NOTE: This version contains updated numbers in the top half of Table 1 after correcting a bug in the simulation code. Results are qualitatively the same as they were in the original paper.

Modeling Rhythm in Speech as in Music: Towards a Unified Cognitive Representation

Ruolan Li (rlli@umd.edu)
Program in Neuroscience and Cognitive Science, Department of Linguistics, & UMIACS, University of Maryland
College Park, MD 20742, USA

Thomas Schatz (thomas.schatz@univ-amu.fr)
Laboratoire Informatique et Systèmes & Institute of Language, Communication and the Brain & Turing Center for Living Systems
Aix Marseille University & CNRS, Marseille, France

Naomi H. Feldman (nhf@umd.edu)
Department of Linguistics, & UMIACS, University of Maryland
College Park, MD 20742, USA

Abstract

Rhythm plays an important role in language perception and learning, with infants perceiving rhythmic differences across languages at birth. While the mechanisms underlying rhythm perception in speech remain unclear, one interesting possibility is that these mechanisms are similar to those involved in the perception of musical rhythm. In this work, we adopt a model originally designed for musical rhythm to simulate speech rhythm perception. We show that this model replicates the behavioral results of language discrimination in newborns, and outperforms an existing model of infant language discrimination. We also find that percussives — fast-changing components in the acoustics — are necessary for distinguishing languages of different rhythms, which suggests that percussives are essential for rhythm perception. Our music-inspired model of speech rhythm may be seen as a first step towards a unified theory of how rhythm is represented in speech and music.

Keywords: rhythm; music; speech; language perception; language and music; computational modeling

Rhythm is important in both speech and music. In speech, rhythm is one of the first things infants perceive and learn about their native language (Nazzi, Bertoncini, & Mehler, 1998; Nazzi & Ramus, 2003). For example, newborns discriminate between English and Japanese, which are rhythmically different (Nazzi et al., 1998), but do not discriminate between English and German, which are rhythmically similar, until they are 7 months old (Chong, Vicenik, & Sundara, 2018). In music, rhythm is a primary structural element, and the rhythmic pattern of a tune can be strongly characteristic of a genre, style, or musical culture (London, 2001). Cognitively, musical rhythm correlates with the rhythm of composers' native language (Patel & Daniele, 2003). Neurally, there exist shared pathways for rhythm perception in music and language (see Kotz, Ravignani, & Fitch, 2018 for a review). Moreover, musical training in rhythm improves speech rhythm encoding (Harding, Sammler, Henry, Large, & Kotz, 2019) and language perception in general (Slater, Azem, Nicol, Swedenborg, & Kraus, 2017; Slater et al., 2018). These connections imply the possibility of a cognitive representation of rhythm that is shared in both domains.

In this work, we examine the connection between speech and musical rhythm by applying a model of musical rhythm (Tsunoo, Ono, & Sagayama, 2009) to simulate speech rhythm perception. In addition, we are interested in asking whether features that are important in music rhythm detection also facilitate rhythmic discrimination in language. In music, rhythmic patterns are often marked by short, transient acoustics such as drums, and isolating the percussive components from the music stream using Harmonic-Percussive Source Separation can lead to a better rhythm representation for downstream tasks (Ono et al., 2010; Fitzgerald, 2010). Here, we separately model the harmonic (slow-changing features such as vowels, pitch and intonation contour) and percussive components (fast-changing features such as syllable onsets and consonants) of speech to test whether percussives in speech can represent rhythm, as they do in music.

We use the model to simulate two language discrimination experiments, one between English and Japanese and the other between English and German. We also compare our results to a computational model that has previously been used to replicate a number of language discrimination experiments in newborns (Carbajal, Fér, & Dupoux, 2016; Carbajal, 2018). We find that our model replicates newborns' language discrimination behavior, unlike the baseline model. Importantly, however, the model is successful only when using a representation with percussives.

Methods

We test models on two pairs of languages and compare the models' discrimination with that of newborns. In the behavioral study of newborn language discrimination (Nazzi et al., 1998), French 3-day-old infants are exposed to spoken sentences by multiple speakers in one language until they are habituated; then, the stimuli change to a new speaker either in the same language or a new language. A difference in infants' response between the two conditions is taken as evidence that they can distinguish the two languages. As in Carbajal et al. (2016),

we simulate the behavioral paradigm by training each model on 4 French speakers, with 15 minutes per speaker, which serves as the brief exposure the infants have to their native language before they are tested in the lab. After training, the model is presented with utterances of different languages, and a machine ABX score is computed on these utterances to simulate discrimination between languages. The training and test data are selected from the Wall Street Journal corpus (Paul & Baker, 1992) and the Globalphone corpus (Schultz, 2002).

