Predicting epenthetic vowel quality from acoustics

Adriana Guevara-Rukoz, Erika Parlato-Oliveira, Shi Yu, Yuki Hirose, Sharon Peperkamp, Emmanuel Dupoux

1 École Normale Supérieure - PSL Research University / EHESS / CNRS / INRIA, France
2 School of Medicine, Federal University of Minas Gerais, Brazil
3 Laboratoire de Phonétique et Phonologie, UMR7018 CNRS / Sorbonne-Nouvelle, France
4 Graduate School of Arts and Sciences, The University of Tokyo, Japan

Abstract

Past research has shown that sound sequences not permitted in our native language may be distorted by our perceptual system. A well-documented example is vowel epenthesis, a phenomenon by which listeners hallucinate non-existent vowels within illegal consonantal sequences. As reported in previous work, this occurs for instance in Japanese (JP) and Brazilian Portuguese (BP), languages for which the ‘default’ epenthetic vowels are /u/ and /i/, respectively. In a perceptual experiment, we corroborate the finding that the quality of this illusory vowel is language-dependent, but also that this default choice can be overridden by coarticulatory information present on the consonant cluster. In a second step, we analyse recordings of JP and BP speakers producing ‘epenthesized’ versions of stimuli from the perceptual task. Results reveal that the default vowel corresponds to the vowel with the most reduced acoustic characteristics and whose formants are acoustically closest to formant transitions present in consonantal clusters. Lastly, we model behavioural responses from the perceptual experiment with an exemplar model using dynamic time warping (DTW)-based similarity measures on MFCCs.

1. Introduction

When languages borrow words from one another, the borrowed words tend to be adapted to the local phonology. For instance, Brazilian Portuguese phonotactic constraints disallow most obstruent-obstruent and obstruent-nasal sequences, while those of Japanese disallow consonant clusters and consonants in coda position (with the exception of geminates and nasal consonants). Foreign words containing these illegal sequences may be broken up by the insertion of so-called ‘epenthetic’ vowels (e.g., BP: “football” → /futibol/, JP: “ice cream” → /nisukurimu/). This phenomenon has been shown to also happen during on-line perception: listeners perceive vowels within illegal consonantal sequences [1, 2, 3, 4, 5, 6, 7, 8, 9]. This suggests that phonotactic constraints of the native language play an active role during speech perception and induce repair of illegal forms such that they are recoded into the nearest legal one. The specific mechanisms of this repair process are still largely unknown. In particular, what determines the quality of the epenthesized vowel? Past work has shown that perceptual epenthesis is language-dependent (e.g., /i/ in BP, /u/ in JP), but also that it may be influenced by local acoustic properties, i.e., by coarticulation [7]. Here, we study these two effects together, and report, firstly, on a perception experiment with BP and JP listeners. Next, we conduct acoustic analyses of the production of possible epenthetic vowels in a subset of the same participants. Lastly, we present an exemplar-based computational model of speech perception which attempts to model phonotactic repairs based on acoustics.

2. Perception experiment

We assess patterns of perceptual epenthesis by BP and JP native listeners on stimuli containing an illegal cluster. We investigate (1) the preferred epenthetic vowel in the two languages (/i/ vs. /u/), and (2) the influence of flanking vowels on responses.

2.1. Methods

Fifty-four items with the structure \( V_1 C_1 C_2 V_2 \), with \( V_1 \) and \( V_2 \) vowels from the set \{/a/, /i/, /u/\}, and \( C_1 C_2 \) a cluster from the set \{/bg/, /bn/, /db/, /dg/, /gb/, /gn/\}, e.g. /labgil/, were recorded by a native speaker of French. Twenty-two native BP listeners and 17 native JP listeners were tested in São Paulo and Tokyo, respectively. None had extensive exposure to languages that allow complex consonantal clusters. At each trial, participants heard a stimulus and had to indicate within 3 seconds which vowel from the set \{/a/, /e/, /i/, /o/, /u/ and none\} they perceived within the consonant cluster.

2.2. Results

Statistical analyses were performed with the R statistical software [10], using MCMC glmm [11, 12]. Effects were considered statistically significant if the 95% highest posterior density (HPD) interval estimated for the variable of interest did not include zero. Please note that we only report effects relevant to hypotheses tested in this work. A full report of all analyses conducted in this paper (as well as additional information) can be found in: https://osf.io/zz88w/.

