue of differences between ADS and IDS and how they affect the process of early language acquisition. Previous studies on the same corpus have shown an adverse effect of IDS when looking at phoneme discrimination [8], as well as a less clear vowel class separation due to an increased intra-class variability [23]. While at segmental level no advantage of IDS was observed, when looking at other linguistic levels, for instance prosodic boundary detection [24] or lexical segmentation [25], IDS does have an advantage. Nevertheless, when combining both acoustic and lexical information for word form learning, an overall detrimental effect for IDS was obtained [26].

For a more holistic view of the learning process, it is important to consider the impact of any acoustic differences between registers on other linguistic processes, such as word segmentation [27], as well as the role that other (higher-level) linguistic knowledge, like lexical information, might play on the category learning itself [28]. Furthermore, any potential adverse influence of IDS, due to higher acoustic variability, might be offset by other considerations: Since infants seem to prefer listening to IDS over ADS [29], a more complete model would have to include also its possible effects on learning, through social interaction and motivation [30].

6. Acknowledgements

The research reported in this paper was partly funded by JSPS Grant-in-Aid for Scientific Research S (16H06319) and MEXT Grant-in-Aid on Innovative Areas #4903 (Co-creative Language Evolution), 17H06382 to RM. BL was also supported by the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement no. 799022.

7. References

[1] A. Fernald, T. Taeschner, J. Dunn, M. Papousek, B. de Boysson-Bardies, and I. Fukui, "A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants," *Journal of Child Language*, vol. 16, no. 3, pp. 477–501, 1989.

[2] Y. Wang, D. M. Houston, and A. Seidl, "Acoustic properties of infant-directed speech," in *The Oxford Handbook of Voice Perception*. Oxford University Press, 2019, pp. 93–116.

[3] P. Kuhl, J. Andruski, I. Chistovich, L. Chistovich, E. Kozhevnikova, V. Ryskina, E. Stolyarova, U. Sundberg, and F. Lacerda, "Cross-language analysis of phonetic units in language addressed to infants," *Science*, vol. 277, no. 5326, pp. 684–686, 1997.

[4] J. E. Andruski, P. K. Kuhl, and A. Hayashi, "The acoustics of vowels in Japanese women's speech to infants and adults," in *Proceedings of the 14th International Congress on Phonetic Sciences*, vol. 3, 1999, pp. 2177–2179.

[5] D. Burnham, C. Kitamura, and U. Vollmer-Conna, "What's new, pussycat? On talking to babies and animals," *Science*, vol. 296, no. 5572, pp. 1435–1435, 2002.

[6] B. McMurray, K. A. Kovack-Lesh, D. Goodwin, and W. McEchron, "Infant directed speech and the development of speech perception: Enhancing development or an unintended consequence?" *Cognition*, vol. 129, no. 2, pp. 362–378, 2013.

[7] A. Cristia and A. Seidl, "The hyperarticulation hypothesis of infant-directed speech," *Journal of Child Language*, vol. 41, no. 4, pp. 913–934, 2014.

[8] A. Martin, T. Schatz, M. Versteegh, K. Miyazawa, R. Mazuka, E. Dupoux, and A. Cristia, "Mothers speak less clearly to infants than to adults: A comprehensive test of the hyperarticulation hypothesis," *Psychological Science*, vol. 26, no. 3, pp. 341–347, 2015.

[9] U. Sundberg and F. Lacerda, "Voice onset time in speech to infants and adults," *Phonetica*, vol. 56, no. 3-4, pp. 186–199, 1999.

[10] K. T. Englund, "Voice onset time in infant directed speech over the first six months," *First Language*, vol. 25, no. 2, pp. 219–234, 2005.

[11] K. Englund and D. Behne, "Changes in infant directed speech in the first six months," *Infant and Child Development*, vol. 15, no. 2, pp. 139–160, 2006.

[12] A. Cristia, "Phonetic enhancement of sibilants in infant-directed speech," *The Journal of the Acoustical Society of America*, vol. 128, no. 1, pp. 424–434, 2010.

[13] K. M. Kurowski and S. E. Blumstein, "Acoustic properties for the perception of nasal consonants," in *Nasals, Nasalization, and the Velum*. Elsevier, 1993, pp. 197–222.

[14] R. Mazuka, Y. Igarashi, and K. Nishikawa, "Input for learning Japanese: RIKEN Japanese mother-infant conversation corpus (COE Workshop session 2)," *IEICE Technical Report*, vol. 106, no. 165, pp. 11–15, 2006.

[15] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, pp. 341–345, 2002.

[16] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *arXiv preprint arXiv:1406.5823*, 2014.

[17] D. Bates, R. Kliegl, S. Vasishth, and H. Baayen, "Parsimonious mixed models," *arXiv preprint arXiv:1506.04967*, 2015.

[18] H. Matuschek, R. Kliegl, S. Vasishth, H. Baayen, and D. Bates, "Balancing type i error and power in linear mixed models," *Journal of Memory and Language*, vol. 94, pp. 305–315, 2017.

[19] B. Ripley, B. Venables, D. M. Bates, K. Hornik, A. Gebhardt, D. Firth, and M. B. Ripley, "Package mass," *Cran R*, 2013.

[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[21] C. Fougeron and P. A. Keating, "Articulatory strengthening at edges of prosodic domains," *The Journal of the Acoustical Society of America*, vol. 101, no. 6, pp. 3728–3740, 1997.

[22] F. Adriaans and D. Swingley, "Prosodic exaggeration within infant-directed speech: Consequences for vowel learnability," *The Journal of the Acoustical Society of America*, vol. 141, no. 5, pp. 3070–3078, 2017.

[23] K. Miyazawa, T. Shinya, A. Martin, H. Kikuchi, and R. Mazuka, "Vowels in infant-directed speech: More breathy and more variable, but not clearer," *Cognition*, vol. 166, pp. 84–93, 2017.

[24] B. Ludusan, A. Cristia, A. Martin, R. Mazuka, and E. Dupoux, "Learnability of prosodic boundaries: Is infant-directed speech easier?" *The Journal of the Acoustical Society of America*, vol. 140, no. 2, pp. 1239–1250, 2016.

[25] B. Ludusan, R. Mazuka, M. Bernard, A. Cristia, and E. Dupoux, "The role of prosody and speech register in word segmentation: A computational modelling perspective," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Short Papers)*, 2017, pp. 178–183.

[26] A. Guevara-Rukoz, A. Cristia, B. Ludusan, R. Thiollière, A. Martin, R. Mazuka, and E. Dupoux, "Are words easier to learn from infant- than adult-directed speech? A quantitative corpus-based investigation," *Cognitive Science*, vol. 42, no. 5, pp. 1586–1617, 2018.

[27] B. Ludusan, A. Seidl, E. Dupoux, and A. Cristia, "Motif discovery in infant-and adult-directed speech," in *Proceedings of the Sixth Workshop on Cognitive Aspects of Computational Language Learning*, 2015, pp. 93–102.

[28] N. Feldman, T. Griffiths, and J. Morgan, "Learning phonetic categories by learning a lexicon," in *Proceedings of the 31st Annual Meeting of the Cognitive Science Society*, 2009, pp. 2208–2213.

[29] R. P. Cooper and R. N. Aslin, "Developmental differences in infant attention to the spectral properties of infant-directed speech," *Child Development*, vol. 65, no. 6, pp. 1663–1677, 1994.

[30] L. Singh, J. L. Morgan, and C. T. Best, "Infants' listening preferences: Baby talk or happy talk?" *Infancy*, vol. 3, no. 3, pp. 365–394, 2002.