Decoding perceptual vowel epenthesis: Experiments & Modelling

Institution Name

AGR

Day Month Year

Abstract

Abstract goes here

Résumé

Abstract goes ici

Allá abajo les ha llovido —aquella nube fugaz que veló el prado verde con sus hilos de oro y plata, en los que tembló, como en una lira de llanto, el arco iris—. Y sobre la empapada lana del asnucho, las campanillas mojadas gotean todavía.

¡Idilio fresco, alegre, sentimental! ¡Hasta el rebuzno de Platero se hace tierno bajo la dulce carga llovida! De cuando en cuando, vuelve la cabeza y arranca las flores a que su bocota alcanza. Las campanillas, níveas y gualdas, le cuelgan, un momento, entre el blanco babear verdoso y luego se le van a la barrigota cinchada.

¡Quién, como tú, Platero, pudiera comer flores..., y que no le hicieran daño!

¡Tarde equívoca de abril!... Los ojos brillantes y vivos de Platero copian toda la hora de sol y lluvia, en cuyo ocaso, sobre el campo de San Juan, se ve llover, deshilachada, otra nube rosa.

Idilio de Abril. Platero y yo.
Juan Ramón Jiménez

Acknowledgements

I want to thank...

Publications

Journal articles

- **Guevara-Rukoz, A.**, Cristia, A., Ludusan, B., Thiollière, R., Martin, A., Mazuka, R., & Dupoux, E. (2018).
  Are words easier to learn from infant- than adult-directed speech? A quantitative corpus-based investigation.
  Cognitive Science. https://doi.org/10.1111/cogs.12616

- Fort, M., Lammertink, I., Peperkamp, S., **Guevara-Rukoz, A.**, Fikkert, P., & Tsuji, S. (2018).
  SymBouki: a meta-analysis on the emergence of sound symbolism in early language acquisition.
  Developmental Science. https://doi.org/10.1111/desc.12659

- **Guevara-Rukoz, A.**, Lin, I., Morii, M., Minagawa, Y., Dupoux, E., & Peperkamp, S. (2017).
  Which epenthetic vowel? Phonetic categories versus acoustic detail in perceptual vowel epenthesis.
  The Journal of the Acoustical Society of America, 142(2), EL211-EL217.

Conference proceedings

- **Guevara-Rukoz, A.**, Parlato-Oliveira, E., Yu, S., Hirose, Y., Peperkamp, S., and Dupoux, E. (2017).
  Predicting epenthetic vowel quality from acoustics.
  Proceedings of Interspeech, 596-600.

Contents

1 Introduction **1**
- 1.1 Vowel epenthesis in loanword adaptations . . . 2
- 1.2 Perceptual vowel epenthesis . . . 4
- 1.3 Processing steps in perceptual vowel epenthesis . . . 5
- 1.4 Justifying the modelling approach . . . 6
- 1.5 Outline of this thesis . . . 7

2 Role of acoustic details in the choice of epenthetic vowel quality **8**
- 2.1 Introduction . . . 8
  - 2.1.1 One-step vs two-step theories . . . 8
  - 2.1.2 Role of acoustics . . . 9
  - 2.1.3 Chapter preview . . . 10
- 2.2 Which epenthetic vowel? Phonetic categories versus acoustic detail in perceptual vowel epenthesis 11
  - 2.2.1 Introduction . . . 11
  - 2.2.2 Methods . . . 12
  - 2.2.3 Results . . . 13
  - 2.2.4 Discussion and conclusion . . . 16
  - 2.2.5 Annexes . . . 18
- 2.3 Predicting epenthetic vowel quality from acoustics . . . 24
  - 2.3.1 Introduction . . . 25
  - 2.3.2 Perception experiment . . . 25
  - 2.3.3 Acoustic analyses . . . 26
  - 2.3.4 Production-based exemplar model . . . 29
  - 2.3.5 Discussion . . . 33
- 2.4 Predicting epenthetic vowel quality from acoustics II: It's about time! 34
  - 2.4.1 Introduction . . . 34
  - 2.4.2 Methods . . . 35
  - 2.4.3 Results . . . 38
  - 2.4.4 Discussion . . . 42
- 2.5 Conclusions . . . 44

3 Modelling speech perception with ASR systems **46**
- 3.1 Introduction . . . 46
- 3.2 Anatomy of our HMM-based speech recogniser . . . 46
  - 3.2.1 Corpora . . . 47
  - 3.2.2 Features . . . 49
  - 3.2.3 Acoustic model . . . 50

3.2.4 Lexicon & language models . . . 52
3.2.5 Decoding . . . 52
3.2.6 Scoring: Assessing native performance . . . 53
3.3 Investigating surface phonotactics . . . 55
3.3.1 Introduction . . . 55
3.3.2 Experiment 1 . . . 55
3.3.3 Experiment 2 . . . 64
3.3.4 Discussion . . . 68
3.4 Investigating acoustic/phonetic match . . . 69
3.4.1 Introduction . . . 69
3.4.2 Experiment 1: Phonological alternations . . . 69
3.4.3 Experiment 2: Allophony . . . 71
3.4.4 General discussion . . . 73
3.5 Conclusions . . . 73

4 General Discussion **74**

A Are words easier to learn from infant- than adult-directed speech? **79**

Chapter 1

Introduction

Beginning from the time spent in her mother's womb, a baby is exposed to sounds and, importantly, to speech. Initially a "universal listener", in her early months she is initially able to discriminate a wide variety of sounds [CITE]. However, as she is exposed to only one or a few languages that are spoken around her, she progressively loses the "universal listener" title. Indeed, through this exposure, her perceptual system progressively becomes attuned to sounds and structures useful for decoding what will progressively become her native language(s) [CITE Werker].

That is, at around 10 months of age, the infant's perceptual system shows signs of becoming optimised for receiving and processing native input. On one hand, a native phonemic inventory is established. Phonemes are generally defined as the smallest phonetic units capable of conveying a lexical distinction in a language. It is therefore essential for the infant to be able to notice when two different acoustic signals correspond to the same or different phonemes, in order to properly access lexical meaning. For this to happen, her perceptual system partitions the acoustic space into areas corresponding to the phonemes relevant for her native language [CITE].

On the other hand, the infant also becomes aware of which phoneme combinations occur in the language and which ones do not [CITE Saffran, ...]. In particular, she progressively acquires the phonotactics for her native language, namely, the constraints determining what constitutes well-formed native sound combinations [CITE: Jusczyk, Pisoni?...]. This is useful, for instance, when trying to find word boundaries in fluent speech, as it is more probable for a rare phoneme combination to occur between words than within words [CITE].

Through the optimisation process that results in the acquisition of the native-phoneme inventory and native phonotactics, the infant's perceptual system becomes specialised in her native language, allowing her to better tackle problemes such as word learning.

A few years later, the infant is now an adult with an attuned perceptual system. And while this perceptual attunement allows her to communicate efficiently in her native language, issues arise when she now attempts to acquire nonnative languages. Nonnative speech is processed by her perceptual system, which has been optimised for her native language. The nonnative speech is therefore processed according to her native segmental and suprasegmental inventories, and phonotactics, which do not match those intended by the speech source. For our adult listener this will often result in nonnative speech being misperceived; phonemes and prosodic elements

might be replaced (assimilation) or deleted (ellipsis). Additional phonemes might even be inserted (epenthesis).

It seems to be generally accepted in the nonnative speech perception literature that misperceptions of nonnative speech result from minimal modifications of the original speech. However, there are multiple proposals relating to the exact nature of these modifications. For instance, are these phonetically-minimal? Phonologically-minimal? Is it a multi-step process?

It is my belief that the various proposals put forward must be formally defined in order to be tested empirically. In this dissertation, I select one of the proposals advanced by the psycholinguistics literature, namely the perception-based reverse inference models of nonnative speech perception, and test a proof-of-concept computational implementations. I present various methodologies for qualitatively and quantitatively evaluating the reverse inference proposal, focusing on the phenomenon of perceptual vowel epenthesis. Following on [CITE peperkamp], the data arising from the computational models is compared to data from psycholinguistics experiments. In these, nonnative speech perception is evaluated using psycholinguistics paradigms which tap onto online (i.e., real-time, individual) perception of nonwords, in order to reduce the influence of confounds such as orthography and semantics. In other words, I subject the proposed computational models to tasks analogous to those completed by human participants and analyse their behaviour both quantitatively and qualitatively. Do we find perception-based mechanisms to be necessary to predict perceptual vowel epenthesis? If so, do they suffice?

In order to investigate the underlying mechanisms of nonnative speech perception, we are required to refer to work in two relevant fields, loanword adaptations and nonnative misperceptions. The former focuses on how words from a source language are modified when introduced to a borrowing language. Rather embedded in the field of theoretical linguistics, this literature offers an indirect window to perception, as loanwords are the product of various complex language transmission phenomena. On the other hand, the experimentally inclined literature of psycholinguistics on nonnative speech perception offers more direct data for undestanding misperceptions. However, data may be less readily available than for loanwords. I will now present both literatures, focusing on various mechanisms that have been proposed to explain the phenomenon of vowel epenthesis within both fields.

1.1 Vowel epenthesis in loanword adaptations

The phenomenon of vowel epenthesis was first studied in the context of loanword adaptations, namely the modification of words from a source language when they are introduced to a borrowing language. Consider the bold vowels in the following examples:

- English “strike” /stɹaɪk/ → Japanese /s**u**t**o**raik**u**/
- French “baguette” /bagɛt/ → Japanese /baget**ːo**/
- English “snob” /snɒb/ → Spanish /**e**snob/

In this context, vowel epenthesis consists in the insertion of a vowel in the (borrowing) surface form of a word that did not contain said vowel in the underlying

representation in the source language. Epenthesis often occurs when the introduced word does not respect the phonotactics of the borrowing language (e.g., illegal consonant clusters, illegal syllabification). The foreign word is imported to the borrowing language in a diachronic process involving multiple individuals and possibly various methods of word transmission (e.g., through written materials, orally, etc). Additionally, adaptations can be influenced by orthography, when available [CITE Daland 2015; Ito?]. The assumption being that words are introduced by highly proficient speakers of the nonnative language that are, therefore, able to access a faithful underlying representation of the source word.

Concerning the question of how the nonnative source word is transformed into the adapted loanword, authors from the loanword literature advance the hypothesis that the underlying mechanisms are phonological and abstract in nature.

For instance, in a grammar-based view, the adapted loanword corresponds to the best match to the source word after candidate adaptations are passed through a grammatical filter [CITE: Hyman (1970), Lovins (1975), Yip (1993), Jacobs & Gussenhoven (2000), Shinohara (2004)]. In this case, the grammar comprises faithfulness and markedness constraints arranged in a certain order, with some authors suggesting that some modifications might be more degrading that others [Steriade, 2001]. How constraints are arranged may or may not be compatible with how native sets of constraints are ordered. It is assumed that the underlying representation of the source word is accessible and is used in the derivation of the adapted loanword. As such, the faithful underlying representation only becomes adapted when the grammar computes a surface representation from said underlying form (e.g., during production).

Another similar mechanism was proposed by [CITE La Charité & Paradis (2005)], where adaptations are based on minimal featural changes. The source word is adapted by highly proficient bilinguals, who can manipulate and compare the phonological systems of both the source and the borrowing languages. In this case, adaptation consists in these highly proficient bilinguals choosing the modifications of the source word that result in the least featural changes (e.g., voicing) between the source adaptation and the resulting loanword.

In many proposals, the role of perception has been assumed to be minimal, or at least secondary, in loanword adaptation [CITE e.g., LaChParadis1997?]. However, this view is far from being a consensus. For instance, as briefly mentioned earlier, [Steriade, 2001] proposed that some grammar-based modifications resulted in more perceptually-salient modifications of the source word, resulting in a greater amounts of degradation. Also, [CITE silverman1992] proposed his multiple scansion theory of loanword adaptation. In this theory, the underlying representation in the borrowing language is retrieved from a phonetic, possibly acoustic, form of the source word. In a first step, a perceptual-level representation is established, where the underlying representation is built with preliminary segmental and prosodic structure. It is only at a second step, at the operational level, that the phonology of the borrowing language is applied to the word (e.g., through a grammatical filter), resulting in modifications such as vowel epenthesis, when necessary. For similar proposals, see [CITE Kenstowicz 2001/2003, KenstowiczSuchato, Yip1993/2006].

Alternative structure: Change to (1) Purely phonological and (2) Perceptual

1.2 Perceptual vowel epenthesis

In the late nineties various theories of nonnative speech perception [CITE Best1995, Kuhl1995] and second language phoneme acquisition [Flege1995] surfaced in the psycholinguistics literature. There seemed to be an overlap between the observed nonnative speech misperceptions and patterns observed in loanwords; may perception directly account for loanword adaptation? It is in this context that appeared proposals such as that of [CITE PeperkampDupoux2003], which put greater emphasis on the role of misperception on loanword adaptation. The proposal being that, when perceiving nonnative input, a phonetic decoding module takes into account segmental, suprasegmental, and syllabic inventories of the native language in order to derive a phonetically-minimally modified representation[1] of the acoustic input. It is this misperceived version that is then converted into an underlying representation by a phonological decoding module; this cemented underlying representation is unfaithful with respect to the original underlying representation in the nonnative language. In this context, when listeners experience perceptual vowel epenthesis, they hallucinate vowels not initially present in the nonnative speech. In the cases studied in this work, this happens as a way to break phonotactically illegal clusters. For instance, in Japanese most consonant clusters, such as /bz/, are phonotactically illegal. When hearing nonwords containing these clusters, such as /ebzo/, Japanese listeners may hallucinate an epenthetic /u/[2] within the cluster, yielding /ebuzo/ as the percept [Dupoux et al., 1999]. Epenthesis of /u/ by Japanese listeners is not only evident in their behaviour but also in their brain responses; they have difficulties differentiating the clusters produced by a French speaker from their epenthesized counterparts (e.g., /ebzo/ *vs.* /ebuzo/) while also showing different event-related potentials compared to native French speakers [Dehaene-Lambertz et al., 2000]. Importantly, experimental data suggests that epenthesis is a pre-lexical process happening early in speech perception [Dupoux et al., 2001]. Epenthesis has also been attested in languages other than Japanese [cite: Dupoux (1999), Dehaene (2001), Dupoux (2001), Monahan (2009), Dupoux (2011), Mattingley (2015)]; it has been studied in other languages such as Korean [cite: de Jong & Park (2012), Durvasula (2015), Shin & Iverson (2011)], Brazilian Portuguese [cite: Dupoux (2011)], Spanish [cite: Hallé (2014)], English [cite: Berent (multi), Davidson], and Mandarin Chinese [cite: Durvasula (2018)].

While research on perceptual vowel epenthesis uses the literature on loanword adaptation as a a source of inspiration and a source of informed predictions, it is importqnt to note that the phenomenon of vowel epenthesis is not defined equally in both fields. Remember that loanword adaptation is a diachronic process, involving complex interactions between several groups of individuals, with varying degrees of source language fluency. In contrast, the psycholinguistic approach focuses on online perception and, while data from multiple participants is collected, the modifications observed on the output of perception are assumed to occur at the level of the individual, as there is no interaction between participants. Participants are not expect to be proficient in the nonnative language, which minimises the influence of

[1] Whether this representation is acoustic, articulatory [CITE cf Berent2015], and/or gestural [CITE: BrowmanGoldstein, Best1995, BestTyler2007] in nature is not discussed here.

[2] A more faithful phonetic transcription of the vowel is [ɯ] but following previous work the phonological notation /u/ will be used in the remainder of the thesis.

linguistic knowledge on misperceptions (e.g., orthography).

1.3 Processing steps in perceptual vowel epenthesis

From now on we will focus on perceptual vowel epenthesis, referring to loanwords as a source of inspiration for experimental setups and as a source of informed predictions. Concerning the process of vowel epenthesis, we can identify two types of proposed pipelines that differ in the amount of processing steps that the nonnative input is subjected to during perception: these are two-step and one-step theories of nonnative speech perception, illustrated in Figure . While these names are somewhat transparent, we will now explain in more detail the differences between the two types of proposals.