In order to assess the influence of \( V_1 \) and \( V_2 \) (henceforth: flanking vowels) on epenthetic vowel quality (/i/ or /u/), we fitted models with fixed effects Language (BP vs. JP), Number of Same Flanking Vowels (NSFV) (none vs. 1; none and 1 vs. 2) and their interaction, with Participants as random effect. We also included the fixed effect Coronal \( C_1 \) (non-coronal vs. coronal) and the resulting interactions when analysing /u/ responses, as the insertion of default /u/ after coronal consonants yields phonotactically illegal sequences in Japanese. Fixed effects were contrast coded with deviance coding and, in the case of the trinomial variable NSFVs, comparisons were achieved by creating dummy variables “none vs 1” with weights \{-0.5, 0.5, 0\}.
for levels none, 1 and 2, respectively, and “Less than 2 vs. 2” with weights [-0.25, -0.25, 0.5] for levels none, 1 and 2.

Overall, BP and JP participants experienced vowel epenthesis in 81% and 87% of the trials, respectively. We focus our analysis on these trials and, in order to allow for comparisons with the model from Section 4 below, we exclude trials for which the reported epenthetic vowel was /a/ (1%) or /e/ (BP: 0%, JP: 3%). Percentages for the remaining responses of interest (/i/, /o/, and /u/) can be seen in the left-hand part of Table 1.

### 2.2.1. /i/-epenthesis

A main effect of Language shows that BP participants perceived an epenthetic /i/ more often than JP participants (post hoc: −277.1, HPD interval: [−389.2, −167.1]). Moreover, the propensity to respond /i/ was influenced by flanking vowels, as indicated by a main effect of NSFV: Participants gave more /i/ responses when one flanking vowel was /i/ (204.1, [80.8, 283.7]), and even more so when both flanking vowels were /i/ (368.9, [208.4, 443.4]).

### 2.2.2. /u/-epenthesis

We found a main effect of Language; BP participants epenthized /u/ less often than JP participants (265.2, [191.0, 347.2]). The significant main effect of NSFV shows that participants were overall more prone to perceiving an epenthetic /u/ if one (137.0, [90.6, 185.4]) or both (300.3, [230.4, 387.7]) flanking vowels were /u/. Lastly, there was also a main effect of Coronal C1 (−43.0, [−75.8, −9.8]); participants perceived /u/ less often after coronal than after labial and velar consounds. However, neither the interaction between Coronal C1 and Language (−63.9, [−132.1, 3.8]), nor the triple interactions with NSFV (−17.8, [−124.9, 94.1], −14.8, [−298.4, 304.3]) were significant; thus, JP participants were not more prone to avoiding /u/-epenthesis after coronal consonants than BP participants.

### 3. Acoustic analyses

In both BP and JP, the shortest vowel corresponds to the default epenthetic vowel, i.e. /i/ in BP [13] and /u/ in JP [14]. Here, we compare epenthetic vowels /i/, /u/, and /o/ on three acoustic parameters: (1) vowel duration, (2) vowel intensity, and (3) Euclidean distance between vowel formsants and formant transitions in consonantal clusters. We hypothesize that, for both languages, the default vowel is the one (1) that is shortest, (2) that has the lowest intensity, and (3) whose formants are closest to the formant transitions present in consonantal clusters.

#### 3.1. Methods

Seventeen BP and 17 JP participants from the perception experiment were also recorded producing 162 stimuli obtained by crossing the 54 V1C1C2V2 frames of the experimental items with the three vowels /i/, /o/, and /u/ (e.g. /ab/gi/ → /abgi/). Items were read aloud in carrier sentences, with stress and pitch accent on the first syllable for BP and JP speakers, respectively. The recordings were manually segmented and transcribed by a trained phonetician. Recordings with errors or unwanted noise were excluded from the analyses. Acoustic measurements were automatically extracted from the speech signal using the R package wrassp [15].

