UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

Language Discrimination May Not Rely on Rhythm: A Computational Study

Permalink

https://escholarship.org/uc/item/9qx164nx

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 46(0)

Authors

Famularo, Ruolan Leslie
Aboelata, Ali
Schatz, Thomas
et al.

Publication Date

2024

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Language Discrimination May Not Rely on Rhythm: A Computational Study

Ruolan Leslie Famularo[1,2,3], **Ali Aboelata**[2], **Thomas Schatz**[4], **Naomi H. Feldman**[3,5]
rlli@umd.edu, aaboelat@terpmail.umd.edu, thomas.schatz@lis-lab.fr, nhf@umd.edu
[1] Program in Neuroscience and Cognitive Science, University of Maryland, College Park, MD 20742, USA
[2] Department of Computer Science, University of Maryland, College Park, MD 20742, USA
[3] Department of Linguistics, University of Maryland, College Park, MD 20742, USA
[4] Aix Marseille Univ, CNRS, LIS, Marseille, France
[5] Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742, USA

Abstract

It has long been assumed that infants' ability to discriminate between languages stems from their sensitivity to speech rhythm, i.e., organized temporal structure of vowels and consonants in a language. However, the relationship between speech rhythm and language discrimination has not been directly demonstrated. Here, we use computational modeling and train models of speech perception with and without access to information about rhythm. We test these models on language discrimination, and find that access to rhythm does not affect the success of the model in replicating infant language discrimination results. Our findings challenge the relationship between rhythm and language discrimination, and have implications for theories of language acquisition.

Keywords: rhythm; language discrimination; speech perception; language acquisition; computational modeling

Introduction

Humans are able to notice a switch between some languages but not others. Even in newborns, language pairs such as English and Japanese are discriminated while pairs like English and Dutch are not (Nazzi, Bertoncini, & Mehler, 1998), a phenomenon we refer to as language discrimination. To date, the only acoustic measurements from the speech signal that have been directly correlated with humans' discrimination behavior are global measures such as the overall percentage of vowels (%V) and the variability of consonantal interval durations (ΔC) in an utterance (Ramus, Nespor, & Mehler, 1999). These global measures have been argued to correlate with the organized temporal structure of vowels and consonants in a language, which linguistically constructs speech rhythm. As a result, rhythm has long been assumed to drive infants' language discrimination (Mehler et al., 1988; Nazzi et al., 1998; Ramus et al., 1999; Nazzi & Ramus, 2003; but see Gasparini, Langus, Tsuji, & Boll-Avetisyan, 2021), but this has not been directly tested.

Nonetheless, the rhythmic basis for language discrimination has already generated various theoretical predictions. As language discrimination was one of the few observations we have about newborns and rhythm has been assumed as its underlying cause, the acquisition of speech rhythm was argued to be one of the earliest processes in language acquisition (Nazzi & Ramus, 2003). The sensitivity to and acquisition of speech rhythm, which primarily rely on language discrimination as behavioral evidence, have also been incorporated into theories of early language acquisition and critical period (Werker & Hensch, 2015; Gervain, Christophe, & Mazuka, 2020). Within these frameworks, rhythm is placed early in language development, and knowledge of rhythm is considered to help with later acquisition of phonetic categories, lexical stress, and word segmentation. In other words, the order of acquisition and relationship between different aspects of language is built upon the assumption that rhythm drives newborns' language discrimination.

In computational modeling, rhythm has also been implicitly assumed to be the cause for language discrimination, even without controlled comparisons. For example, in Dominey and Ramus (2000), when a small recurrent neural network (RNN) model replicated infant language discrimination, the replication was directly attributed to the recurrent structure successfully capturing the sequential rhythmic information of vowel and consonantal alternation. Recently, a computational model has been built to simulate the cognitive process of language discrimination (Carbajal, Fér, & Dupoux, 2016; Carbajal, 2018). This model used acoustic features that characterize speech information on the slower and suprasegmental level to represent the use of rhythm in language discrimination, and successfully replicated language discrimination results in infants. The success of this model was attributed to the inclusion of rhythm in the acoustic features, although the authors never tested controls that did not have access to rhythm.