Reminiscent of Silverman's multiple scansion theory for loanword adaptations [**?**], two-step theories of nonnative speech perception divide the perception process in two stages. According to these proposals, the quality of the epenthetic vowel is determined by a language-specific grammar after an initial parsing of the nonnative input. For [Berent et al., 2007], the identity of the segments present in the nonnative input is retrieved in an initial step, yielding a *phonetic form*. The native grammar then assesses the phonotactic legality of this phonetic form in a second step. If a phonotactic violation is found, the grammar, which combines both language-specific and universal components, repairs the phonetic form by inserting a vowel. The output of this final step is the *phonological representation*. Another proposal, that of [Monahan et al., 2009] also consists in two steps, but with some differences. During the first step the identity of the segments in the input is retrieved and segments are grouped into syllables, following native phonotactics. Some syllables will contain indeterminate segments (e.g., /ebzo/ will have been parsed as /e.bV.zo/). In a second step, the quality of the indeterminate segments, in this case the epenthetic vowel, is chosen amongst vowels that are of low sonority and can undergo devoicing. The quality of the vowel might not be determined if an optimal match is not found. The two proposals that we have summarised share the fact that the categorisation of the segments that are not the epenthetic vowel occurs in a first step and it is not modified during the second step, where the identity of the epenthetic vowel is determined.

In contrast, with one-step proposals, authors such as [Dupoux et al., 2011] and [Wilson and Davidson, 2013] argue that the identity of the epenthetic vowel is determined in the process of parsing the input, simultaneously to the categorisation of all other segments. The phonotactic legality of the input is therefore assessed at the same time as the categorisation happens. Notably, the input is not processed as a linear sequence of sounds; syllabic structure is taken into account during the parsing process [Kabak and Idsardi, 2007].

[Wilson and Davidson, 2013] qualify the process as a process of reverse inference within a Bayesian framework, where the perceptual system computes $P(w|X)$ the posterior probability of candidate percepts w given the auditory input X. These are estimated, for each candidate percept, from the product of $P(X|w)$ the likelihood of the acoustics given the percept and $P(w)$ the prior probability of the percept, defined as its phonotactic acceptability. Mathematically, this can be formulated as

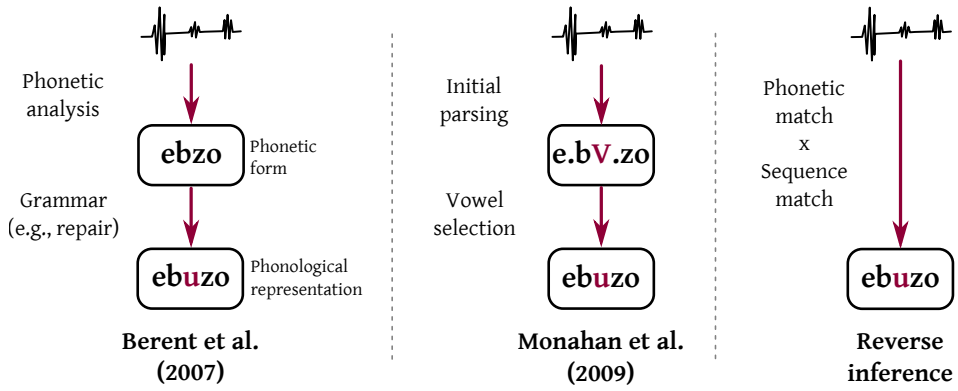

Figure 1.1: *Processing of the nonnative stimulus /ebzo/ by Japanese listeners, according to two-step and one-step proposals for perceptual vowel epenthesis.*

in equation 1.1. Then, in a maximum *a posteriori* (MAP) estimation scenario, the final percept $\widehat{w}$ corresponds to the percept with the highest posterior probability, as shown in equation 1.2. Alternatively, the final percept may be estimated by weighted sampling, where weights are defined by the posterior probabilities.

$$P(w|X) \propto P(X|w) \cdot P(w) \tag{1.1}$$

$$\widehat{w} = \underset{w}{arg\,max} \{P(X|w) \cdot P(w)\} \tag{1.2}$$

In other words, for one-step models, parsing becomes an optimisation problem where the optimal output is the one maximising the acoustic match to the input and the likelihood of the phonemic sequence in the native language. [Durvasula and Kahng, 2015] adds to the aforementioned proposals by suggesting that listeners are decoding nonnative speech through a process of reverse inference that not only optimises the output according to phonetic representations and surface phonotactics, but also according to native phonological alternations (i.e., mappings between underlying and surface representations).

1.4 Justifying the modelling approach

> ... But why make models?
>
> *Derek Zoolander, probably.*

In this thesis I will evaluate computational implementations of one-step theories of nonnative speech perception. Notably, I will investigate models from the field of automatic speech recognition (ASR) which are a direct implementation of the Bayesian model shown in equation 1.1[3]. While only the one-step family of theories will be evaluated in this preliminary work, I encourage further research to be

[3]The equivalent of a weighted sampling procedure was preferred over MAP estimation for percept selection, since participant responses in previous experimental work on epenthesis tended to show variation and were not deterministic.

done with similar methodologies in order to investigate all of the various co-existing proposals.

Indeed, using computational models in order to investigate competing theories is beneficial in several ways. Firstly, the need to translate the theories into model implementations forces model ideators to provide a mathematically and/or algorithmically well defined model. This is in contrast to more vague and ambiguous verbally defined theories that leave more espace to reader interpretations. Having more rigorous model definitions also allows to better understand competing theories (what is the exact nature of the input? which grammar constraints are applied and how? ...), meaning that it is easier to compare proposals and see where they differ significantly or not.

Secondly, obtaining a computational implementation of a theory means that it is then possible to derive predictions from the models in question. It is then possible to qualitatively and quantitatively examine these predictions, and compare them to what is observed in behavioural data.

1.5 Outline of this thesis

[TO-DO while writing discussion]

Chapter 2

Role of acoustic details in the choice of epenthetic vowel quality

2.1 Introduction

As presented in the Introduction, work on loanword adaptation and online speech perception shows that listeners epenthesize or delete vowels from nonnative input when it does not conform to native non-native phonotactics. While this statement seems to be generally accepted, the mechanisms underlying these phenomena are subject to more debate. In this chapter we will investigate the mechanisms underlying variations of epenthetic vowel quality.

2.1.1 One-step vs two-step theories

We saw that theories such as those by [Berent et al., 2007, Monahan et al., 2009] view perceptual vowel epenthesis as a two-step process. According to these proposals, the quality of the epenthetic vowel is determined by a language-specific grammar after an initial parsing of the nonnative input. In contrast, one-step theories such as those proposed by [Dupoux et al., 2011, Wilson and Davidson, 2013, Durvasula and Kahng, 2015] argue that parsing is an optimisation problem where the optimal output maximises the acoustic/phonetic match to the input and the likelihood of the phonemic sequence in the native language.

How can we confront and test these one-step and two-step proposals? For this, we can dissect the phenomenon of perceptual vowel epenthesis and split it into two subproblems:

1. When does epenthesis occur?

2. What vowel is epenthesized?

Concerning the first subproblem, neither one-step theories nor two-step theories give explicit predictions concerning the rate of epenthesis. It is even unclear if the two-step theories exposed above allow for epenthesis to *not* happen. In the case of [Berent et al., 2007], not epenthesizing a vowel would require directly yielding the phonetic form, without repairs being performed by the grammar. While [Berent et al., 2007] hypothesizes that this may happen in tasks requiring participants to pay more attention to phonetics, it is unclear in which cases listeners

would directly retrieve the phonetic form within a same task, for similar stimuli. In the case of [Monahan et al., 2009], lack of epenthesis would involve a different syllabification of the input than when epenthesis happens. Therefore, a priori, epenthesis should always happen if the input is syllabified according to native phonotactics. In the case of reverse inference one-step theories [Dupoux et al., 2011, Wilson and Davidson, 2013, Durvasula and Kahng, 2015], lack of epenthesis might occur if the optimal match between the nonnative input and the native output is more strongly driven by acoustic/phonetic match than by sequence acceptability.

We now turn to the second subproblem, epenthetic vowel quality. For a given phonemic sequence containing a phonotactic violation, two-step theories would predict that epenthetic vowel quality is determined after an initial categorisation step. As such, we do not expect different tokens of a same type to yield epenthetic vowels of different quality. On the other hand, this would be possible for one-step accounts, since the acoustic details are included in the computation of the optimal output. Examining modulations in epenthetic vowel quality, therefore, allows to empirically tease apart one-step and two-step theories summarised above.

2.1.2 Role of acoustics

Results by [Dupoux et al., 2011] support one-step theories, since the authors were able to modulate the identity of the epenthetic vowel perceived by Japanese and Brazilian Portuguese listeners from stimuli that had the exact same segmental structure. For instance, Japanese listeners could be let to epenthesize /i/ more often instead of their default /u/ within consonant clusters (and vice versa for Brazilian Portuguese listeners). What are the factors determining whether participants more readily epenthesized /i/ or /u/?

The modulations in epenthetic vowel quality observed in [Dupoux et al., 2011] were due to acoustic cues present in the stimuli. Stimuli with the same segmental sequence were constructed by excising the medial vowel from /ebuzo/ and /ebizo/, yielding /eb(u)zo/ and /eb(i)zo/. These items differ in the coarticulation cues remaining in the consonants, but they have identical segmental structure (/ebzo/). Participants epenthesized /i/ more often from /eb(i)zo/ than from /eb(u)zo/, and similarly for /u/.

Remember that, for the two-step proposals above, the quality of the epenthetic vowel is determined in a second step, after the identity of the segments has been blocked. Both /eb(u)zo/ and /eb(i)zo/ would have /ebzo/ as a phonetic form (following [Berent et al., 2007]) or they would be parsed as /e.bV.zo/ (following [Monahan et al., 2009]). We cannot predict the modulations in epenthetic vowel from these initial parsings. However, in a one-step processing the acoustic details (in this case, coarticulation) could be taken into account when computing the acoustic match between the input and possible output phoneme sequences.

In this view, the representation used as input for the computation is acoustic in nature. This is in contrast to proposals of input as featural representation (e.g., binary, geometric). For instance, it has been hypothesized that the phenomena of epenthetic vowel copy (i.e., when the epenthetic vowel shares quality with neighbouring vowels) is due to a transfer of phonological features from neighbouring vowels and/or consonants towards an undeterminate epenthetic vowel [Rose and Demuth, 2006, Uffmann, 2006]. These phonological explanations of epenthetic

vowel quality would therefore predict that, in auditory stimuli where the quality of the coarticulation and the quality of neighbouring vowels would be in conflict, the quality of the epenthetic vowel would mostly be determined by the neighbouring vowels.

2.1.3 Chapter preview

In this chapter we will investigate perceptual vowel epenthesis in order to tackle two main questions:

- How does the influence of acoustic details on epenthetic vowel quality compare to other influences such as those of an abstract grammar?

- Stimuli in [Dupoux et al., 2011] were made by excising vowels. Can we reproduce the modulations of epenthetic vowel quality caused by coarticulation using naturally produced stimuli?

- And finally, if acoustic factors are essential when choosing epenthetic vowel quality, does this mean that they are sufficient to do so?

In **section 2.2** a perceptual experiment aims at disentangling the contributions of phonetic categories and acoustic details on epenthetic vowel quality. Participants are asked to report their choice of epenthetic vowel (if any) within consonant clusters in stimuli where the acoustic information contained in the cluster may be in disagreement with the identity of neighbouring vowels. Information theoretic measures allow us to quantify the influence of both neighbouring phonetic categories and acoustic details.

In **section 2.3** we explore the possibility of predicting epenthetic vowel quality in Brazilian Portuguese (BP) and Japanese (JP) using a production-based exemplar model of perception. This type of model predicts the quality of a vowel epenthesized within the cluster of a stimulus based solely on the acoustic similarity of said /CC/ cluster to /CVC/ exemplars produced by native speakers of BP or JP. From this modelling approach we can evaluate the influence of pure acoustics on effects such as default epenthetic vowel quality and modulations induced by neighbouring vowels, with naturally produced stimuli that have not been manipulated.

In **section 2.4** we modify the production-based exemplar models from section 2.3. Several modifications are applied, mostly based on increased performance during the parameter optimisation phase. However, a notable modification is the normalisation of features by speaker (our input is still acoustic in nature, but it is closer to phonetics than previously). Also, we include in our models the possibility to add a duration-mismatch penalty, based on the finding that default epenthetic vowels are also those that are shorter. We examine the effect of the presence or absence of the duration penalty on default epenthetic vowel choice and modulations of epenthetic vowel quality by neighbouring vowels.

2.2 Which epenthetic vowel? Phonetic categories versus acoustic detail in perceptual vowel epenthesis

The following section is a modified version of the following journal article:
Guevara-Rukoz, A., Lin, I., Morii, M., Minagawa, Y., Dupoux, E., & Peperkamp, S. (2017). Which epenthetic vowel? Phonetic categories versus acoustic detail in perceptual vowel epenthesis. Journal of the Acoustical Society of America, 142(2), EL211-EL217.
Stimuli were designed and recorded by I. Lin. Experimental data for the identification task were collected by M. Morii and Y. Minagawa, using scripts by A. Guevara-Rukoz. The ABX experiment was designed and run by I. Lin. Statistical analyses, phonetic transcriptions, and acoustical analyses were performed by A. Guevara-Rukoz. The initial manuscript draft was prepared by E. Dupoux, S. Peperkamp, and A. Guevara-Rukoz. E. Dupoux and S. Peperkamp supervised the entirety of the study.
Modifications with respect to the original paper: additional figures, annexes.

Abstract This study aims to quantify the relative contributions of phonetic categories and acoustic detail on phonotactically-induced perceptual vowel epenthesis in Japanese listeners. A vowel identification task tested whether a vowel was perceived within illegal consonant clusters and, if so, which vowel was heard. Cross-spliced stimuli were used in which vowel coarticulation present in the cluster did not match the quality of the flanking vowel. Two clusters were used, /hp/ and /kp/, the former containing larger amounts of resonances of the preceding vowel. While both flanking vowel and coarticulation influenced vowel quality, the influence of coarticulation was larger, especially for /hp/.

2.2.1 Introduction

Our auditory perceptual system is tuned to the sound system of our native language, resulting in impoverished perception of nonnative sounds and sound sequences [Sebastián-Gallés, 2005]. For instance, in Japanese, a vowel can only be followed by a moraic nasal consonant or by a geminate consonant. As a consequence, Japanese listeners tend to perceive an illusory, epenthetic, /u/ within illegal consonant clusters [Dupoux et al., 1999, Dehaene-Lambertz et al., 2000, Dupoux et al., 2001, Monahan et al., 2009, Dupoux et al., 2011, Guevara-Rukoz et al., 2017] and it is evident in loanword adaptation as well (e.g. the word 'sphynx' is borrowed in Japanese as /sufiNkusu/). Similar effects have been documented in other languages, with different epenthetic vowels (/ɨ/ in Korean [Kabak and Idsardi, 2007, Berent et al., 2008, de Jong and Park, 2012]; schwa in English [Berent et al., 2007, Davidson and Shaw, 2012]; /i/ in Brazilian Portuguese [Dupoux et al., 2011, Guevara-Rukoz et al., 2017]; and /e/ in Spanish [Hallé et al., 2014]). Even within languages, there sometimes is variation in the quality of the epenthetic vowel; for instance, in Japanese, the epenthetic vowel can in certain contexts be /i/ or /o/ [Mattingley et al., 2015, Guevara-Rukoz et al., 2017].