The rhythm model that we adopt from Tsunoo et al. (2009) is designed in the following way. The model is parameterized by a set number (6) of templates, each of which represents a recurring rhythmic pattern from speech. Each template is composed of 40 frames (920 ms), where each frame is an independent multivariate Gaussian distribution. Instances of the template in the speech stream are assumed to be drawn from the corresponding multivariate Gaussians. Using a dynamic programming algorithm (Ney, 1984), a speech stream of arbitrary length can be optimally aligned to the templates. Each template can be matched to a stretch of speech between 0.5 and 2 times its length, which allows the model to match rhythmic patterns of flexible length. We extract spectral features from the speech data using Short-Time Fourier Transform on every 46 ms (one frame) of speech, with a 23 ms moving window. Following Tsunoo et al. (2009), the spectral features in the 0–8 kHz range are averaged into eight 1 kHz-wide bins, leading to 8 dimensions per frame.

Using Harmonic-Percussive Source Separation, we separate the temporally continuous components of speech (harmonics) from the spectrally continuous component (percussives). We train and test models using one of the three representations: harmonics, percussives, or natural (both harmonics and percussives).

Training is done through Expectation Maximization and results in an optimized set of parameters for the templates. At test, each utterance is aligned to the models' templates using the dynamic programming algorithm, and the average log likelihood of the test utterance under the model is calculated. The discriminability of the log-likelihoods for utterances from different test languages is assessed by computing machine ABX discrimination scores (Schatz et al., 2013; Schatz, 2016).

As a baseline, we also train the model proposed by Carbajal et al. (2016). This model uses 64-dimensional features composed of 7 MFCC features with pitch track and 56 Shifted Delta Coefficients (Torres-Carrasquillo et al., 2002). This captures short-time information in speech as well as local changes within a 200 ms window, including intonation. The model's shift towards each test utterance, so-called i-vectors (Dehak, Kenny, Dehak, Dumouchel, & Ouellet, 2010), are calculated and ABX tasks are run like the above, but with i-vectors.

Results

The top section of Table 1 gives the results of simulating language discrimination behavior in newborns using the musically-inspired model from Tsunoo et al. (2009). When models are trained using natural and percussive features, performance aligns exactly with newborns' behavior. The model trained with harmonic features, however, does not discriminate between either pair of languages.

Table 1: ABX accuracy for both models. Perfect discrimination is 100%; chance discrimination is 50%. Newborns discriminate English and Japanese, but not English and German.

Rhythm Model	Harmonic	Percussive	Natural
Eng. vs. Jap.	44.03%	64.94%	55.12%
Eng. vs. Ger.	48.75%	47.14%	46.67%

Baseline Model	Full	MFCC	MFCC+pitch
Eng. vs. Jap.	70.20%	57.35%	60.49%
Eng. vs. Ger.	99.82%	51.80%	65.24%

In contrast, the baseline model discriminates between both pairs of languages. Its performance for English and German is near perfection, which is not similar to newborns' behavior, as infants are not able to discriminate between these two languages until 7 months (Chong et al., 2018). One possible explanation for the behavior of the baseline model is its access to intonation (i.e. pitch and its contour along an utterance). While input to the baseline model explicitly included the pitch track and its local change, young infants have limited ability to discriminate between languages using intonational cues, as they are not able to use intonation cues to discriminate between English and German until 7 months of age (Chong et al., 2018). If intonation accounts for the behavior of the baseline model, then a version of this model without access to intonation would behave more like newborns.

To test this, we simulated the baseline model again using two sets of features. The first set included only MFCCs without direct pitch information, and the second set included MFCCs and the pitch track. We found that the version with MFCCs behaved like newborns this time, while the version with the pitch track still distinguished between English and German. These results confirm that taking away intonational cues from the model's input leads to more newborn-like performance. In line with this observation, our model likely does not have access to the pitch information at all since we binned the spectral output in 1 kHz-wide bins, which likely caused its newborn-like behavior. This suggest that our model is closer to the representation of human newborns, in which rhythm, not intonation, is primarily used for language discrimination.

Discussion

In this work, we simulated the perception and representation of speech rhythm using a music-inspired model. We found that the model can discriminate the same language pairs as newborns, but only when percussives are present. We also found evidence that our model, similar to young infants, is not sensitive to intonational cues. The model's success at mod-

eling speech rhythm perception, combined with its previous success in capturing musical rhythm, supports a unified representation of rhythm in speech and music. Also, connected with the evidence that percussives represent rhythm in music well (Tsunoo et al., 2009), our results suggest that percussives are relevant for rhythm representation, unlike harmonics.