In this paper, we challenge the claim that rhythm is important or even necessary in language discrimination. We replicate the model in Carbajal et al. (2016) using features in which rhythmic information was removed. Additionally, following the RNN precedence in Dominey and Ramus (2000), we test a newer RNN model that was built for real speech. Lastly, we used acoustic features directly as representations for testing, without model and training. We test all of our models using acoustic features extracted directly from the test stimuli, as well as acoustic features that are temporally scrambled to remove rhythmic information. In all models tested, we find that language discrimination is not any less successful with scrambled stimuli compared with stimuli that retained rhythmic information. This suggests rhythm is not necessary for language discrimination, contradicting longstanding assumptions. The computational simulations also suggest that short-time acoustics is enough for language discrimination, suggesting that human newborns could also use segmental information alone to achieve language discrimination. Our

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.* ©2024 The Author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

results have implications for theories of early language acquisition, challenging previous assumptions about what about rhythm is acquired by infants as well as when this happens.

Simulation overview

In our simulations, we focus on replicating a series of newborn discrimination studies. In the set of behavioral studies (e.g., Nazzi et al., 1998), French newborns (3-day-old infants) passively listened to sentences of one language for a few minutes. Then, the sentences switched to a novel language. If the infants' sucking rate increased, then the infants were said to discriminate between the two languages. In our computational simulations, we conceptually replicate this experiment through the machine ABX task, using the various representations from trained models or acoustic features to compute distances within a language and to a novel language. Below, we introduce each of the models as well as the test paradigm.

Models

We choose two models based on the literature of language discrimination. Firstly, directly based on Carbajal et al. (2016), we used the same i-vector model as was used in their experiments, which is a generative model that performs unsupervised clustering. Secondly, we used a RNN as a conceptual replication of the computational model by Dominey and Ramus (2000). In Dominey and Ramus (2000), a small RNN demonstrated human-like language discrimination from heavily annotated speech. Here, we used a state-of-the-art model with real speech input, which learns through predictive coding.

i-vector model First, we replicate the model used by Carbajal et al. (2016). A Gaussian Mixture Model (an unsupervised clustering model) was trained on a small amount of French, representing young infants' limited exposure to their native language. Then, a low-dimensional shift in means was calculated from each trained Gaussian cluster to each test stimulus. This low-dimensional shift, called an i-vector (Dehak, Torres-Carrasquillo, Reynolds, & Dehak, 2011), represents the shift from the training data to the test, which is the shift from French to the new language in this setting.

The i-vector model itself cannot access the order of individual frames in the acoustic features, which are only 25 ms long. In Carbajal et al. (2016), the acoustic features were expanded with shifted delta coefficients (SDCs), which contain information about the local change of the acoustic features within the neighboring 200 ms. Such expanded acoustic features allowed the model to capture rhythmic sequences. We replicated this model as the "original" model. To examine the language discrimination outcome of the model when rhythmic information is not accessible, we also trained and tested models with only the acoustic features from the original model (which are mel-frequency cepstral coefficients; Davis & Mermelstein, 1980; "MFCC" model). As the MFCC model cannot access the order of the acoustic features, it can only perform language discrimination using information other than rhythm. Following the original study, the MFCCs were extracted with a 25 ms window and 10 ms moving window. For each type of model, we trained four models on disjoint sets of training data. The training data for each model are approximately one-hour long. All training data were drawn from the French Globalphone corpus (Schultz, 2002).

RNN model In addition to the i-vector model above, which is generative and probabilistic, we also test a RNN that learns through predictive coding. We use a small version of an Autoregressive Predictive Coding model (Chung & Glass, 2020), with a vector quantizing layer to facilitate learning bottleneck (Chung, Tang, & Glass, 2020). This self-supervised model learns by predicting the upcoming speech material some tens of milliseconds ahead. We train the RNN model on French speech, representing the young infants' exposure to their native language. After training, the embeddings from the RNN model contain information useful for tasks like speaker identification, phoneme classification, and emotion classification. In this project, we use the embedding to test the model's language discrimination.

We used models containing 3 recurrent layers of 32 hidden units per layer and codebook size 16 in vector quantizing. The network predicts 9 frames (i.e. 90 ms) in the future. We chose these hyperparameters to build a smaller model than the original model in Chung et al. (2020). Since the original model was used for speech technology tasks and included much more training data, we scaled down the model size here to fit a smaller training set in order to prevent overfitting. For each model, we report language discrimination results after 10k, 100k, and 500k minibatch updates to examine the effect of learning on language discrimination. We used embeddings from the second layer of the RNN for language discrimination tests. All models are trained on the entire set of the French Globalphone data, which contains 23.2 hours of speech.

The input to the RNN model was composed of mel spectrograms with 25 ms window length and 10 ms moving window. Different from typical usage of the spectrogram, which contains 80 frequency channels, here, we report results with only 8 frequency channels to reduce frequency resolution, through selecting broader-band frequency filters along the mel scale. Although we obtain qualitatively similar results with 80-channel spectrograms, we select the 8-channel spectrogram here to further reduce segmental information such as formants.