The factors that determine the quality of the epenthetic vowel are still unclear. There is evidence that local acoustic cues in the form of vowel coarticulation play a role. Specifically, using artificial consonant clusters obtained by completely removing an inter-consonantal vowel, [Dupoux et al., 2011] found that the quality of the removed vowel – traces of which are present in the neighboring consonants – influences the quality of the epenthetic vowel. Other studies, however, have argued for an influence of phonological factors, such as the legality of the resulting repair at the phonotactic level [Mattingley et al., 2015] or the presence of phonological alternations in the language [Durvasula and Kahng, 20

Determining the source of epenthetic vowel quality is important at a theoretical level, because it can shed light on the computational mechanisms underlying the perception of speech sounds. For instance, [Dupoux et al., 2011] argued that coarticulation effects cannot be accounted for by two-step models, in which the repair of illegal sequences follows that of phoneme categorization, while they are in accordance with one-step models, in which phoneme categorization takes phonotactic probabilities into account. [1] However, [Dupoux et al., 2011] only assessed the presence of acoustic effects, without investigating a possible role of categorical effects. Here, our aim is to quantify the relative contributions of categorical and acoustic effects on epenthetic vowel quality by directly comparing these two types of effect.

We focus on perceptual vowel epenthesis following /h/. This case is ideally suited for our objective as in Japanese loanwords these fricatives are typically adapted by adding a 'copy' of the preceding vowel when they occur in a syllable coda. For instance, *'Bach'*, *'(van) Gogh'*, and *'Ich-Roman'* are adapted as /bahːa/, /gohːo/, and /ihːiroman/. In work on loanword adaptations, cases of vowel copy in epenthesis have been explained as a result of the spreading of phonological features from the preceding vowel onto the epenthetic vowel (i.e., vowel harmony), for instance in Shona, Sranan, and Samoan [Uffmann, 2006], and Sesotho [Rose and Demuth, 2006]. In speech perception, however, this pattern could be based either on phonetic categories, i.e. the preceding vowel itself, or on acoustic detail, i.e. traces of this vowel that are present in /h/, as laryngeal fricatives such as /h/ contain acoustic information relative to formants of surrounding vowels [Keating, 1988]. Using an identification task, we tease apart these two explanations by independently manipulating the categorical context in which /h/ occurs and the acoustic realization of this segment, using cross-splicing. As a control, we also use stimuli with /k/, which are expected to give rise to more default /u/-epenthesis because they contain less coarticulation.

2.2.2 Methods

2.2.2.1 *Participants*

Twenty-five native Japanese speakers were tested in Tokyo, Japan (mean age 24 ± 3.5; 13 female). All were students at Keio University, and none had lived abroad.

2.2.2.2 *Stimuli*

We constructed a set of 20 base items, 10 disyllabic ones of the form $V_1C_1C_2V_1$ and 10 matched trisyllabic ones of the form $V_1C_1V_1C_2V_1$, with V_1 a vowel in the set /a, e, i, o, u/ (henceforth: flanking vowel), C_1 /h/ or /k/, and C_2 a fixed consonant, /p/, e.g. /ahpa/, /ekpe/, /ohopo/, /ikipi/. Three trained phoneticians, native speakers of Dutch, American English and Argentinian Spanish, respectively, recorded all items with stress on the first syllable. All /kp/ stimuli presented release bursts. For each disyllabic item, we used one token per speaker as a natural control stimulus. By systematically replacing the $/C_1C_2/$-cluster in these items by the same cluster out of the other disyllabic items produced by the same speaker but with a different vowel, we created spliced test stimuli such as $/ah_opa/$, and $/ek_ipe/$, where the small vowel denotes vowel coarticulation present in the consonant cluster. Similarly, by replacing the $/C_1C_2/$-cluster in the disyllabic items by the same cluster out of the second token of the same items, we created spliced control stimuli in which the vowel coarticulation matched the flanking vowel, e.g. $/ah_apa/$, $/ek_epe/$. We also created trisyllabic fillers in which the middle vowel either matched or mismatched the flanking vowel, e.g. /ahapa/, /ekepe/, /ahopa/, /ekipe/ (these were also

[1] Note that due to a typo the summary in the first-to-last paragraph of this article erroneously states the opposite.

created by splicing, as they served as test stimuli in an experiment not reported in this article). Overall, each speaker thus contributed 40 test stimuli (5 flanking vowels x 4 vowel coarticulations x 2 consonant clusters), 20 control stimuli (5 flanking vowels x 2 consonant clusters, all both in a natural and a spliced form), and 50 fillers. Ten additional training items were recorded by a fourth speaker. Their structure was similar, but included only phonotactically legal nasal + stop sequences with or without an intervening copy vowel (e.g., /ampa/, /enepe/).

2.2.2.3 *Procedure*

Participants were tested individually in a soundproof room. At each trial, they heard a stimulus over headphones and were asked to identify the vowel between the two consonants, if any. They were provided with a transcription of the item on screen, containing a question mark between the two consonants (e.g. *"ah?pa"*) in latin characters (as non-CV syllables cannot be transcribed using Japanese characters), as well as the list of possible responses: *"none, a, i, u, e, o"*. Participants responded by pressing labelled keys on a keyboard. Participants were familiarised with the procedure with 10 training trials in which they received on-screen feedback.

The 330 stimuli were presented in a pseudo-randomised order: Consecutive stimuli were produced by different speakers, and a stimulus could not be followed by a stimulus with the same combination of vowel coarticulation and consonant. Trials were presented in two blocks, with each stimulus appearing once per block, for a total of 660 trials. The experiment lasted approximately 40 minutes.

2.2.3 Results

Test and control trials with responses that were either too fast (before the medial portion of the stimulus could be perceived and processed, <400 ms) or too slow (> 3 SD: 3238 ms) were excluded from the analyses. This concerned 736 trials (4.5%).

2.2.3.1 *Control items*

Participants experienced perceptual epenthesis in 57% of control items in which the flanking vowel and coarticulation are of the same quality (/hp/: 52%, /kp/: 61%). Recall that in loanwords, the default epenthetic vowel is /u/, while after voiceless laryngeal fricatives it is a copy of the preceding vowel. Focusing on trials with an epenthetic response, we examined whether the choice of epenthetic vowel reflected this pattern.

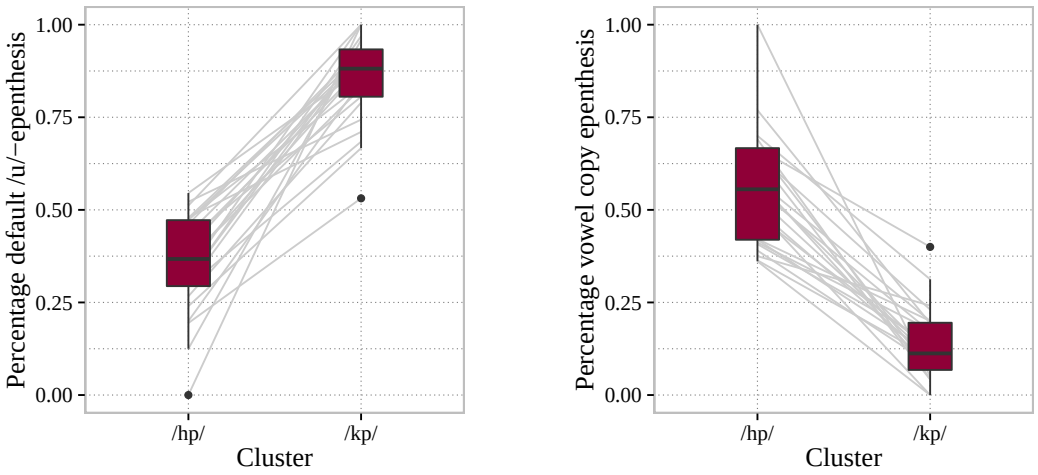

Figure 2.1: *Percentage of default /u/-epenthesis (left) and vowel copy epenthesis (right) for control items. Box plots display the distribution of the scores across speakers (median, quartiles and extrema), with gray lines connecting data points corresponding to a single participant.*

First, a generalised mixed-effects model with a declared binomial distribution [Bates et al., 2015] was used to examine a possible effect of consonant cluster on default /u/-epenthesis. Thus, we analyzed the proportion of default /u/, using participant, speaker, experimental block, and trial as random effects, and consonant cluster (/kp/ *vs.* /hp/; contrast coded) as fixed effect. This model was compared to a reduced model with no fixed effect. The full model was found to explain significantly more variance than the reduced model ($\beta = -4.2$, $SE = 1.2$, $\chi^2(1) = 9.9$, $p < 0.01$), showing that participants experienced significantly less default /u/-epenthesis in /hp/- than /kp/-items (39% *vs.* 86% of all trials with epenthesis, respectively).

Next, we examined whether epenthesized vowels shared the quality of the flanking vowel more often in /hp/- than in /kp/-clusters. Given that for items with flanking vowel /u/ it is impossible to know if /u/-epenthesis is due to vowel copy or to default epenthesis, these items were excluded. As before, a generalised mixed-effects model with a declared binomial distribution was used. We analyzed the proportion of vowel copy (i.e., whether the flanking vowel and epenthetic vowel shared quality), using participant, speaker, experimental block, and trial as random effects, and consonant cluster (/kp/ *vs.* /hp/; contrast coded) as fixed effect. Comparing this full model to a reduced model with no fixed effects revealed a significant effect of consonant cluster ($\beta = 3.7$, $SE = 1.2$, $\chi^2(1) = 7.4$, $p < 0.01$). Therefore, participants epenthesized a vowel that matched the flanking vowel more often in /hp/-clusters (53%) than in /kp/-clusters (13%).

Thus, analysis of control items revealed that, similarly to the loanword pattern, participants perceived the vowel /u/ more often in /kp/- than in /hp/-clusters, and they perceived a vowel copy more often in /hp/- than in /kp/-clusters.

2.2.3.2 *Test items*

Figure 2.2 shows trial counts, separated according to response category, consonant cluster, flanking vowel, and vowel coarticulation for test and control trials. Within the individual rectangles, vertical lines are indicative of a larger influence of flanking vowels compared to vowel coarticulation. Horizontal lines, by contrast, are indicative of a larger influence of vowel coarticulation. Finally, uniform colouring indicates that neither flanking vowels

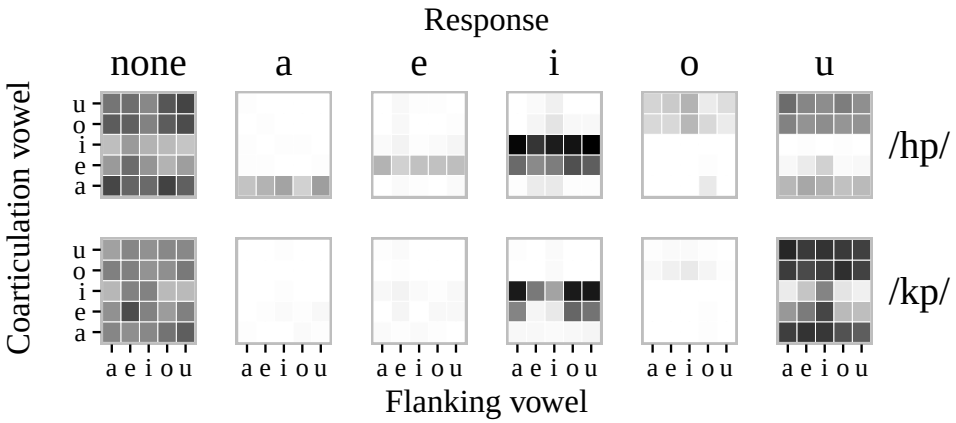

Figure 2.2: *Counts of responses for the test items and spliced control items. Top: /hp/-items; bottom: /kp/-items. Within each rectangle, flanking vowels and vowel coarticulation are given in the horizontal and vertical axes, respectively. Darker colours indicate higher counts.*

nor vowel coarticulation have the upper hand in influencing the quality of the epenthetic vowel. Note that except for the rectangles with "none" and "u" responses where colouring is more uniform, horizontal lines are more visually prominent than vertical lines. Thus, the epenthetic vowel's quality generally depends mostly on acoustic details present in the consonant cluster.

Focusing on the test trials eliciting epenthesis (/hp/: 62%, /kp/: 66%), we quantify the respective influence of flanking vowel and vowel coarticulation (explanatory variables, EV) on the epenthetic vowel (response variable, RV), using two measures from information theory, **mutual information** (MI) and **information gain** (IG) [**?**, see]for a comprehensive description of these measures]daland2015. MI and IG are derived from **entropy**, which is the 'uncertainty' in the value of a RV at a given trial. The lower the entropy $H[X]$ of a variable X, the easier it is to predict the outcome of a trial. The **MI** $I[X;Y]$ of variables X and Y represents the reduction in 'uncertainty' of the trial outcome for RV X, given that the value of EV Y is known (and vice versa). This corresponds to the maximum amount of influence that Y can have over X, without removing contributions from other variables. **IG** $H[X|Z] - H[X|Y,Z]$ represents the minimum amount of influence of variable Y on X. This corresponds to the reduction in uncertainty as to the value of X that arises from knowing the value of Y, after removing all uncertainty explained by variable Z.

As in [Daland et al., 2015], we compute **accidental information** introduced to MI and IG, which corresponds to inaccuracies introduced to our measurements by the process of inferring underlying probability distributions from samples, i.e., sampling error (as when one does not obtain 50 tails and 50 heads when flipping a fair coin 100 times). We can estimate the accidental information by recomputing MI and IG after having removed the dependencies between the EV and the RV. We can do so by shuffling the values of the EV within each participant. For instance, in order to compute the accidental information introduced to MI and IG for the EV "vowel coarticulation", we randomly shuffle the vowel coarticulation labels of all of our trials, per participant, while leaving the EV "flanking vowel" untouched. We then compute MI and IG as for the real data. In order to obtain a better estimate of accidental information from an average value, we do this 1000 times (i.e., Monte Carlo shuffling process).

To recapitulate, for both coarticulation vowel and flanking vowel, we compute 'sam-

ple' and 'accidental' MI and IG. The 'true' values of these measures are obtained by removing mean accidental information from sample information. Following [Daland et al., 2015], we consider the set of shuffled datasets (i.e., 'accidental' MI and IG) as probability distributions given by the null hypotheses that neither coarticulation nor the flanking vowel influence the responses.

Table 2.1: *Quantified influence of vowel coarticulation and flanking vowel on vowel epenthesis measured with information gain (IG) and mutual information (MI). Ranges for Monte Carlo simulations of the null hypothesis (i.e. accidental information) are given in square brackets. Values are given in bits.*

	Vowel coarticulation				Flanking vowel			
	IG		MI		IG		MI	
	data	null	data	null	data	null	data	null
/hp/	.90	.04 [.02, .05]	.93	.01 [0, .02]	.07	.03 [.02, .05]	.09	.01 [0, .02]
/kp/	.47	.03 [.02, .05]	.53	.01 [0, .02]	.07	.03 [.02, .04]	.13	.01 [0, .02]

As shown in Table 2.1, all sample lower bounds are greater than their respective accidental information gains on all 1000 shufflings, for which the ranges are given in parentheses. Therefore, the 'true' lower bounds for both coarticulation and flanking vowel influence on epenthesis are greater than 0 with $p < 0.001$, showing that both coarticulation and flanking vowel quality influence participant responses. However, the amount of influence differs greatly: a larger information gain is yielded by considering vowel coarticulation than by considering the flanking vowel. This is true both for /hp/-items, which are heavily coarticulated, and for /kp/-items, where coarticulation is mainly only present in the burst, even though the influence of coarticulation on epenthetic vowel quality is higher for the former (/hp/: $[0.86, 0.92]$ *vs.* /kp/: $[0.44, 0.52]$). (The range of variation within shuffles of accidental information was about .03; thus any difference of .06 or bigger is significant, including differences between MI and IG values, respectively). In summary, both vowel coarticulation and the flanking vowel influence epenthetic vowel quality, but this influence is greater for vowel coarticulation; response patterns are more predictable when the value of this variable is known than when the value of the flanking vowel variable is known.