Our simulations add to the evidence regarding the cues that newborns use to discriminate languages. Newborns can discriminate rhythmically different languages even when the speech is resynthesized in a monotone manner, where all intonation information is lost. As reviewed earlier, newborns are also not sensitive to intonation enough to discriminate between English and German (Chong et al., 2018). Together, this evidence suggests that intonation may be separately represented and acquired from rhythm, with newborns relying on rhythm more than intonation.

While the application of harmonics and percussives to speech processing is new in this project, a similar dichotomy is seen in some previous observations in the literature. In Slater et al. (2017), percussionists and vocalists are found to have better neural encoding for fast-changing acoustics and harmonic structure in speech, respectively. Whereas the neural representation of speech rhythm has generally been associated with the acoustic envelope of speech (e.g., as reviewed by Poeppel & Assaneo, 2020), our study highlights percussives—a cue that is not well captured by the acoustic envelope—as important for rhythm perception. Further research into the relationship between percussives and neural encoding of rhythm may reveal how rhythm is represented in the brain.

Acknowledgments

We thank Bob Slevc and Shihab Shamma for their feedback and comments throughout this project. This research was supported by NSF grants BCS-1734245 and BCS-2120834.

References

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

Carbajal, M. J., Fér, R., & Dupoux, E. (2016). Modeling language discrimination in infants using i-vector representations. In *CogSci.*

Chong, A. J., Vicenik, C., & Sundara, M. (2018). Intonation plays a role in language discrimination by infants. *Infancy*, *23*(6), 795–819.

Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., & Ouellet, P. (2010). Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(4), 788–798.

Fitzgerald, D. (2010). Harmonic/percussive separation using median filtering. In *Proceedings of the International Conference on Digital Audio Effects (DAFx)* (Vol. 13).

Harding, E. E., Sammler, D., Henry, M. J., Large, E. W., & Kotz, S. A. (2019). Cortical tracking of rhythm in music and speech. *NeuroImage*, *185*, 96–101.

Kotz, S. A., Ravignani, A., & Fitch, W. T. (2018). The evolution of rhythm processing. *Trends in Cognitive Sciences*, *22*(10), 896–910.

London, J. (2001). *Rhythm.* Oxford University Press. doi: 10.1093/gmo/9781561592630.article.45963

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

Ney, H. (1984). The use of a one-stage dynamic programming algorithm for connected word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *32*(2), 263–271.

Ono, N., Miyamoto, K., Kameoka, H., Le Roux, J., Uchiyama, Y., Tsunoo, E., ... Sagayama, S. (2010). Harmonic and percussive sound separation and its application to MIR-related tasks. In *Advances in Music Information Retrieval* (pp. 213–236). Springer.

Patel, A. D., & Daniele, J. R. (2003). An empirical comparison of rhythm in language and music. *Cognition*, *87*(1), B35–B45.

Paul, D. B., & Baker, J. (1992). The design for the Wall Street Journal-based CSR corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992* (pp. 357–362).

Poeppel, D., & Assaneo, M. F. (2020). Speech rhythms and their neural foundations. *Nature Reviews Neuroscience*, *21*(6), 322–334.

Schatz, T. (2016). *ABX-discriminability measures and applications* (Unpublished doctoral dissertation). Université Paris 6.

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 *Proc. INTERSPEECH.*

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

Slater, J., Azem, A., Nicol, T., Swedenborg, B., & Kraus, N. (2017). Variations on the theme of musical expertise: Cognitive and sensory processing in percussionists, vocalists and non-musicians. *European Journal of Neuroscience*, *45*(7), 952–963.

Slater, J., Kraus, N., Woodruff Carr, K., Tierney, A., Azem, A., & Ashley, R. (2018). Speech-in-noise perception is linked to rhythm production skills in adult percussionists and non-musicians. *Language, Cognition and Neuroscience*, *33*(6), 710–717.

Torres-Carrasquillo, P. A., Singer, E., Kohler, M. A., Greene,

R. J., Reynolds, D. A., & Deller Jr, J. R. (2002). Approaches to language identification using Gaussian mixture models and shifted delta cepstral features. In *Proc. INTERSPEECH.*

Tsunoo, E., Ono, N., & Sagayama, S. (2009). Rhythm map: Extraction of unit rhythmic patterns and analysis of rhythmic structure from music acoustic signals. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 185–188).