Acoustic Feature Representations In addition to the two models, which both represent the test stimuli in terms of their trained (i.e., "native") language, French, we also test language discrimination using only acoustic representations that were not trained with any model. For this, we directly take the mel-frequency spectrograms that are input to the RNNs as the representation for test.

Test

The ABX test We use the machine ABX task (Schatz et al., 2013; Schatz, 2016). In each individual trial, three random utterances A, B, and X are drawn from the tested language

pair. The cosine distance $d(\cdot)$ is taken between (A,X) as well as (B,X). Suppose X and A are from the same language, then if the distance $d(A,X)$ is smaller than $d(B,X)$, the machine is correct in this trial, and vice versa. Since we have a large number of utterances, we randomly sampled A, B, and X for each trial and made 2000 independent draws for each model and condition. Each language in the tested language pair is the correct answer half of the time (i.e., 1000 trials).

For the i-vector model, since each utterance produces one i-vector of fixed length, the ABX task was directly applied to the i-vectors. For the RNN model and acoustic representations, since representations depend on the duration of the utterance, in each trial, one second[1] of speech material was randomly sampled from each test utterance. Dynamic time warping was used to align the two one-second samples in the distance function $d(\cdot)$.

Conditions To examine whether rhythm is necessary for language discrimination, we applied scrambling to remove rhythmic information in the test stimuli for each model and condition. During scrambling, the obtained acoustic features, each computed over 25 ms of the speech signal, were shuffled within the utterance randomly. This scrambling procedure is designed to remove any information contained in slow temporal regularities, such as rhythm. If scrambling does not remove the language discrimination effect, then rhythm is not necessary for language discrimination.

In behavioral experiments, speech stimuli were often low-pass filtered to remove segmental information such as formants (Mehler et al., 1988; Nazzi et al., 1998). In our simulations, we used both original and low-pass filtered speech as test stimuli. For low-pass filtering, we used a 4-th order Butterworth filter with cutoff frequency at 400 Hz. We have stronger predictions on the low-pass filtered condition, since there exist direct behavioral results to compare with (Nazzi et al., 1998). Additionally, we report results on natural speech, as many studies on newborns and infants a few months older used natural speech and found similar results (Moon, Cooper, & Fifer, 1993; Dehaene-Lambertz & Houston, 1998; Nazzi, Jusczyk, & Johnson, 2000).

Simulation I

We first test our models on the stimuli used in Carbajal (2018), the UCAM Bilingual Corpus.[2] This corpus contains speech of bilingual speakers of English and another language, with read sentences from one bilingual speaker per language pair. We picked three language pairs due to their close relationship with behavioral studies: English-Dutch, English-Italian, and English-French. Among the three, English-Dutch was directly tested in behavioral studies like Nazzi et al. (1998). For English-Italian, the low-pass filtered stimuli from this language pair were indirectly discriminated from an experiment testing a mixture of languages in Nazzi et al. (1998), where infants noticed a switch from a mixture of English and Dutch to another mixture from Spanish and Italian. Lastly, English-French is different from the other pairs due to French being the native language of the young infants. We are not aware of work that directly tested this language pair in newborns, but it was found to be distinguished by French 2-month-olds through language preference (Dehaene-Lambertz & Houston, 1998). In addition to the related behavioral experiments, the global acoustic correlates in Ramus et al. (1999), %V and ΔC, also predict no discrimination for English-Dutch, but discrimination for English-Italian and English-French. We consider a model's behavior to be human-like if it discriminates English-Italian and English-French better than English-Dutch.

Our results for Simulation I are shown in Figure 1. First of all, we replicated Carbajal (2018). In the Full version of the i-vector model, the pattern of results (English-Italian and English-French were significantly better discriminated than English-Dutch) matches the original study. Additionally, the numerical ABX error rates were close to the those in Carbajal (2018), which suggests that the replication was successful.

Next, looking at the ablated i-vector models that cannot access rhythmic information, we note that the MFCC model consistently showed the same direction of effects, which challenges the original conclusion that i-vector models needed suprasegmental information to perform human-like language discrimination. In the Scrambled condition, the results were more mixed on English-Italian, but English-French was consistently better discriminated than English-Dutch.

Similarly, in the RNN model as well as the acoustic features, human-like language discrimination behavior was observed when the stimuli were low-pass filtered. All models showed human-like discrimination with low-pass filtered speech regardless of their access to rhythmic information.