2.2.4 Discussion and conclusion

We used an identification task to assess the quality of epenthetic vowels perceived by Japanese listeners in illegal consonant clusters with varying amounts of coarticulation. Our findings can be summarized as follows: First, we were able to replicate the perception of illusory vowels within phonotactically illegal clusters by Japanese listeners (64% of all test trials).[2][3] Second, when the flanking vowel and coarticulation match, the quality of

[2]Note that whereas previous studies examined perceptual epenthesis within clusters with at least one voiced consonant, we presently focused on completely voiceless clusters, a context in which the high vowels /i/ and /u/ may be devoiced in Japanese [Han, 1962, Vance, 1987]

[3]As pointed out by an anonymous reviewer, the differences in rates of epenthesis by speaker (Dutch: 68%, Am. English: 58%, Arg. Spanish: 66%) are consistent with an important role for acoustic factors in epenthesis, suggesting that participants interpret speakers' acoustic cues instead of responding based on abstract phonological categories [?, cf.]]wilson2014. This can also be seen in more detail when decomposing Figure 2.2 by speaker, as in the annex Figure 2.3

the perceived vowel patterned in the same way as in loanword adaptation data. That is, for /kp/-clusters, the predominant epenthetic vowel was the standard default vowel for Japanese (/u/), while for /hp/-clusters, it was a copy of the flanking vowel. Finally, and most importantly, in items where the coarticulation and flanking vowel differed, the quality of the epenthetic vowel was significantly influenced by both variables, but the influence of the former was much larger than that of the latter, especially in the case of /hp/. Our discussion focuses on this last finding.

Before discussing its theoretical relevance, let us comment on the numerically small – yet significant – influence of flanking vowel on epenthesis for /hp/-clusters, where vowel coarticulation is maximal. This result suggests a contribution of categorical variables on epenthetic vowel quality (i.e., copy effect). A similar effect, though, was also found for /kp/-clusters, for which loanword adaptation patterns provide no particular reason to propose the existence of a categorical copy phenomenon; indeed, in loanwords, coda-/k/ generally triggers default /u/-epenthesis. Therefore, it is possible that this effect results from a response bias due to task demands: given a perceptually uncertain stimulus, the flanking vowel could prime a 'copy' response, for instance, because it was visually available on-screen at each trial (e.g. "ah?pa"). Further work using different tasks is necessary to examine the perceptual reality of this 'vowel copy' effect.

Keeping in mind that this work focuses on the choice of epenthetic vowel, while not directly addressing questions related to why phonologically-illegal clusters are repaired, or what the role of phonotactics in epenthesis is, the finding that the quality of the epenthetic vowel is influenced more by coarticulation than by the flanking vowel calls for a perceptual repair mechanism in which acoustic details are taken into consideration. Two-step models in which epenthetic repair is performed after the consonant cluster in the acoustic input has been represented in terms of discrete phonetic categories are therefore ruled out. Rather, like [Dupoux et al., 2011], we argue in favor of one-step models, in which epenthetic vowel quality is based on the similarity between local acoustic cues and prototypical properties of each vowel in the language, such that the closest matching vowel gets selected for insertion. This mechanism can account both for the coarticulation-induced vowel copy effect in items with a /hp/-cluster, as the voiceless glottal fricative /h/ contains strong coarticulation from the adjacent vowels [Keating, 1988] also see Annex Figure 2.4, and for the default /u/-epenthesis effect in items with a /kp/-cluster – which exhibit a lower degree of coarticulation – as /u/ is the phonetically shortest vowel in the language [Han, 1962] and is prone to be devoiced in certain contexts (see footnote).

Focusing on cases where the quality of the epenthetic vowel varies *within language* as a function of the type of cluster, previous studies have investigated whether language-specific phonotactic or phonological properties play a role for the quality of the epenthetic vowel. In Japanese, for instance, dental stops cannot be followed by /u/, and in loanwords this phonotactic constraint gives rise to adaptation by means of /o/-epenthesis (e.g. *'batman'* $\rightarrow'$ *batoman'*). Using identification tasks, both [Mattingley et al., 2015] and [Guevara-Rukoz et al., 2017] report that the perceptual equivalent of this effect is only marginally present in Japanese listeners [**?**, 10-12% of /o/-epenthesis in /d/-initial clusters; see also]for the absence of such an effect in a discrimination task]monahan2009. Thus, so far there is only weak evidence that the mechanism of phonotactic repair takes into account the legality of the resulting CVC-sequence. A stronger effect of cluster-dependent perceptual epenthesis has been reported in Korean listeners, who repair /eʃma/ and – to a lesser extent – /ec^{h}ma/ with an epenthetic /i/ instead of the default epenthetic vowel /ɨ/ [Durvasula and Kahng, 2015]. This is argued to be due to the existence of an allophonic rule that palatalizes /s/ and /t^{h}/ before /i/, yielding [ʃi] and [c^{h}i], respectively. It is also possible, however, that this effect is (partly) due to coarticulation; for instance, acoustic cues in /ʃ/ and /c^{h}/ might be more suggestive of /i/ than of /ɨ/.

To conclude, we directly compared the relative contributions of acoustic and categorical effects on epenthetic vowel quality, and found that the former override the latter. This result thus strengthens those of [Dupoux et al., 2011], who also established the presence of acoustic effects but without investigating possible categorical effects. More research is needed to investigate whether our findings generalize to other cases of perceptual epenthesis. This question can be addressed by two complementary approaches. One would be to run additional experiments with cross-spliced stimuli, as in the present study. Another one would be to measure the effective amount of coarticulation in experimental stimuli of previous studies, using a computational implementation of a one-step repair mechanism (see [Dupoux et al., 2011] and [Wilson et al., 2014] for propositions, and [Schatz, 2016] and later chapters of this thesis for implementations using Hidden Markov Models).

2.2.5 Annexes

Here we provide additional results for (a) differences in patterns of epenthesis for items recorded by different speakers, (b) acoustic analyses of coarticulation, and (x) an ABX discrimination task.

2.2.5.1 Identification results separated by speaker

Our three recorded speakers did not share the same native language, causing their recorded items to differ in their acoustic details. A consequence of this is that response patterns of Japanese participants had subtle differences according to the speaker producing the stimuli, as seen in Figure 2.3. For instance, most "o" responses were prompted by stimuli recorded by the Dutch speaker, while most "e" responses arose from stimuli by the American English speaker. Importantly, as mentioned in the discussion, rates of epenthesis and choice of epenthetic vowel varied according to speaker, which further supports our hypothesis that Japanese participants attended to acoustic details when experiencing perceptual epenthesis. It would be interesting to see whether these differences are due to phonetics specific to the native language of the speakers or to personal idiosyncracies, since here we only recorded one speaker per native language.

2.2.5.2 Acoustic analyses

In order to examine the acoustic properties of our stimuli, we annotated them using Praat [Boersma et al., 2002] and we automatically extracted the first three formants from all vowels in the stimuli (V_{1a} and V_{1b} from $V_{1a}CpV_{1b}$ items, and V_2 from CV_2p items used to construct $V_1CV_2pV_1$ items), and also from /h/ and /k/ in /Cp/ clusters (coarticulation). Their distribution in F1 x F2 space can be seen in Figure 2.4. As might be expected, the vowel triangle formed by vowels /i, a, u/ is discernible when plotting full vowels in F1 x F2 space. This is not the case, however, when plotting coarticulation contained by consonants /h/ and /k/. Visually, it appears that the distinction between front vowels /i, e/ and the rest (/a, o, u/ is better maintained in /h/ than in /k/. We used Linear Discriminant Analysis (LDA) to perform classification of the points plotted in Figure 2.4, using as input features a vector containing the first three formants F1, F2, and F3 of each datapoint. We trained the LDA classifier first using data from full vowels as training data, in order to classify coarticulation from the consonants. The classifier accuracy is 38.0% for coarticulation in /h/ and 33.3% for coarticulation in /k/. The corresponding classification patterns can be found in the top part of Figure 2.5. As we can see, classification patterns are similar; /i, e/ coarticulation is classified as /e/, while /a, o, u/ are mostly classified as /a/. Furthermore, we used LDA with cross-validation (i.e., from a set of n items, an item is classified based on LDA performed on all other $n-1$ items). When the set of

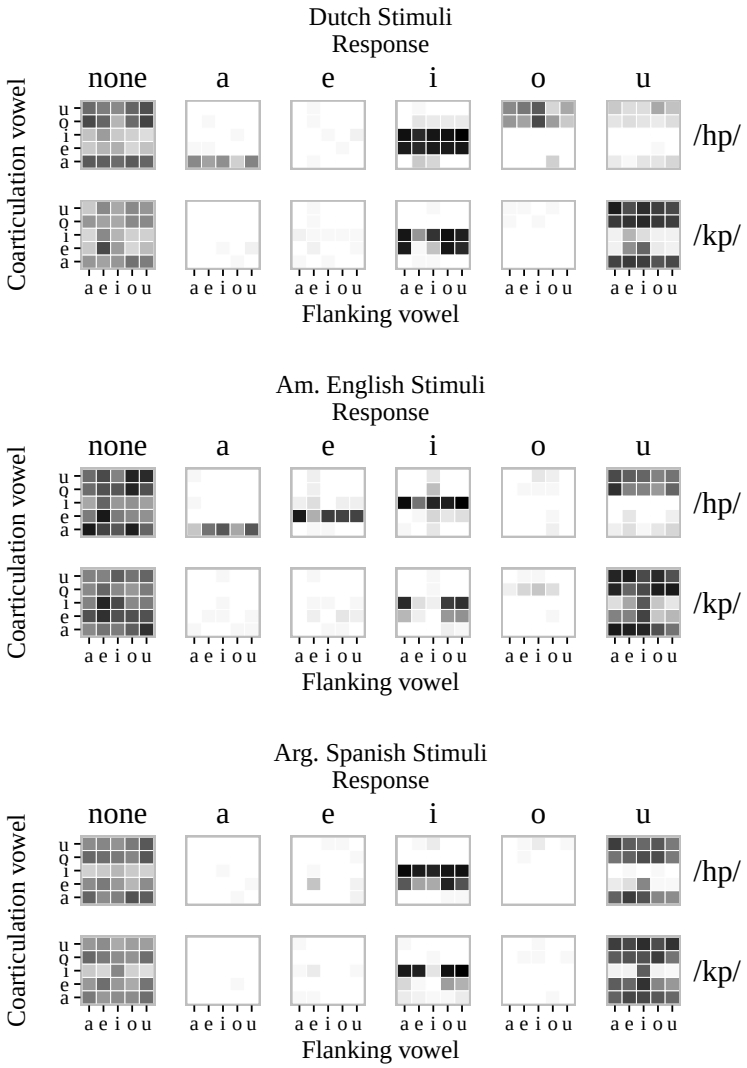

Figure 2.3: *Counts of responses for the test items and spliced control items, separated by speaker. For each speaker: top: /hp/-items; bottom: /kp/-items. Within each individual rectangle, flanking vowels and vowel coarticulation are given in the horizontal and vertical axes, respectively. Darker colours indicate higher counts, with colours normalized within each speaker.*

interest was that of coarticulation in /h/, the accuracy was of 51.7%, while it was 36.7% for coarticulation in /k/. The resulting classifications can be seen in the lower section of Figure 2.5; for /h/ the classification patterns are more similar within members of a taxonomic group (e.g., /i, e/) than for /k/. Thus, while coarticulation from both types of clusters can be mapped onto the original vowel space similarly well (or badly, depending on the perspective), it would be easier to deduce the quality of neighbouring vowels from the coarticulation cues contained within an /hp/ rather than a /kp/ cluster, especially with regards to the separation of the front vowels /i, e/ from /a, o, u/, as can be seen in the lower part of Figure 2.5 and the right panels of Figure 2.4.

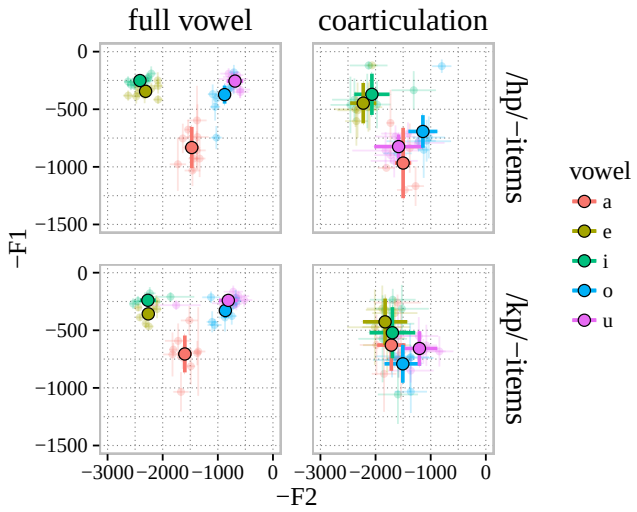

Figure 2.4: *Visualisation in $F1 \times F2$ space of vowels (left panels), and coarticulation found in the first consonant of C_VC clusters (right panels), of items used in the experiment. Dimmer dots and lines respectively show median formant values and median formant bandwidths within the vowel or consonant. Dots circled in black and thicker lines show global means.*

2.2.5.3 Supplementary experiment: ABX task

In this additional experiment we assessed the perception of illegal consonantal clusters in Japanese using an ABX discrimination task, which, contrary to the vowel identification task used in Experiment 1, does not require an explicit categorization of the item's segments. As in previous work [Dupoux et al., 1999, Dupoux et al., 2011], we used different speakers for stimuli A, B, and X, such that the task could not be performed on the basis of low-level acoustic information.

Participants Twenty-six native Japanese listeners were recruited in Paris, France. While testing for this experiment was done outside of Japan, we recruited only participants with little experience with French or other languages in which consonant clusters are allowed. For instance, many participants were recently arrived exchange students or family members of professionals that had been transferred to Paris.

