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, Fé, & 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

<table>
<thead>
<tr>
<th>Rhythm Model</th>
<th>Harmonic</th>
<th>Percussive</th>
<th>Natural</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eng. vs. Jap.</td>
<td>48.10%</td>
<td>63.60%</td>
<td>54.45%</td>
</tr>
<tr>
<td>Eng. vs. Ger.</td>
<td>48.15%</td>
<td>50.85%</td>
<td>50.84%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Baseline Model</th>
<th>Full</th>
<th>MFCC</th>
<th>MFCC+pitch</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eng. vs. Jap.</td>
<td>70.20%</td>
<td>57.35%</td>
<td>60.49%</td>
</tr>
<tr>
<td>Eng. vs. Ger.</td>
<td>99.82%</td>
<td>51.80%</td>
<td>65.24%</td>
</tr>
</tbody>
</table>

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.

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


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