When the stimuli were not low-pass filtered, however, the results were more mixed. While English-French was consistently discriminated better compared with English-Dutch, English-Italian showed mixed results across the three types of models. Since the results were more consistent in the low-pass filtered condition, which contains less acoustic detail of the test stimuli, it is possible that acoustic details in the English-Italian data may have lead to different representations across three model representations. Additionally, across all the results, we observed a consistent advantage in discriminating English-French — this language pair was significantly better discriminated than English-Dutch, our baseline, in almost all models and conditions. In the i-vector and RNN models especially, the error rate for English-French was often much lower than the other two languages, especially under certain conditions when rhythmic information was not accessible. This can be explained by a native advantage, namely that the models were better at discriminating the language that they were trained on from a novel language. Nonetheless, this advantage was also observed in the acoustic features, where no training was done. Therefore, it is also possible that the speech material

[1] We also ran tests with two and three seconds of speech material, with qualitatively similar results.

[2] The EMIME project, https://www.emime.org.

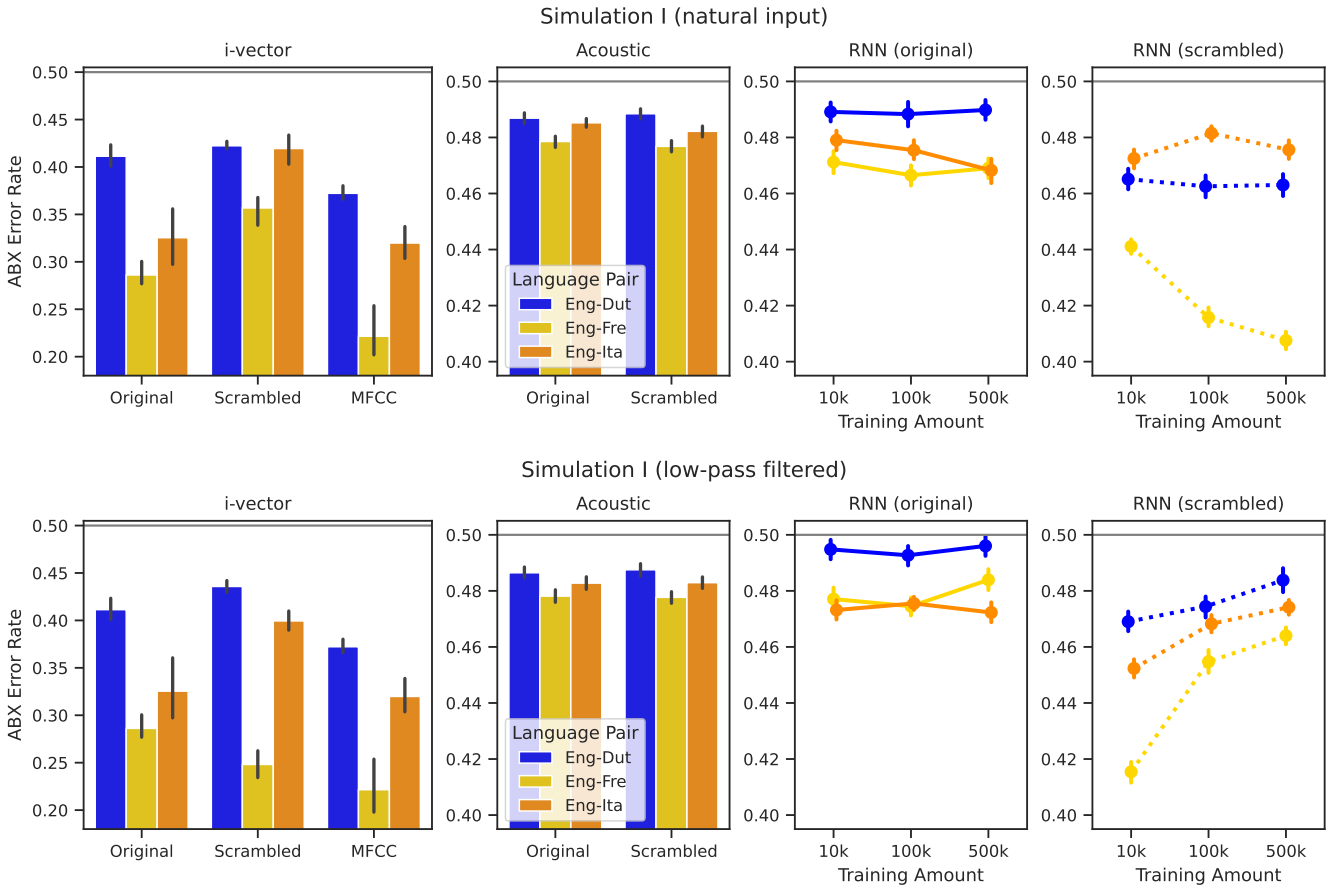

Figure 1: Results for Simulation I. In the i-vector model and acoustics, the x-axis denotes different conditions. In the RNN model, the x-axis denotes the same model trained for different amount of steps. The y-axis in all plots shows the performance of the model on language discrimination through ABX error rate, where lower scores indicate better performance. Error bar for the i-vector models indicate the standard error obtained from four models trained on different speech data. For the acoustics and RNN models, the error bars indicate standard error obtained from 20 independent ABX tests. The grey horizontal line indicates chance level for the ABX task (50%). Note that the y-axis for the i-vector model used a different scale due to much lower error rates in the i-vector model.