Stimuli From the stimuli used for the identification task, we extracted items relevant for pairs shown in Table 2.2. We defined four types of AB pairs with constant flanking vowels, based on the nature of the items in the pair:

- Natural cluster items (N) correspond to natural control stimuli from the identification task, disyllabic $V_1C_1C_2V_1$ items which have not been spliced.
- Spliced cluster items (Sp) correspond to the identification task test stimuli, disyllabic $V_1C_1(V_2)C_2V_1$ items for which the $C_1(V_2)C_2$ cluster has been spliced from a $V_2C_1C_2V_2$ item.
- Full vowel items (FV) correspond to trisyllabic fillers from the identification task, $V_1C_1V_2C_2V_1$ items for which the $C_1V_2C_2$ cluster has been spliced from a $V_2C_1V_2C_2V_2$ item

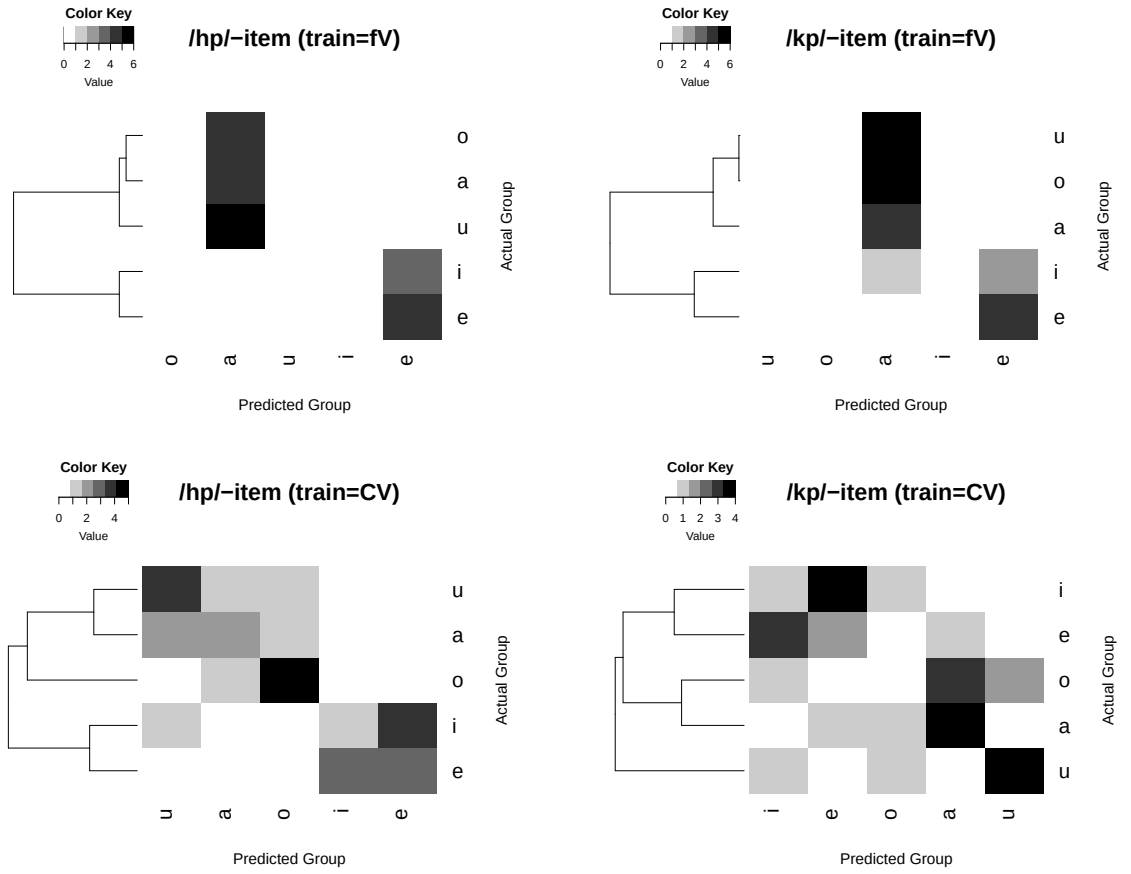

Figure 2.5: *Classification of consonants /h/ and /k/ based on formant values of their coarticulation cues. Classification was performed based on category descriptions dictated by formant values of full vowels (top) or in a cross-validation manner (bottom). Consonants are labeled according to the quality of coarticulation cues, therefore of neighbouring vowels ("Actual Group", rows). Classification labels are shown in columns ("Predicted Group"), with darker colours indicating higher counts. Dendrograms on the left-hand side of each heatmap show the grouping of consonants according to their similarity in the classification patterns. Diagonals show identity. Please note that the order of the vowels differs between the four panels, since it is set by the dendrograms.*

Table 2.2 also shows how well participants are predicted to discriminate items in the AB pairs depending on how phonotactically illegal stimuli might be repaired. Participants might break the illegal consonant cluster by adding a vowel identical to flanking vowels (Flank.), by adding a vowel of the same quality as the coarticulation (Coart.), or they might simply add /u/ by default (Default). Participants might also not experience epenthesis at all (No Epenth.).

Procedure Participants were tested in a soundproof room wearing headphones. On each trial, participants heard two different stimuli of categories A and B, followed by a third stimulus X, belonging to either category A or B. Within each trial, all three stimuli had a $V_1C(V_2)pV_1$ structure, with V_1 and C remaining constant. The three tokens were produced by different speakers and were presented with an ISI of 500 ms. An ITI of 1 s separated a participant's response from the following trial.

Within each triplet, A always contained either a natural or a spliced cluster, while

Table 2.2: Types of AB pairs for Experiment 2. The discrimination accuracy is predicted according to the following hypotheses about epenthetic vowel quality: (1) it is determined by flanking vowel quality (Flank.); (2) it is determined by coarticulation cues (Coart.); (3) participants experience default /u/ epenthesis (Default) epenthesis; (4) participants do not experience epenthesis (No Epenth.). Cases of good discrimination are marked with plus signs.

Type	A	B	# pairs	Example	Flank.	Coart.	Default	No Epenth.
N-Sp	natural	spliced	40	$/ahpa/ - /ah_ipa/$	−	+	−	−
Sp-Sp	spliced	spliced	100	$/ah_ipa/ - /ah_epa/$	−	+	−	−
N-FV	natural	full V	10	$/ahpa/ - /ahapa/$	−	−	+	+
Sp-FV	spliced	full V	50	$/ah_ipa/ - /ahipa/$	+	−	+	+

B always contained either a full vowel or a spliced cluster. Table 2.2 shows the four different types of AB pairs that were thus tested, together with the expected discrimination accuracy based on different hypotheses about how epenthetic vowel quality is determined.

In total, there were 200 AB pairs. Since there are four possible presentation orders for each pair and its corresponding third item X (i.e., ABX_A, BAX_A, ABX_B, BAX_B), there are 800 possible unique trials. In order to reduce the duration of the experiments, participants were divided into two groups exposed to counterbalanced halfs of the total set of trials.

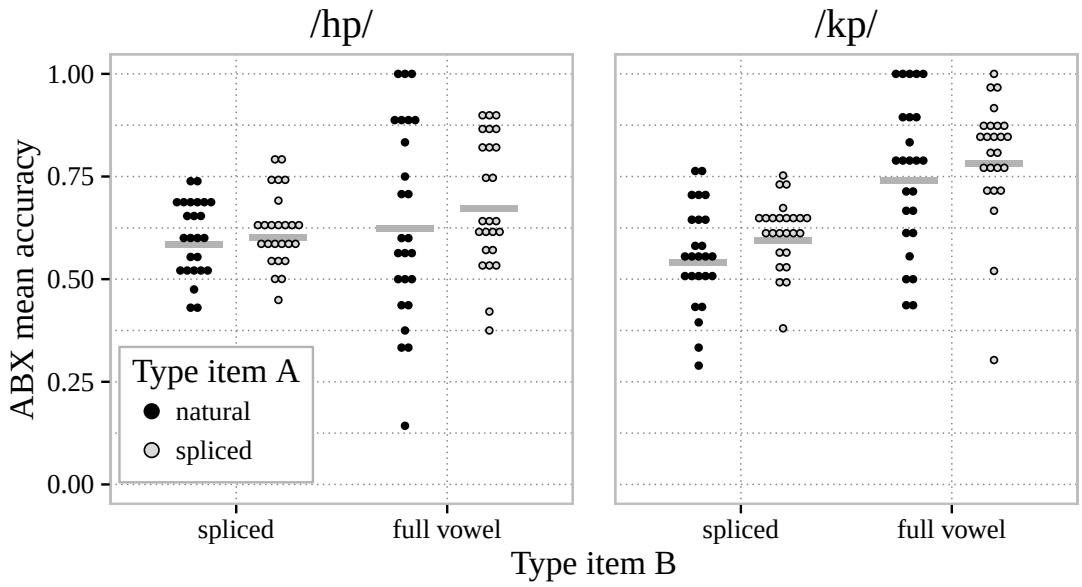

Figure 2.6: *Discrimination accuracy at the ABX task on /hp/ (left) and /kp/ (right) items. Dot plots show the distribution of average scores (one dot per participant). Horizontal grey lines show mean accuracy for each AB pair type.*

Results Trials in which the response was given before all items in the ABX triplet had been played were excluded (1063 trials representing 11% of all trials). The remaining data were analysed using a generalised linear mixed-effects model in R (*lme4*; [Bates et al., 2015]) with a declared binary distribution. The binomial response variable of interest for each trial was Accuracy (*correct vs. incorrect*); were included as fixed

effects CONSONANT (*/h/ vs. /k/*), TYPE A (*natural vs. spliced*), TYPE B (*full vowel vs. spliced*), as well as the interactions between every pair of fixed effects. All fixed effects were contrast-coded. PARTICIPANT, ITEM A, ITEM B, and TEST GROUP were included as random effects. Significance testing was done through model comparison: the full model including all fixed and random effects was compared to reduced models, in which one of the fixed effects was absent.

The full model did not explain significantly more data variance than a model excluding the fixed effect CONSONANT, suggesting that participant accuracy was not significantly different for /hp/ and /kp/ trials ($\beta = 0.17$, $SE = 0.16$, $\chi^2(1) = 1.1$, $p > 0.05$).

We did not find evidence of accuracy being lower or higher when an ABX trial contained a natural cluster item instead of a spliced cluster item (TYPE A; $\beta = 0.20$, $SE = 0.12$, $\chi^2(1) = 2.52$, $p > 0.05$). Moreover, there was no significant interaction between CONSONANT and TYPE A ($\beta = -0.10$, $SE = 0.17$, $\chi^2(1) = 0.37$, $p > 0.05$), nor between TYPE A and TYPE B ($\beta = 0.08$, $SE = 0.15$, $\chi^2(1) = 0.26$, $p > 0.05$).

By contrast, accuracy was significantly enhanced by the presence of an item with a full vowel cluster (i.e., *Sp-FV* and *N-FV* pairs) (TYPE B, $\beta = 0.93$, $SE = 0.16$, $\chi^2(1) = 29.8$, $p < 0.0001$). This increase in accuracy appears to be exacerbated in pairs with /kp/-containing items relative to pairs with /hp/-containing items, as the interaction between CONSONANT and TYPE B was also significant ($\beta = -0.7$, $SE = 0.15$, $\chi^2(1) = 19.0$, $p < 0.0001$).

These results are compatible with predictions given by the DEFAULT and NO EPENTH. hypotheses in Table 2.2, i.e., better discrimination for N-FV and Sp-FV pairs. After looking at response patterns from the identification task, this should come as no surprise. Indeed, most of participant responses were "none" (36% of test trials) and "u" (32% of test trials). This ABX task is therefore not sensitive enough to detect differences between the more subtle modulations in epenthetic vowel quality caused by flanking vowels and coarticulation cues. However, we can examine the correlation between ABX discriminability and response patterns for the identification task, in order to verify that results from the latter experiment are not solely due to task-specific demands (e.g., participants focusing on phoneme identity).

Correlation with identification results In order to assess the role of perceptual assimilation on stimulus discrimination, we derived a measure of perceptual distance from response patterns given in the identification task, and examined if this distance predicted the outcome in the ABX discrimination task. To do so, we computed for each item a six-dimensional numerical vector of the shape $x = [x_1, ..., x_6]$, with values corresponding to the percent responses to categories a, e, i, o, u, and *none*, respectively. The distance $d(x, y)$ between two items x and y was computed as the normalized Euclidian distance between their associated vectors:

$$d(x, y) = \frac{\sqrt{\sum_i (x_i - y_i)^2}}{\sqrt{2}}$$

One data point was obtained per AB pair, giving a total of 200 datapoints (cf. Table 2.2).

Multiple regression analysis was used to test if assimilation patterns from the identification task significantly predicted participants' accuracy during the ABX task. A scatterplot summarizes the results in Figure 2.7. The model included as independent variables the normalized perceptual distance between two items (range = [0;1]), and the consonant cluster (/hp/ or /kp/). These two predictor variables explained 52% of the variance ($R^2 = 0.52$, $F(3, 196) = 73.63$, $p < 0.0001$). Consonant cluster ($t < 1$) and the interaction of the two independent variables ($t < 1$) were not significant. On the other

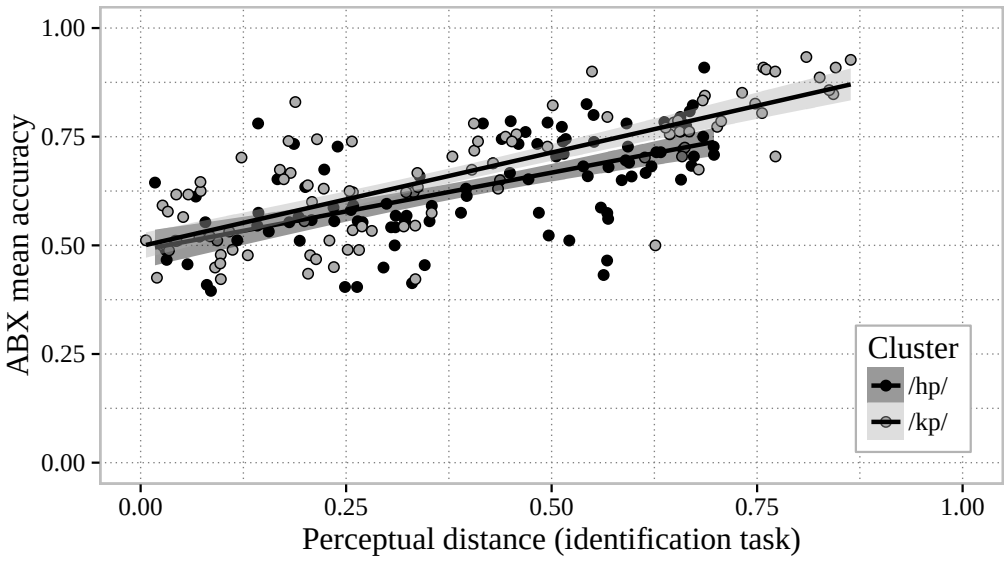

Figure 2.7: *Correlation between the perceptual distance derived from the identification task responses and the accuracy at the ABX discrimination task.*

hand, perceptual distance significantly predicted accuracy during the ABX task ($t = 13.8$, $p < 0.0001$); the less similar the response patterns to both items in the AB pair, the easier their discrimination in the ABX task. These results suggest that adaptation patterns attested in the identification task are not task-dependent.

2.3 Predicting epenthetic vowel quality from acoustics

The following section is a modified version of the following article:
Guevara-Rukoz, A., Parlato-Oliveira, E., Yu, S., Hirose, Y., Peperkamp, S., and Dupoux, E. (2017). Predicting epenthetic vowel quality from acoustics. Proceedings of Interspeech, 596-600.
Stimuli were designed and recorded by E. Parlato-Oliveira and E. Dupoux. Experimental and production data were collected by E. Parlato-Oliveira and Y. Hirose. Phonetic transcriptions were provided by S. Yu. Statistical analyses and exemplar-based models were run by A. Guevara-Rukoz. The initial manuscript draft was prepared by E. Dupoux, S. Peperkamp, and A. Guevara-Rukoz. E. Dupoux supervised the entirety of the study.
Modifications with respect to the original paper: additional figures.

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.

2.3.1 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" → /aisukuri:mu/). This phenomenon has been shown to also happen during on-line perception: listeners *perceive* vowels within illegal consonantal sequences [Dupoux et al., 1999, Dehaene-Lambertz et al., 2000, Dupoux et al., 2001, Berent et al., 2007, Kabak and Idsardi, 2007, Monahan et al., 2009, Dupoux et al., 2011, Mattingley et al., 2015, Durvasula and Kahng, 2015]. 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 [Dupoux et al., 2011]. 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.

2.3.2 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.

2.3.2.1 Methods

Fifty-four items with the structure $V_1C_1C_2V_2$, with V_1 and V_2 vowels from the set {/a/, /i/, /u/}, and C_1C_2 a cluster from the set {/bg/, /bn/, /db/, /dg/, /gb/, /gn/}, e.g. /abgi/, 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.

2.3.2.2 Results

Statistical analyses were performed with the R statistical software [R Core Team, 2016], using MCMC glmm [Hadfield, 2010, Plummer et al., 2006]. 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 section (as well as additional information) can be found in: https://osf.io/zr88w/.

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.

Response patterns are shown on Figure 2.8. 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: 1%, JP: 3%). Percentages for the remaining responses of interest (/i/, /o/, and /u/) can be seen in the lefthand part of Table 2.3.

Table 2.3: Percentage of responses.

	Human data			Model		
	i	o	u	i	o	u
BP	80.39	0.64	18.97	52.73	6.22	41.05
JP	18.37	5.64	75.98	49.34	0.13	50.52

/i/-epenthesis Figure 2.9 shows the proportion of /i/-epenthesis. A main effect of Language shows that BP participants perceived an epenthetic /i/ more often than JP participants (posterior mode: -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]).