#### 3.2. Results

For each of the continuous response variables examined in this section, we fitted an MCMC glm with fixed effects Language (BP vs. JP), Medial Vowel (/i/ vs. /u/; /i/ and /u/ vs. /o/), and their interaction, with Participant and Item as random effects. Fixed effects were contrast coded with deviance coding and, in the case of the trinomial variable Medial Vowel, the comparisons were achieved by creating a dummy variable “/i/ vs /u/” with weights [−0.5, 0, 0.5] for levels /i/, /o/ and /u/, respectively, and one for “High Vowels vs. /o/” with weights [−0.25, 0.5, −0.25]. Multiple pairwise comparisons, using Least Squares Means (LSMEANS) and Tukey’s adjustment, were performed using the R package lsmeans [16].

#### 3.2.1. Vowel duration

The measured duration of each medial vowel V3 (in seconds) was log-transformed to account for distribution skewness. We found a main effect of Medial Vowel (“/i/ vs /u/”: 0.04, [0.02, 0.06]); “High Vowels vs. /o/”: 0.32, [0.30, 0.34]), showing that, overall, /o/ is longer than /u/, which is longer than /i/. The interaction of Language and Medial Vowel was also significant (“/i/ vs /u/”: −0.15, [−0.18, −0.10]), reflecting the fact that in BP, /i/ is shorter than /u/ (mean /i/: 57.2 ms; mean /u/: 64.5 ms; adjusted p < 0.05) while in JP, /i/ is shorter than /u/ (mean /i/: 69.6 ms; mean /u/: 72.0 ms; adjusted p < 0.05).

#### 3.2.2. Vowel intensity

We compared the mean intensity of the medial vowels V3 in decibels (dB). There was a main effect of Medial Vowel, with /i/ having on average lower intensity than /u/ (0.8, [0.7, 1.1]), and high vowels having lower intensity than /o/ (2.1, [1.9, 2.4]). Of interest is the fact that the former effect is larger for JP than for BP (Language x “/i/ vs /u/”: 0.38, [0.09, 0.90]), meaning that while /i/ is the vowel with least intensity in BP (mean: 72.8 dB vs. 73.2 for /u/, adjusted p < 0.05), the reverse is not true for JP (mean: 69.7 dB for /u/ vs. 68.7 for /i/, adjusted p < 0.05). This might be due to an overall higher degree of vocal constriction during the production of /i/ compared to /u/.

#### 3.2.3. Vowel formants

We extracted median formant values (F1, F2, and F3, in Bark) from medial vowels V3 and computed their Euclidean distance to the transitions found within their respective clusters (e.g. the /i/ in /abiga/ was compared to transitions in /bg/ from the French recording of /abga/). These Euclidean distances were square-root transformed to account for skewness. There was a main effect of Medial Vowel, as on average distance was shorter for /u/ than for /i/ (−0.06, [−0.08, −0.04]), while it was longer for /o/ relative to both /i/ and /u/ (0.28, [0.25, 0.30]). Of interest is the significant interaction between Language and Medial Vowel (“/i/ vs /u/”: −0.31, [−0.36, −0.28]), reflecting the fact that in BP /i/ formants were closer to cluster transitions than /u/ formants (mean: /i/: 2.8 vs. /u/: 3.1, adjusted p < 0.05), while the reverse held in JP (mean: /i/: 2.9 vs. /u/: 2.2, adjusted p < 0.05).

---

Table 1: Percentage of responses.

<table>
<thead>
<tr>
<th>Human data</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>BP</td>
<td>i</td>
</tr>
<tr>
<td></td>
<td>o</td>
</tr>
<tr>
<td>JP</td>
<td>i</td>
</tr>
<tr>
<td></td>
<td>o</td>
</tr>
</tbody>
</table>
Figure 1: Responses from the perception experiment (left) and model predictions (right), for both BP (top) and JP (bottom). Numbers indicate trial counts, with darker cell backgrounds representing higher values. Within each 3 x 3 grid, trials are separated according to $V_1$ (columns) and $V_2$ (rows). Within each individual rectangle, the horizontal axis relates to whether $C_1$ is coronal (/d/) or not, while the vertical axis corresponds to possible responses. For instance, BP participants experienced /l/-epenthes in in all 78 trials involving /iC_1C_2a/ stimuli for which $C_1$ was not the coronal consonant /d/.

Figure 2: Medial vowels and clusters in F1 x F2 space.