for English-French was acoustically more different.

The test corpus used in Simulation I has two potential limitations. The first one comes from the speaker distribution. The UCAM corpus only contains one speaker per language pair, which makes it hard to discern speaker-specific characteristics from language-specific characteristics. Additionally, since the corpus contains only bilingual speakers, it may not be comparable to the monolingual productions that infants were tested on (e.g., in Nazzi et al., 1998). As analyses from multiple languages confirm that bilingual production of speech has distinct rhythm measures compared with monolingual speakers of either language (Lin & Wang, 2008; Li & Post, 2014), using bilingual speakers may confound the results in unpredictable ways. Secondly, with this corpus the behavioral experiments cannot be easily compared due to different language pairs being used. In behavioral studies, English-Italian was only tested when mixed with two other languages (Dutch and Spanish), so it remains uncertain which languages drive the effect of language discrimination. In natural speech, newborns' discrimination of English-Italian only resulted in a very weak effect (Mehler et al., 1988; Mehler & Christophe, 1995). Additionally, to our knowledge, there is no published data on the language discrimination between English and French in newborns. This motivates us to extend the current paradigms to directly simulate some behavioral experiments in the infant language discrimination literature, which we report in the next set of simulations.

Simulation II

In the previous simulation, we replicated Carbajal (2018) in low-pass filtered speech, but results on natural speech were more mixed. Additionally, all the representations tested showed a drastic advantage in the English-French pair, which cannot be explained by a native advantage alone. This lack of explainability may have been due to confounds specific to the test corpus, namely the language pairs available and the single-speaker setting. In the second set of simulations,