/u/-epenthesis Figure 2.9 shows the proportion of /u/-epenthesis. We found a main effect of Language; BP participants epenthesized /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 C_1 (-43.0, $[-75.8, -9.8]$): participants perceived /u/ less often after coronal than after labial and velar consonants. However, neither the interaction between Coronal C_1 and Language (-63.9, $[-132.1, 3.0]$), nor the triple interactions with NSFV (-17.8, $[-124.9, 84.1]$, -14.8, $[-238.4, 304.3)$ were significant; thus, JP participants were not more prone to avoiding /u/-epenthesis after coronal consonants than BP participants.

2.3.3 Acoustic analyses

In both BP and JP, the shortest vowel corresponds to the default epenthetic vowel, i.e. /i/ in BP [Escudero et al., 2009] and /u/ in JP [Han, 1962]. Here, we compare epenthetic vowels /i/, /u/, and /o/ on three acoustic parameters: (1) vowel duration, (2) vowel

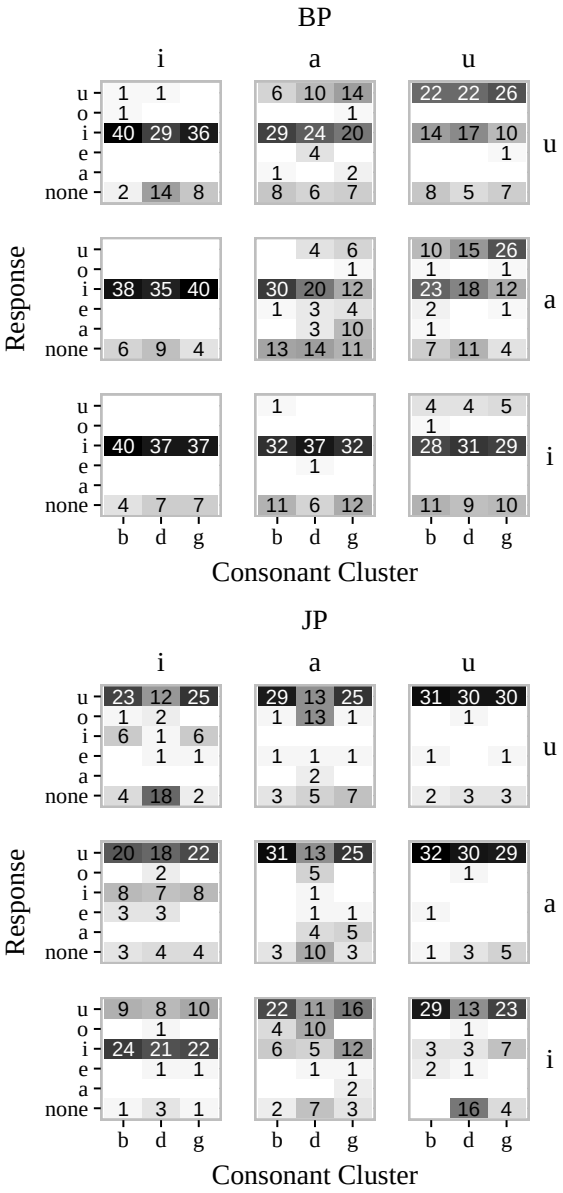

Figure 2.8: *Responses for all trials from the perception experiment for both BP (top) and JP (bottom), including trials with responses not given by the exemplar model ("none", "a", "e"). Numbers indicate trial counts, with darker cell backgrounds representing higher values. Within each of the two 3 x 3 grid, trials are separated according to V_1 (columns) and V_2 (rows). Within each individual rectangle, the horizontal axis shows the first consonant of the consonant cluster, while the vertical axis corresponds to possible responses.*

intensity, and (3) Euclidean distance between vowel formants and formant transitions in consonant 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.

2.3.3.1 Methods

Seventeen BP and 17 JP participants from the perception experiment were also recorded producing 162 stimuli obtained by crossing the 54 $V_1C_1\_C_2V_2$ frames of the experimental

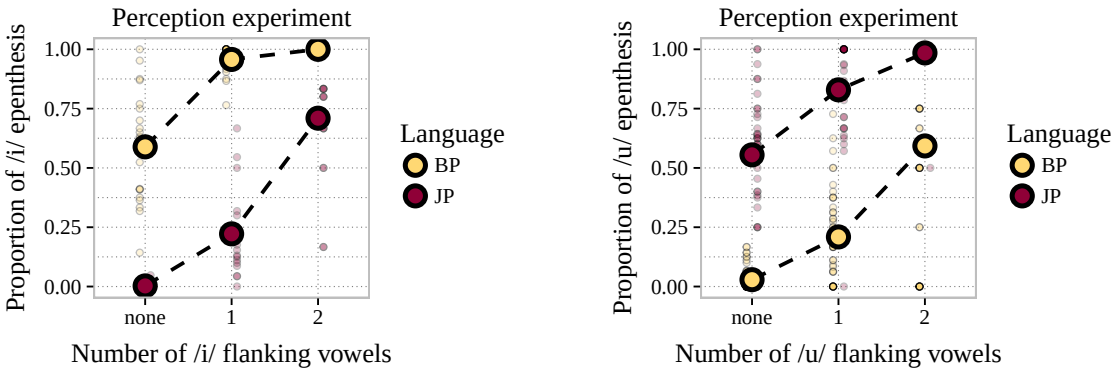

Figure 2.9: *Proportion of /i/-epenthesis (left) and /u/-epenthesis (right) exhibited by BP and JP participants in the perception experiment. Big dots show mean values, while smaller dots show individual values for human participants.*

items with the three vowels /i/, /o/, and /u/ (e.g. */ab_gi/* → */abigi/*, */abogi/*, */abugi/*). 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 [Bombien et al., 2016].

2.3.3.2 Results

For each of the continuous response variables examined in this section, we fitted an MCMC glmm 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 [Lenth, 2016].

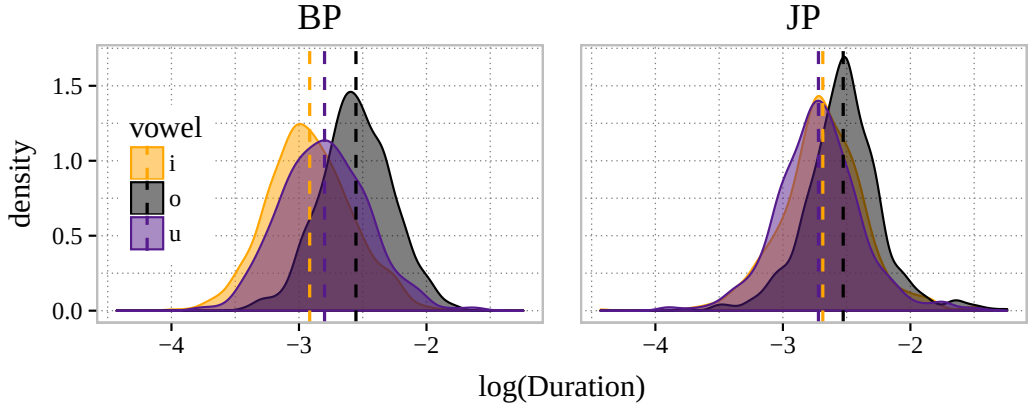

Figure 2.10: *Distribution of log'd vowel duration (in s) of medial vowels /i, o, u/ produced by BP and JP participants. Dashed lines show mean values.*

Vowel duration The measured duration of each medial vowel V_3 (in seconds) was log-transformed to account for distribution skewness. The resulting distributions can be seen in Figure 2.10. 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, /u/ is shorter than /i/ (mean /u/: 69.6 ms; mean /i/: 72.0 ms; adjusted $p < 0.05$).

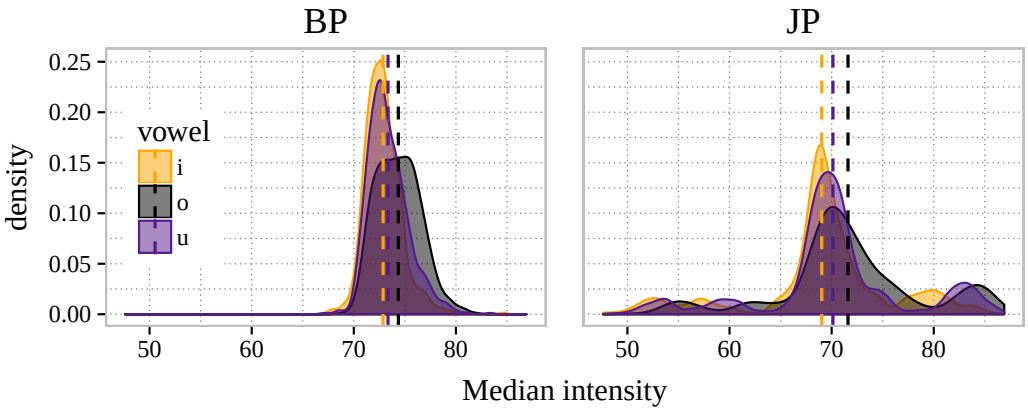

Figure 2.11: *Distribution of median intensity (in dB) of medial vowels /i, o, u/ produced by BP and JP participants. Dashed lines show mean values.*

Vowel intensity We compared the mean intensity of the medial vowels V_3 in decibels (dB). The associated distributions can be seen in Figure 2.11. 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/.

Vowel formants We extracted median formant values (F1, F2, and F3, in Bark) from medial vowels V_3 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/). The resulting distributions can be seen in Figure 2.12. 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 $p < 0.05$).

2.3.4 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'

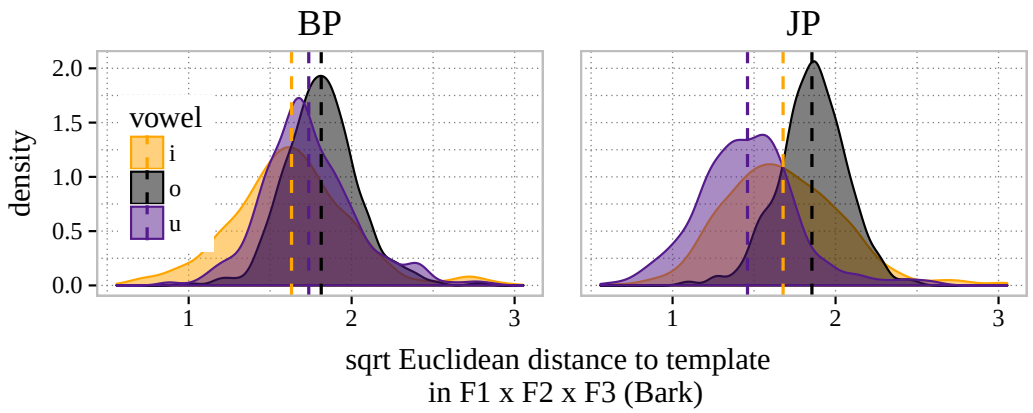

Figure 2.12: *Distribution of square root Euclidean distance to template in F1 x F2 x F3 space (frequencies in Bark) of medial vowels /i, o, u/ produced by BP and JP participants. Dashed lines show mean values.*

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 $V_1C_1C_2V_2$ 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.

2.3.4.1 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[4], 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. /abgi/) and their corresponding $V_1C_1V_3C_2V_2$ epenthesized versions (e.g. /abigi/, /abogi/, /abugi/) using Dynamic Time Warping (DTW) [Sakoe and Chiba, 1978, Giorgino, 2009]. 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 epenthesized versions was defined as

$$s = e^{-cd} \tag{2.1}$$

where d is the DTW distance, and c is a parameter determining the weight of the DTW distance on classification [Nosofsky, 1992]. When $c = 0$, DTW is disregarded and all possible classification categories are equally probable. Higher values of c result in higher probabilities being given to items with smaller d. In order to control for unequal number of tokens in each category, classification was performed by computing the mean similarity

[4]Due to a mistake in the feature computation pipeline, this meant that the log energy and the 12 first MFC coefficients were concatenated, not that the first coefficient of 13 coefficients was replaced by the log energy, as was originally intended.

within each category. From there we sampled a classification label weighting category probabilities by the resulting mean similarity weights. Parameter c was individually optimised for each language by performing leave-one-out cross-validation (maximum accuracy: 0.50 with $c = 0.5$ for BP, and 0.63 with $c = 2.2$ for JP; chance level at 0.33).

2.3.4.2 Results

The same statistical models from Section 2, but without a random effect for Participant, were used.

The perception model was able to accurately predict participant responses for 59.1% (BP) and 58.4% (JP) of trials. Figure 2.14 shows a detailed distribution of the responses. As shown in the righthand part of Table 2.3, the model rarely predicted /o/ responses, as expected based on acoustic analyses; however, it is surprising that most /o/ responses were predicted for BP rather than for JP. This might be due to overlap of /u/ and /o/ in the formant space of BP, which is visible in Figure 2.13.

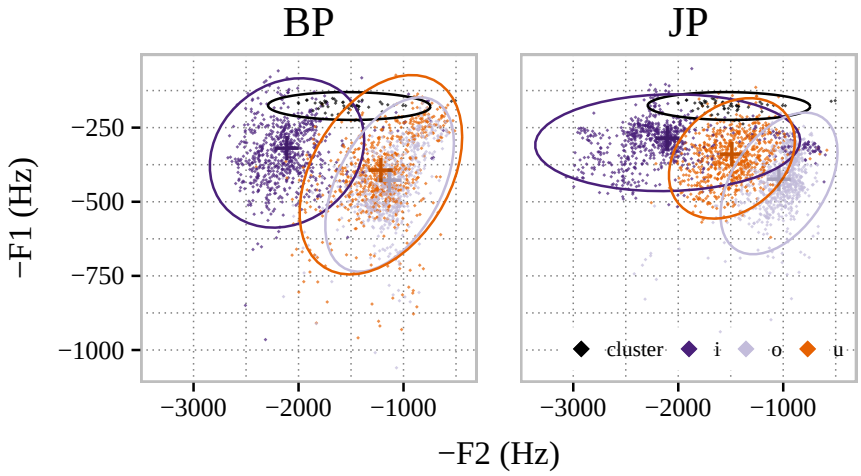

Figure 2.13: *Medial vowels and clusters in F1 x F2 space.*

Concerning /i/ and /u/, numerically the model predicted more /i/ responses than /u/ responses for BP, and the opposite for JP. However, these differences are not as clear as they are for our human data, where in both languages the default vowel is chosen four times more often than the non-default high vowel.

/i/-epenthesis The left panel of Figure 2.15 shows the proportion of /i/-epenthesis for human participants and the corresponding exemplar models. We found a main effect of Language (-52.0, $[-86.9, -23.8]$) and a main effect of NSFV (*none vs.* 1: 93.3, $[53.0, 125.0]$; Less than 2 *vs.* 2: 245.7, $[150.8, 328.2]$). Thus, our model is able to reflect the higher frequency of /i/ as epenthetic vowel in BP compared to JP participants, as well as the influence of flanking vowels on /i/-epenthesis in both BP and JP.

/u/-epenthesis The right panel of Figure 2.15 shows the proportion of /u/-epenthesis for human participants and the corresponding exemplar models. We found a main effect of NSFV (*none vs.* 1: 37.8, $[16.0, 62.3]$; Less than 2 *vs.* 2: 190.7, $[131.0, 240.6]$) but not of Language (-15.9, $[-50.2, 10.1]$). Thus, while our model was able to qualitatively reproduce the influence of flanking vowels on epenthetic vowel quality for /u/, it was unable to reflect the fact that JP listeners perceive /u/ more often than BP listeners.