4. Production-based exemplar model

We built an exemplar model of the perception of phonotactically illegal consonant clusters by BP and JP listeners, exclusively based on acoustics. We used all participants’ productions from Section 3 as the inventory of exemplars available to our model. This is a simple way of representing the acoustics that a BP/JP native listener may have been exposed to during language development. As an analogy to the perception experiment from Section 2, the model classified each $V_1C_1C_2$ template as $V_1C_1iC_2V_2$, $V_1C_1oC_2V_2$, or $V_1C_1uC_2V_2$, based on the similarity of the template to exemplars of these three categories available in the inventory. We examined whether the model was able to predict participants’ epenthetic patterns, in particular, whether it was able to mimic preferences for default vowels and capture the modulation of these preferences induced by flanking vowels.

4.1. Methods

Recordings from Section 3 were converted into sequences of 39-dimensional feature vectors consisting of 12 Mel-frequency cepstral coefficients (MFCCs) and energy features, with delta and delta-delta coefficients. Coefficient values were standardised. We computed the optimal alignment between all $V_1C_1C_2V_2$ templates (e.g. /abig/) and their corresponding $V_1C_1iV_2C_2V_2$ epenthesis versions (e.g. /abigi/, /abogii/, /abugi/) using Dynamic Time Warping (DTW) [17, 18]. In order to ensure that the resulting distances were not mainly influenced by spectral differences of flanking vowels $V_1$ and $V_3$, we only compared $C_1C_2$ clusters to $C_1V_3C_2$ sections. Note, however, that coarticulation cues from flanking vowels are expected to be present within the clusters.

For the simulation, we built a classifier that assigns any given template to one category in the set $\{V_1C_1iC_2V_2, V_1C_1oC_2V_2, V_1C_1uC_2V_2\}$, based on acoustic similarity. Similarity $s$ between templates and epenthized versions was defined as

$$s = e^{-cd}$$

where $d$ is the DTW distance, and $c$ is a parameter determining the weight of the DTW distance on classification [19]. When
Less than none in BP compared to JP participants, as well as the influence of able to reflect the higher frequency of /i/ as epenthetic vowel on epenthetic vowel quality for /u/, it was unable to reflect the (C and a main effect of NSFV (none = 2

so that /u/ more often than /o/ in coronal contexts, even though both /du/ and /di/ are phonotactically illegal sequences in JP. These results, which are reminiscent of previous work [6, 8], suggest that constraints on perception given by surface phonotactics can be overruled by constraints relative to matching input acoustics [7]. In fact, if this were not the case, novel sound sequences would have never arisen in JP loanwords (e.g. party is adapted as /pacti/, not /pacti/).

Finally, we presented one exemplar model per language, based on productions by BP and JP participants, respectively. These models reproduced the major effects found in the perception experiment — namely, the preferred epenthetic vowel, and some significant effects of language and flanking vowels — although with a high level of noise. This noise level may be due to the relatively low number of tokens that were available as exemplars, the fact that the DTW procedure removes temporal cues (recall that we found that default vowels tend to be of shorter duration), and/or the fact that MFCC features do not appropriately capture speaker invariance. We interpret the results as providing a proof of principle that some of the salient effects regarding perceptual epenthesis can be accounted for on purely acoustic grounds. Future research is needed to improve on the model, whose predictions deviated from the perceptual data on several counts (e.g., 6% /o/-epenthesis for BP, but less than 1% for JP; failure to produce more /i/-epenthesis for JP than BP). These improvements could involve more phonetically and/or temporally informed features (e.g., spectrotemporal representations [22]), state-of-the-art large-scale approaches with HMM or DNN systems, or physiologically-inspired models of speech perception (e.g., based on cortical oscillations [23]).

To conclude, a triple approach combining perception experiments, acoustic analyses, and modeling allows us to gain insight into the mechanisms underlying perceptual epenthesis, and, more generally, repairs of illegal phonological structure during speech perception.

6. Acknowledgements

This research was funded by the European Research Council (ERC-2011-AdG-295810 BOOTHON). It was also supported by the Agence Nationale pour la Recherche (ANR-10-LABX-0087 IEC, ANR-10-IDEX-0001-02 PSL*). We thank Ewan Dunbar, Page Piccinini, and Neil Zeghidour for discussion.
7. References