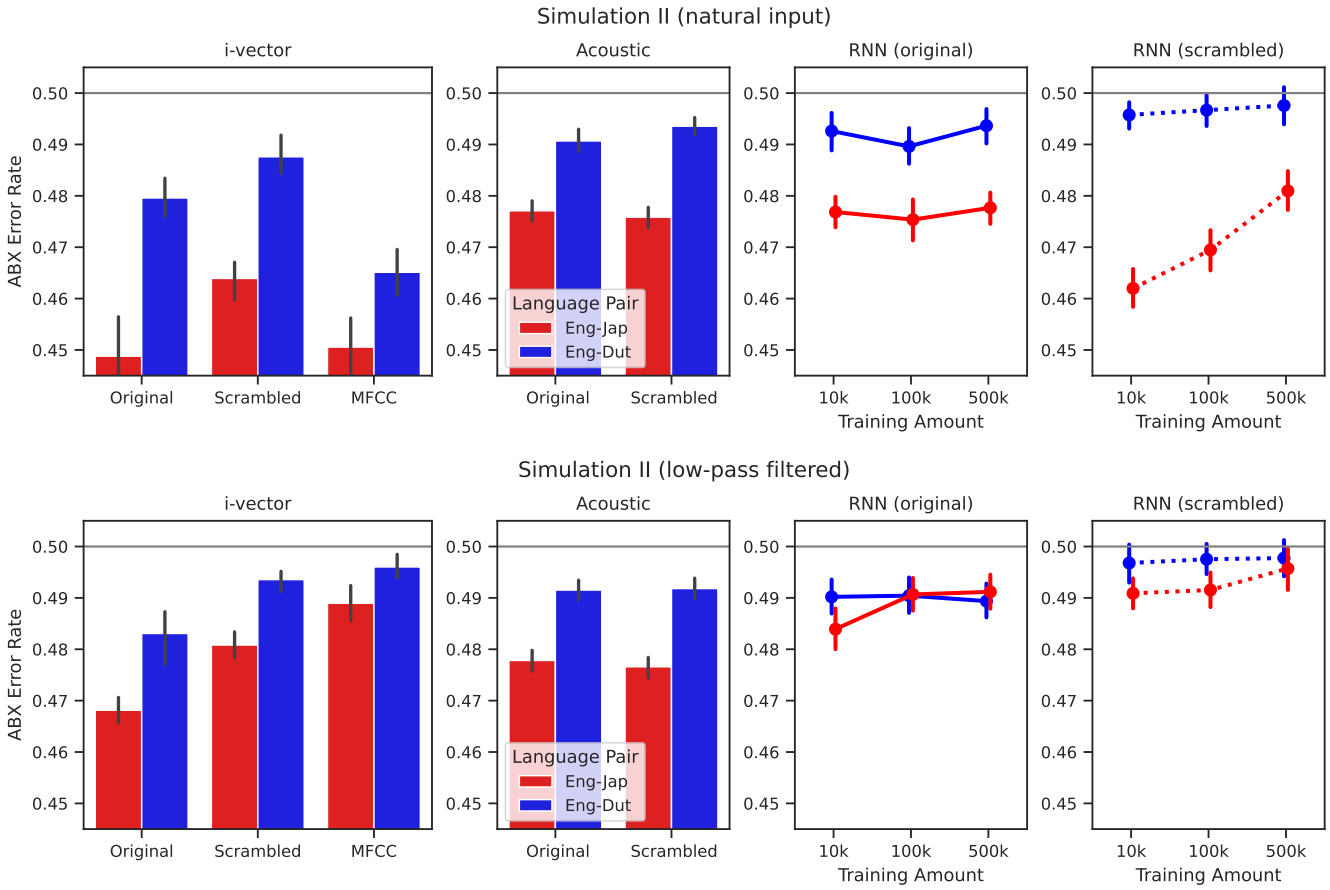

Figure 2: Results for Simulation II. All plotting choices followed those in Figure 1.

we use a different corpus that contains more speakers and languages as test data, and test two language pairs directly used in the behavioral study reviewed above (Nazzi et al., 1998) — English-Japanese and English-Dutch. For a model to have human-like performance, English-Japanese should be significantly better discriminated than English-Dutch.

To test the models, we used utterances from the Common Voice corpus (v. 13.0, Ardila et al., 2019). Among all sentences, we discarded sentences with downvotes (i.e., rated by corpus contributors to have noise, dis-fluency, or bad otherwise; around 20%). We selected utterances that are between 4 and 10 seconds long, and the utterances were shuffled to avoid selecting multiple clips from the same speaker. Then, for each test language, we manually listened to the audio files and selected the first 100 utterances without significant noise, disfluencies, or obvious non-native accents. Through the manual selection process, the criteria filtered out roughly 50% of the utterances. All utterances were root-mean-square normalized before any further processing.

The results are shown in Figure 2. Under low-pass filtered speech, the i-vector model showed human-like discrimination in all conditions, regardless of accessibility to temporal information. The same holds true for the acoustic features. For the RNN model, there was a null effect regardless of scrambling, although in the least-trained version of the model (10k minibatch udpates), a marginal but not significant effect was observed. The lack of discrimination in the RNN can potentially be explained by out-of-distribution problems. Since the training data (natural speech) and the test data (low-pass filtered speech) were different acoustically, the RNN model, which never saw any low-pass filtered data during training, may generate embeddings that are too noisy for language discrimination. This, along with the variability of the speaker composition and recording quality, might have made this task too hard for the RNN. Supporting this explanation, in all RNN models that were tested on low-pass filtered stimuli, any observed discrimination effect became weaker as the training amount increased, which suggests that as the RNN model fits to the training data, it contains less language-specific details in low-pass filtered speech. Overall, despite the null effect from the RNN model, for any effect that was significant in the other models, temporal scrambling did not remove the effect, which suggests that rhythm was not necessary for human-like language discrimination in any of the models.

In natural speech, all models and conditions demonstrated human-like performance. This suggests that across all three representations, rhythm is not necessary in language discrimination. Surprisingly, in the RNN model, the crosslinguis-

tic effect (i.e. the difference between English-Japanese and English-Dutch) was greater when the stimuli were scrambled. This seems unintuitive, but can be explained if global properties, as Ramus et al. (1999) found, were directly used in language discrimination. For example, if %V was directly used for discrimination, then scrambling would make the frames containing vocalic intervals distribute in a more uniform manner, thus decreasing the error rate of language discrimination when %V differs between the pair of languages (i.e., in the case of English-Japanese), and increasing the error rate when %V does not differ much (i.e. in English-Dutch).

Discussion

In this paper, we examined the long-standing claim that rhythm drives language discrimination by building computational models of speech perception with and without access to rhythm. Our results suggest the models' success in discriminating languages like human newborns was independent of its access to rhythm, indicating that rhythm was not crucial to language discrimination. These results challenge the relationship between language discrimination and speech rhythm, and have implications for language acquisition theories that center on the early acquisition of rhythm.