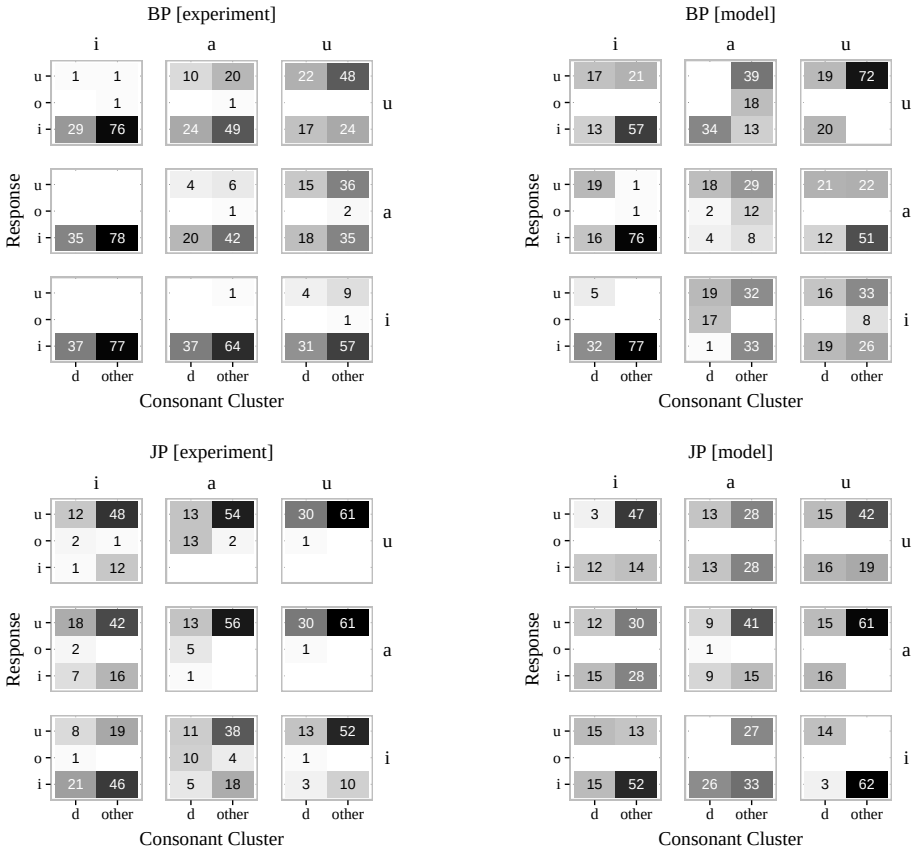

Figure 2.14: *Responses from the perception experiment (left) and model predictions (right), for both BP (top) and JP (bottom), on trials common to the human and model experiments. Numbers indicate trial counts, with darker cell backgrounds representing higher values. Within each 3 x 3 grid, trials are 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 /i/-epenthesis in all 78 trials involving /iC_1C_2a/ stimuli for which C_1 was not the coronal consonant /d/.*

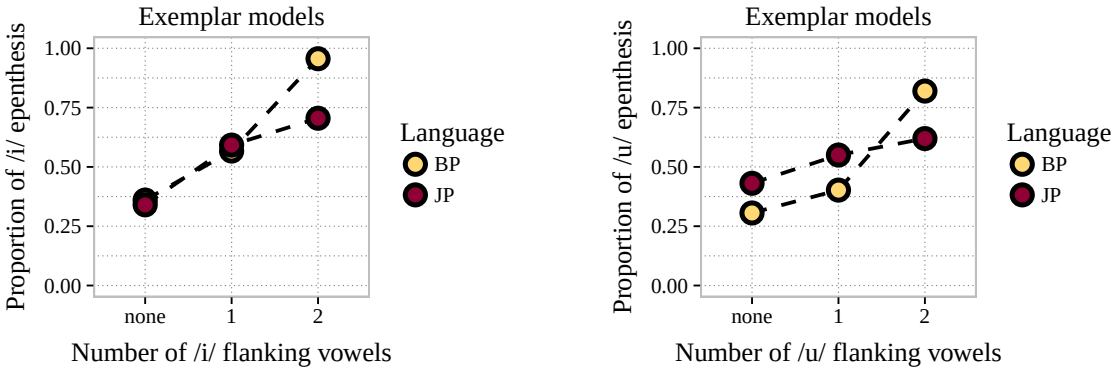

Figure 2.15: *Proportion of /i/-epenthesis (left) and /u/-epenthesis (right) exhibited by exemplar-based models. Dots show mean values.*

There was also a main effect for Coronal C_1 (-66.7, $[-95.6, -39.5]$) but no interaction of this effect with Language ($40.2, [-28.0, 91.8]$); similarly to the perception data, this reflects an overall lower propensity for the model to 'epenthesize' /u/ after coronal consonants. The triple interaction NSFV x Coronal C_1 x Language was significant for both "levels" of NSFV ($94.1, [19.9, 210.6]$, $525.8, [275.5, 666.1]$). Closer inspection suggests that this reflected the model's inability to predict higher percentages of /u/-responses by both BP and JP participants after coronal consonants when both flanking vowels were /u/.

2.3.5 Discussion

Examining epenthetic vowel quality preferences by BP and JP speakers in a perception task, we corroborated previous findings ([Dupoux et al., 1999, Dupoux et al., 2011]) that, like in loanword adaptations, the default epenthetic vowel during speech perception is /i/ for BP and /u/ for JP. Our acoustic analyses suggest that the choice of epenthetic vowel is acoustically driven. That is, in BP, /i/ is shorter and spectrally closer to the formant transitions in our stimuli than /u/ (and /o/), while the reverse holds in Japanese. As such, it may not be necessary to rely on phonological explanations of epenthetic vowel quality as in [Rose and Demuth, 2006, Uffmann, 2006], were we to find that these are shared characteristics of default epenthetic vowels in a variety of languages. We also found an influence of flanking vowels on epenthetic vowel quality, similar to what was reported in [Dupoux et al., 2011]. Indeed, participants gave fewer default responses when the quality of the flanking vowels was in disagreement with the default choice, resulting in more "vowel copy" epenthesis (i.e. perceiving a vowel of the same quality as that of a flanking vowel). Furthermore, we found that this effect of flanking vowels is additive, as it is even more prominent when both flanking vowels are of the same quality.

Interestingly, phonotactics did not influence JP participants' responses as may have been expected; while /o/ was almost exclusively perceived after coronal consonants, this was almost always the case for stimuli with V_1 = /*a*/ (cf Figure 2.8). In fact, for all combinations of flanking vowels, participants responded /i/ and/or /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 [Monahan et al., 2009, Mattingley et al., 2015], suggest that constraints on perception given by surface phonotactics can be overruled by constraints relative to matching input acoustics [Dupoux et al., 2011]. In fact, if this were not the case, novel sound sequences would have never arisen in JP loanwords (e.g. *party* is adapted as /paːti/, not /paːtɕi/).

Finally, we presented results from one exemplar model per language, based on productions by BP and JP participants, respectively. These models reproduced some effects found in the perception experiment — mainly the influence of 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 /u/-epenthesis for JP than BP). These improvements could involve more phonetically and/or temporally informed features (e.g., spectrotemporal representations [Chi et al., 2005]), 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

[Hyafil et al., 2015]).

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.

2.4 Predicting epenthetic vowel quality from acoustics II: It's about time!

2.4.1 Introduction

In the previous section we introduced a production-based exemplar model of perception for which input representations were solely acoustic. We used this relatively primitive model to simulate a perceptual experiment probing perceptual vowel epenthesis by BP and JP listeners. We showed that results from the model shared some qualitative similarities with those from the perceptual experiment, the most notable being the models' ability to reproduce modulations of flanking vowels on the quality of the epenthetic vowel. Putting these results together with the main results from section 2.2 (i.e., higher influence of coarticulation than flanking vowel quality on epenthetic vowel quality), we concluded that modulations of epenthetic vowel quality such as those observed in our perceptual task and in [Dupoux et al., 2011] were due to acoustic details. On the other hand, the production-based exemplar models were not able to adequately reproduce effects related to default epenthetic vowels. Recall that for human participants we saw a majority of /i/- and /u/-epenthesis for BP and JP, respectively.

However, acoustic analyses of recordings made by native BP and JP speakers showed that default epenthetic vowels were not only the closest in formant space to acoustic cues contained in cluster transitions, but they were also the shortest vowels in the inventory /i, o, u/. As such, we can hypothesize that what is particular about default epenthetic vowels is not limited to their spectral characteristics; vowel duration may be as important, if not more, when computing the less costly vowel insertion. Viewing nonnative speech misperceptions as an optimisation problem, where the output is obtained by applying the phonetically minimal modification to the nonnative input [Peperkamp and Dupoux, 2003, Dupoux et al., 2011, Steriade, 2001], we can posit the importance of duration match. In the case of perceptual vowel epenthesis, where the output presents additional segments relative to the input, it would seem logical that, for hypothetically equal spectral properties, shorter segments would be preferred compared to longer segments. For instance, we wouldn't expect JP listeners to epenthesize long vowels instead of short vowels.

We would therefore want our models to take duration mismatches into consideration when computing the similarity between the nonnative input and stored exemplars. This was not the case for models in section 2.3, because the distance between two items was computed using Dynamic Time Warping (DTW), which by design disregards duration mismatch. The goal of the following section was to introduce a duration-mismatch score that could be combined with the original distance score provided by DTW, in order to produce a distance metric that reflects both the spectral and durational proximity of two items.

Additionally, we performed various changes (highlighted throughout the methods section). Most notably, feature standardisation was performed by speaker in the version of the model described bellow, which in a way equates to the model being aware of speaker identity when computing item similarity. As a consequence, this newer version of the model is not purely acoustic, as it is a step towards speaker invariant auditory represen-

tations. Considering these changes, we will address the following questions: First, before even introducing a duration-mismatch penalty, can our newer models reproduce default epenthetic vowels and flanking vowel-related modulations of epenthetic vowel quality? Secondly, what about models with a duration-mismatch penalty?

2.4.2 Methods

2.4.2.1 Features

In order to ensure feature compatibility with future experiments (due to differences in file formats), features were recalculated using the Kaldi speech recognition toolkit [Povey et al., 2011], introducing slight changes regarding the parameters used in section 2.3. As in section 2.3, audio recordings of items used as stimuli in the perceptual experiment and those used for the acoustic analyses were converted into sequences of 39-dimensional feature vectors consisting of 13 Mel-frequency cepstral coefficients (MFCCs), with delta and delta-delta coefficients. In contrast to our previously used features, here our first coefficient did not correspond to the log of the total frame energy, but to the zeroth cepstral coefficient. Since the zeroth coefficient corresponds to the sum of the log of the 40 mel values, it is roughly equivalent to the log energy. We applied this change purely due to the change in the tools used for computing features.[5] Additionally, we added 3 coefficients (and their corresponding delta and delta-delta coefficients) adding pitch information to our features: normalized-pitch, delta-pitch, voicing-feature. The final 48 coefficient values were standardised to have zero-mean and unit-variance within coefficient and within each speaker.[6] While pitch features and delta and delta-delta coefficients were computed, their use in the model was evaluated according to whether performance was better or not during parameter optimisation (see below).

2.4.2.2 Classification

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 to the exemplars recorded by native speakers of BP and JP. We simulated the perception experiment by classifying each template n times, n being the number of total valid trials for that template in the perceptual experiment. Details of the classifier are given below.

Dynamic Time Warping As in section 2.3, we computed the optimal alignment between all $V_1C_1C_2V_2$ templates (e.g. /abgi/) and their corresponding $V_1C_1V_3C_2V_2$ epenthesized versions (e.g. /abigi/, /abogi/, /abugi/) using Dynamic Time Warping (DTW) [Sakoe and Chiba, 1978] with the R package *dtw* [Giorgino, 2009]. 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 feature frames corresponding to C_1C_2 clusters and $C_1V_3C_2$ sections. Note, however, that coarticulation cues from flanking vowels are expected to be present within the clusters.

As input for the DTW distance computation at each speech frame, we used either the entire 48-dimensional feature vectors (MFCCs + pitch features + delta + delta-delta), or we omitted pitch features and/or delta + delta-delta coefficients. The final selection of the features to be used for our models was determined during parameter optimisation.

[5] As a reminder, however, due to a mistake when computing features in section 2.3, those unconventional features consisted of the log energy, the first 12 MFCCs (including the zeroth coefficient) and the corresponding deltas and delta-deltas.

[6] Normalisation had not been done within speakers in the previous version of the model, as we aimed to have acoustics-based models with the least amount of abstraction.

Concerning DTW specifics, in section 2.3 we used the commonly used step pattern for which, at position $x_{i;j}$, the only possible steps are towards positions $x_{i+1;j}$ (horizontal step), $x_{i;j+j}$ (vertical step), or $x_{i+1;j+1}$ (diagonal step). In this default setting (named "symmetric2" in the R package *dtw*), a diagonal step is twice as costly as a horizontal or a vertical step, which favours template-query matches with compressions/stretching over more direct matches. Therefore, in this section we chose to opt for a step pattern with the same three possible steps as before (i.e., horizontal, vertical, diagonal), but for which diagonal steps cost as much as horizontal/vertical steps on the final DTW distance. Since distances obtained with this step pattern ("symmetric1") cannot be normalised by being divided by the sum of the lengths of the template and query, we normalise by dividing the cummulative distance by the length of the optimal path.

From the DTW we extract two values per template-query combination: (1) DTW_{dist}, the normalised DTW distance between the template and the query, and (2) DTW_{time}, the proportion of non-diagonal steps taken in the optimal path. This latter value, which was not present in the previous version of the model, is an indicator of the proportion of time dilation and time compression that was required to match the template and the query.

Similarity function We use the same similarity function as in section 2.3, inspired by the exemplar-based generalized context model (GCM) detailed by [Nosofsky, 1992], and make modifications to accommodate for the inclusion of the duration mismatch penalty DTW_{time}. Our goal is to classify the $V_1C_1C_2V_2$ template into a category from the set of vowels /i, o, u/, through the similarity of the template to exemplars $V_1C_1iC_2V_2$, $V_1C_1oC_2V_2$, and $V_1C_1uC_2V_2$, respectively. We obtain $P(R_J|S_i)$, the evidence favouring category J given stimulus i, by averaging the similarity of said stimulus i to all recorded exemplars of category J. Since we do not aim to introduce a language model to the exemplar model, as we want it to be based entirely on acoustics, we do not introduce a term for response bias for category J. All exemplars are weighted equally.

$$P(R_J|S_i) = \frac{\frac{1}{n_J}\sum_{j \epsilon C_J} \eta_{ij}}{\sum_K \frac{1}{n_K}\sum_{k \epsilon C_K} \eta_{ik}} \tag{2.2}$$

where n_J is the number of exemplars of category J and with η defined as in equation 2.3,

$$\eta_{ij} = e^{-c \cdot d_{ij}} \tag{2.3}$$

where c is a parameter determining the weight of η_{ij} on classification. When $c = 0$, DTW is disregarded and all possible classification categories are equally probable. Higher values of c result in higher sampling probabilities being given to items with smaller distance η_{ij}.

η_{ij} is defined as in equation 2.4

$$\eta_{ij} = DTW_{dist} + \alpha \cdot DTW_{time} \tag{2.4}$$

where α is a scaling factor for DTW_{time}. Setting $\alpha = 0$, gives an equation equivalent to the one used in section 2.3.

2.4.2.3 Parameter estimation

We used grid search in order to optimise parameter c for each language, as well as the format of our acoustic features. We classified all $V_1C_1V_3C_2V_3$ exemplars from the BP and JP recordings in section 2.3 to a category within /i, o, u/ in a leave-one-out cross-validation method. Namely, we assessed the accuracy in the classification of these tokens with known labels, using the classification procedure described above:using DT, we measured their

respective similarity to all other exemplars with the same $V_1C_1 - C_2V_3$ skeleton, then for each we sampled the classification label from the three possible categories, weighting the sampling probability by the average distance to exemplars of the three categories.