Our results suggest that the speech signal still contains abundant information sufficient for language discrimination even after rhythm is removed through scrambling or feature engineering. This includes segmental information, such as spectral properties for vowels and consonants that differ across languages, and global information, such as the proportion and distribution of different vowels and consonants. Such segmental and global information in the speech signal correlates with %V (percentage of vowels) and ΔC (variability of consonantal intervals), two known measurements directly correlated with human discrimination. On one hand, %V directly reflects the global property and remains unchanged after scrambling. On the other hand, while ΔC would not be directly retained in scrambled speech, it has been argued that ΔC indirectly characterizes "rhythm" through syllable structures and stress patterns (Ramus et al., 1999; Langus, Mehler, & Nespor, 2017). It has been argued that languages with greater ΔC have more complex syllable structures, which often correlate with a greater variety of consonant clusters (Ramus et al., 1999; Dauer, 1983). In languages with a greater ΔC like English and Dutch, stressed and unstressed syllables also often differ greatly in acoustic cues such as intensity and spectral properties (Dauer, 1983). In the computational models we tested, such acoustic and global differences can be easily represented in a mechanism such as the lower-dimensional shift between training and test (i.e., i-vector) or accumulating local frames to generate a representation of the distribution of acoustic properties through the RNN embeddings. In either case, rhythm, or the sequential structure of segmental information, was not involved in the perceptual mechanism.

Looking across all the error rate values, one can notice that the absolute value of the error rate depends greatly on the model, test corpus, and condition. Particularly, the i-vector models seemingly had lower error rates compared with other models. This can be attributed to the representation of the i-vector model being time-independent, while all other models have representations that change along time, which must be aligned through dynamic time warping. In some cases, especially in Simulation II, error rates were quite high and close to chance (50%). Considering the differences between the corpora in the two simulations, we argue that the increased variability in speakers and recording conditions due to crowdsourcing is a major reason. While the corpus used in Simulation I was recorded in-lab with professional microphones and contained only one speaker per language pair, the corpus in Simulation II was recorded by volunteers on the internet, and contained one speaker per utterance. In the behavioral experiments (e.g., in Nazzi et al., 1998), the stimuli used were controlled in terms of speaker gender and voice quality to sound as similar as possible, and there were only two speakers per language. Arguably, the variability in Simulation II was much greater than that of the stimuli in infant experiments.

Our results are consistent with other work that has challenged the relevance of rhythm to language discrimination or the importance of rhythm in young infants. For example, in De Seyssel, Wisniewski, Dupoux, and Ludusan (2022), a different i-vector model was trained to classify languages typologically. Unlike Carbajal et al. (2016), their model did not have access to rhythm information. Nonetheless, their typological map showed patterns of similarity across languages that are often attributed to rhythm, such as English being closer to Dutch than to Spanish.

Even if speech rhythm is unnecessary for explaining the language discrimination effect in infants, it may still be perceived and learned by infants in other ways. For example, newborns have been found to be sensitive to different cues to rhythm in a language-specific way (Abboub, Nazzi, & Gervain, 2016). However, this type of sensitivity involves perceiving different cues (e.g., duration vs. pitch) in different languages, instead of being sensitive to the specific structure of durational cues in different languages. As such, our results may weaken some theories about the perception of crosslinguistic differences in rhythm. For example, Nazzi and Ramus (2003) argued through a series of language discrimination experiments that infants gradually learn the rhythm of their native language during the first few months. However, as our results call into question the relationship between rhythm and language discrimination, it becomes questionable whether the behavioral evidence implies acquisition of rhythm, or rather, segmental and global properties. As another example, bilingual infants were argued to be slower in their acquisition if the two languages were rhythmically similar (Sundara & Scutellaro, 2011). Based on our results, it may be worth considering alternative features that infants might be relying on, such as the acoustic similarity between the two languages. Overall, our results invite reconsideration of the timing and content of rhythmic acquisition in language development.

Acknowledgments

This work was supported by NSF grant BCS-2120834. T.S.' contribution to the work, carried out within the Institute of Convergence ILCB, was supported by grants from France 2030 (ANR-16-CONV-0002) and the Excellence Initiative of Aix-Marseille University (A*MIDEX).

References

Abboub, N., Nazzi, T., & Gervain, J. (2016). Prosodic grouping at birth. *Brain and Language*, *162*, 46–59.

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., ... Weber, G. (2019). Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

Carbajal, M. J. (2018). *Separation and acquisition of two languages in early childhood: A multidisciplinary approach*. Unpublished doctoral dissertation, Université Paris sciences et lettres.

Carbajal, M. J., Fér, R., & Dupoux, E. (2016). Modeling language discrimination in infants using i-vector representations. In *Proceedings of the Annual Conference of the Cognitive Science Society.*

Chung, Y.-A., & Glass, J. (2020). Generative pre-training for speech with autoregressive predictive coding. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 3497–3501).

Chung, Y.-A., Tang, H., & Glass, J. (2020). Vector-quantized autoregressive predictive coding. *arXiv preprint arXiv:2005.08392*.

Dauer, R. M. (1983). Stress-timing and syllable-timing reanalyzed. *Journal of Phonetics*, *11*(1), 51–62.

Davis, S. B., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *Proceedings of the IEEE*, 357-366.

Dehaene-Lambertz, G., & Houston, D. (1998). Faster orientation latencies toward native language in two-month-old infants. *Language and Speech*, *41*(1), 21–43.

Dehak, N., Torres-Carrasquillo, P. A., Reynolds, D. A., & Dehak, R. (2011). Language recognition via i-vectors and dimensionality reduction. In *Interspeech* (pp. 857–860).

De Seyssel, M., Wisniewski, G., Dupoux, E., & Ludusan, B. (2022). Investigating the usefulness of i-vectors for automatic language characterization. In *Speech Prosody 2022-11th International Conference on Speech Prosody.*

Dominey, P. F., & Ramus, F. (2000). Neural network processing of natural language: I. sensitivity to serial, temporal and abstract structure of language in the infant. *Language and Cognitive Processes*, *15*(1), 87–127.

Gasparini, L., Langus, A., Tsuji, S., & Boll-Avetisyan, N. (2021). Quantifying the role of rhythm in infants' language discrimination abilities: A meta-analysis. *Cognition*, *213*(August 2020), 104757. doi: 10.1016/j.cognition.2021.104757

Gervain, J., Christophe, A., & Mazuka, R. (2020). Prosodic bootstrapping.

Langus, A., Mehler, J., & Nespor, M. (2017). Rhythm in language acquisition. *Neuroscience & Biobehavioral Reviews*, *81*, 158–166.

Li, A., & Post, B. (2014). L2 acquisition of prosodic properties of speech rhythm: Evidence from l1 Mandarin and German learners of English. *Studies in Second Language Acquisition*, *36*(2), 223–255.

Lin, H., & Wang, Q. (2008). Interlanguage rhythm in the English production of Mandarin speakers. In *Proceedings of the 8th Phonetic Conference of China and the International Symposium on Phonetic Frontiers* (pp. 18–20).

Mehler, J., & Christophe, A. (1995). Maturation and learning of language in the first year of life. In *Gazzaniga, M.S. (Ed.) The Cognitive Neurosciences* (pp. 943–954). Bradford Books, Cambridge, MA.

Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertoncini, J., & Amiel-Tison, C. (1988). A precursor of language acquisition in young infants. *Cognition*, *29*(2), 143–178.

Moon, C., Cooper, R. P., & Fifer, W. P. (1993). Two-day-olds prefer their native language. *Infant Behavior and Development*, *16*(4), 495–500.

Nazzi, T., Bertoncini, J., & Mehler, J. (1998). Language discrimination by newborns: Toward an understanding of the role of rhythm. *Journal of Experimental Psychology: Human perception and performance*, *24*(3), 756.

Nazzi, T., Jusczyk, P. W., & Johnson, E. K. (2000). Language discrimination by English-learning 5-month-olds: Effects of rhythm and familiarity. *Journal of Memory and Language*, *43*(1), 1–19.

Nazzi, T., & Ramus, F. (2003). Perception and acquisition of linguistic rhythm by infants. *Speech communication*, *41*(1), 233–243.

Ramus, F., Nespor, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, *73*(3), 265–292.

Schatz, T. (2016). *ABX-discriminability measures and applications*. Unpublished doctoral dissertation.

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 *INTERSPEECH 2013: 14th Annual Conference of the International Speech Communication Association* (pp. 1–5).

Schultz, T. (2002). Globalphone: A multilingual speech and text database developed at Karlsruhe University. In *Seventh International Conference on Spoken Language Processing.*

Sundara, M., & Scutellaro, A. (2011). Rhythmic distance between languages affects the development of speech perception in bilingual infants. *Journal of Phonetics*, *39*(4), 505–513.

Werker, J. F., & Hensch, T. K. (2015). Critical periods in speech perception: New directions. *Annual Review of Psychology*, *66*, 173–196.