We assessed the optimality of parameter values based not only on mean classification accuracy, but also by inspecting median classification accuracy. Indeed, while these two measures are positively correlated, an improvement in median accuracy might not be obvious when inspecting mean accuracy alone. In other words, combinations of parameters with similar mean classification accuracy could differ greatly in median classification accuracy. As a curious side note, while we aimed to optimise parameters when setting $\alpha = 0$ (baseline models with no duration penalty), increasing the value of α during cross-validation with BP and JP recordings decreased classification accuracy. However, the resulting degradation in accuracy tended towards zero when increasing values of parameter c such as those chosen for this work.

We found that classification accuracy was worse when including pitch features, delta + delta-delta coefficients, and both, than when only using 13 MFCCs. Therefore, for all experiments in this section, the acoustic input given to our models consisted of 13-dimensional vectors. Concerning parameter c, we chose optimal values $c = 60$ (mean accuracy: 64.6%; median accuracy: 1) for BP models, and $c = 40$ (mean accuracy: 90.1%; median accuracy: 1) for JP models. This constitutes an improvement compared to the previous version of the model in section 2.3 (50% and 63% mean accuracy for BP and JP, respectively).

2.4.2.4 Data analysis

Statistical analyses were performed with the R statistical software [R Core Team, 2016], using Markov chain Monte Carlo generalised linear mixed-models [Hadfield, 2010, Plummer et al., 2006]. These Bayesian models sample coefficients from the posterior probability distribution conditioned on the data and given priors. We used priors that are standard for mixed-effects multinomial models. Model convergence was assessed by visual inspection of trace plots and the Gelman–Rubin convergence diagnostic [Gelman and Rubin, 1992], using four chains with different initialisations. Effects were considered statistically significant if the 95% highest posterior density (HPD) interval estimated for the coefficient of interest did not include zero. We report both the posterior mode and the 95% HPD interval.

In order to assess the influence of V_1 and V_2 (henceforth: flanking vowels) on epenthetic vowel quality (/i/ or /u/), we chose as fixed effects for our models Language (BP *vs.* JP, sum contrast coded) and Number of Same Flanking Vowels (NSFV; considered as a continuous variable with values 0, 1, or 2 instead of a factor with 3 levels, in order to reduce the number of model parameters and promote convergence), as well as their interaction. As random intercepts we included Cluster and Participant when analysing data from the perceptual experiment, and Cluster when analysing data from the exemplar models. We also added random slopes for Language on Cluster, and NSFV on Participant. The change in statistical models with respect to the previous section was motivated by a will to avoid coefficient inflation due to the sparsity of our data.[7] Because of these changes, we reanalysed the behavioural results using the same statistical model before analysing results from the exemplar models. This allowed a fairer comparison between effects observed in real and simulated datasets. However, we did not expect these results to be qualitatively different from those in section 2.3.

[7] Statistical models in section 2.3 had as fixed factors Language, NSFV, Coronal, all interactions, and random intercepts for Participant (for the perceptual experiment only).

2.4.3 Results

2.4.3.1 Re-analysing results from the perception experiment

/i/-epenthesis The left panel of Figure 2.16 shows the proportion of /i/-epenthesis for human participants, with data collapsed by C_1C_2 cluster[8]. We report the results from our statistical analyses below, even though they are qualitatively equivalent to those presented in section 2.3. We found a significant main effect of LANGUAGE (mode: -6.33, HPD: $[-7.94, -4.40]$), which reflects the fact that BP participants perceived an epenthetic /i/ more often than JP participants. The main effect of NSFV was also significant (mode: 3.72, HPD: $[3.30, 4.27]$); participants epenthesized /i/ more often when more flanking vowels were /i/. The interaction between the fixed effects LANGUAGE x NSFV was not significant (mode: -0.24, HPD: $[-1.15, 0.76]$).

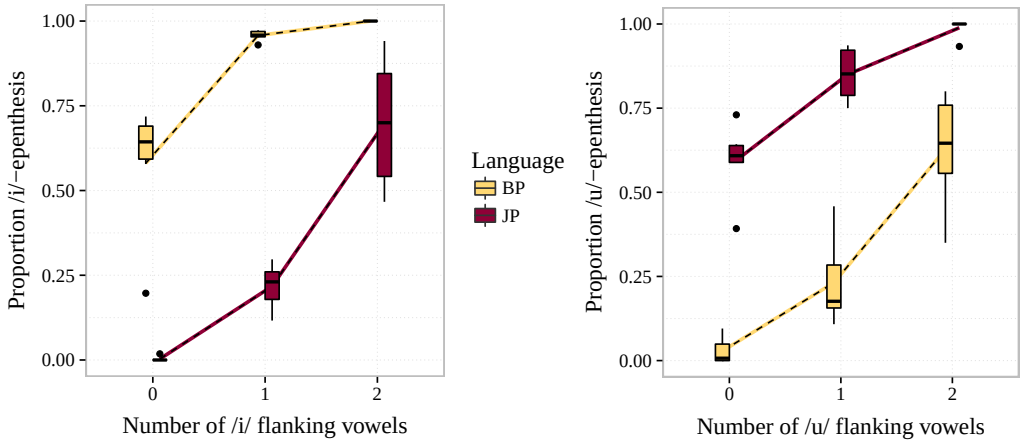

Figure 2.16: *Proportion of /i/-epenthesis (left) and /u/-epenthesis (right) exhibited by BP and JP participants in the perception experiment. The box and whiskers plots display the distribution of proportions across C_1C_2 clusters (median, quartiles and extrema). Dashed lines connect mean values. This representation was preferred over showing distributions across participants, in order to have a direct visual representation of what our statistical models are evaluating, as well as to have the same amount of plotted datapoints for participants and models. Data plotted collapsed by participant can be seen in Figure 2.9*

/u/-epenthesis The proportion of /u/-epenthesis for human participants can be seen on the right panel of Figure 2.16. As for /i/-epenthesis, we report our results but they are qualitatively equivalent to results from section 2.3. We found a significant main effect of LANGUAGE (mode: 5.06, HPD: $[3.47, 6.41]$), reflecting the higher rates of /u/-epenthesis for JP participants compared to BP participants. The main effect of NSFV was also significant (mode: 2.35, HPD: $[2.02, 2.74]$); more /u/ flanking vowels yielded higher rates of /u/-epenthesis. The interaction between the fixed effects LANGUAGE x NSFV was not significant (mode: -0.57, HPD: $[-1.30, 0.18]$).

2.4.3.2 Exemplar models without duration penalty

We used our new exemplar models to simulate the perceptual experiment, setting $\alpha = 0$, effectively setting up models with no duration-mismatch penalty. The resulting classifica-

[8]Please note that, while based on the same data as Figure 2.8, here the datapoints correspond to clusters, not participants.

tion patterns can be seen in Figure 2.17.

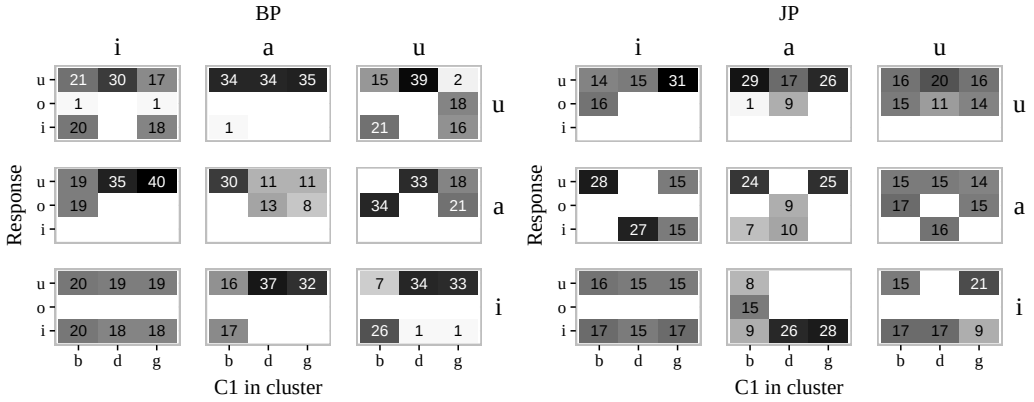

Figure 2.17: *Responses given by exemplar models with no duration-mismatch penalty ($\alpha = 0$) for BP (left) and JP (right). Numbers indicate trial counts, with darker cell backgrounds representing higher values. Within each 3 x 3 grid, trials are separated according to V_1 (columns) and V_2 (rows). Within each individual rectangle, the cluster C_1 is given by the horizontal axis, while the vertical axis corresponds to response categories. For instance, the BP model yielded /u/-epenthesis on 37 trials involving /adC_2i/ items.*

/i/-epenthesis The proportion of /i/-epenthesis given by these models is shown in Figure 2.18. The main effect of LANGUAGE was not significant (mode: 2.85, HPD: $[-1.48, 8.70]$); the models do not appear to reproduce the higher rates of /i/-epenthesis for BP than JP. We did, however, find a main effect of NSFV (mode: 2.04, HPD: $[1.78, 2.35]$); as for human participants, more /i/ flanking vowels result in higher rates of /i/-epenthesis by exemplar models with no duration penalty. The interaction between the fixed effects LANGUAGE x NSFV was not significant (mode: -0.26, HPD: $[-0.91, 0.25]$).

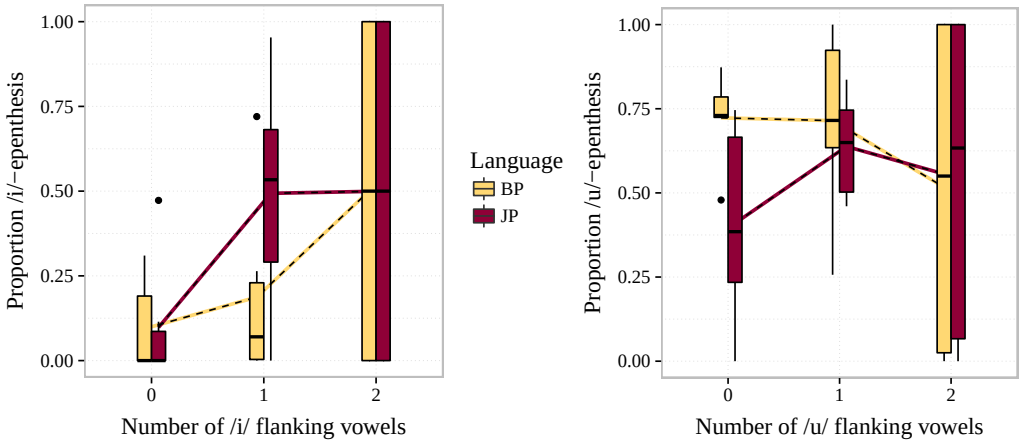

Figure 2.18: *Proportion of /i/-epenthesis (left) and /u/-epenthesis (right) exhibited by BP and JP exemplar models with no duration penalty. The box and whiskers plots display the distribution of proportions across C_1C_2 clusters (median, quartiles and extrema). Dashed lines connect mean values.*

/u/-epenthesis The right panel of Figure 2.18 shows the proportion of /u/-epenthesis for exemplar models with no duration penalty. Neither the main effect of LANGUAGE (mode: -1.60, HPD: $[-4.36, 1.16]$) nor the main efect of NSFV (mode: 0.15, HPD: $[-0.05, 0.34]$) were significant; the models do not appear to reproduce the higher rates of /u/-epenthesis for JP than BP, and they do not show significantly higher rates of /u/-epenthesis with more /u/ flanking vowels in general. However, the interaction LANGUAGE x NSFV was significant (mode: 0.96, HPD: $[0.57, 1.34]$). We therefore performed supplementary analyses to examine the effect of NSFV for each language independently. Using the R package *lme4* [Bates et al., 2015], for each language we fitted a generalised linear mixed model (GLMM) with a declared binomial dependent variable (/u/-epenthesis) with NSFV as the sole fixed effect and CLUSTER as a random effect. We assessed significance through model comparison with a null model without the main effect NSFV. For both the BP and the JP models, we found NSFV to be significant but with opposing effects; while the JP model yielded more /u/-epenthesis with more /u/ flanking vowels ($\beta = 0.52$, $SE = 0.12$, $z = 4.48$, $p < 0.001$), the BP model yielded less /u/-epenthesis with increasing numbers of /u/ flanking vowels ($\beta = -0.44$, $SE = 0.12$, $z = -3.75$, $p < 0.05$).

2.4.3.3 Exemplar models with duration penalty

Adding the duration-mismatch penalty In this experiment our aim was to examine, first of all, the effect of an increased duration-mismatch penalty on our models' ability to mirror human performance at the perceptual task. Remember that, when performing parameter estimation with BP and JP items, increasing the weight of the duration-mismatch penalty DTW_{time} resulted in a decrease in classification performance. For our optimal values of parameter c, when changing from $\alpha = 0$ to $\alpha = 5$, this difference was of 0.9% and 7.8% in mean classification accuracy, and 0 and 0.1% in median classification, for BP and JP, respectively. Apart from α, all model parameters (c, feature coefficient selection) are set as for models without a duration-mismatch penalty.

In order to assess the effect of varying α when simulating our non-native speech perception experiment, we computed the distance between response patterns given by human participants and models, for each item used in the experiment, while varying the value of α. We did this by computing the Euclidean distance between $[h_i, h_o, h_u]$ and $[m_i, m_o, m_u]$, vectors containing the proportion of /i, o, u/ responses given by humans and models, respectively, within each experimental item. We normalised distances in order to constrain their values to the interval $[0, 1]$ (0 corresponding to identical response patterns). The variation of the distance between patterns as a function of α can be seen in Figure 2.19.

Contrary to what we observed when classifying BP and JP items during parameter optimisation, we observe that increasing the weight of DTW_{time} increases the similarity between model and human responses until a certain value, after which the average similarity decreases. In order to examine the best case scenario for models that value duration match between templates and queries, we selected optimal α values for BP and JP, based on the aforementioned response pattern similarity. Therefore, it should be noted that we select the best possible model for each language, and it is this fitted model that we will later analyse. Parameter values which minimise the distance between human responses and model responses were $\alpha = 8$ for BP, and $\alpha = 3$ for JP. The classification patterns obtained with these parameter values can be seen in Figure 2.20. We now turn to our main questions: Do these models better reflect default epenthetic vowel choice and flanking vowel influence than models based solely on spectral features and without duration-mismatch penalties?

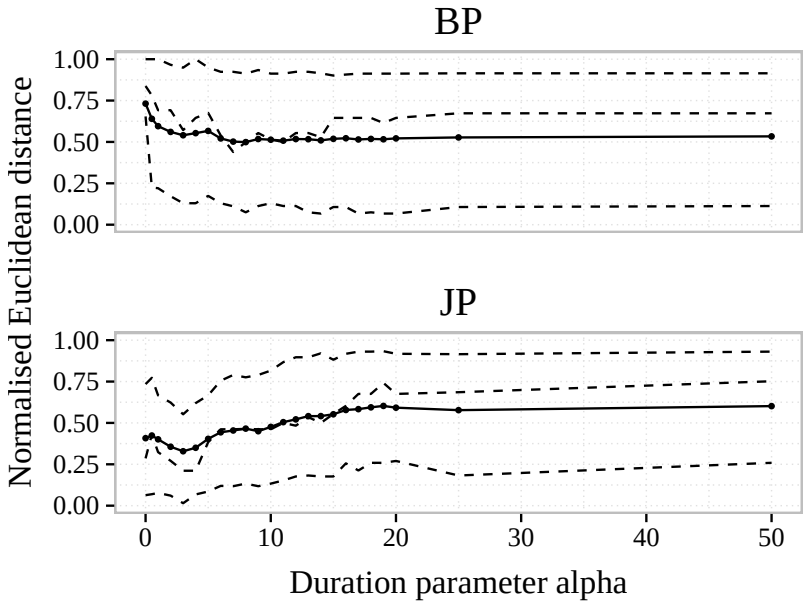

Figure 2.19: *Similarity between human and model responses for varying values of the duration-mismatch parameter α. Solid lines display the mean similarity, dashed lines display the distribution of proportions across items (median and quartiles).*

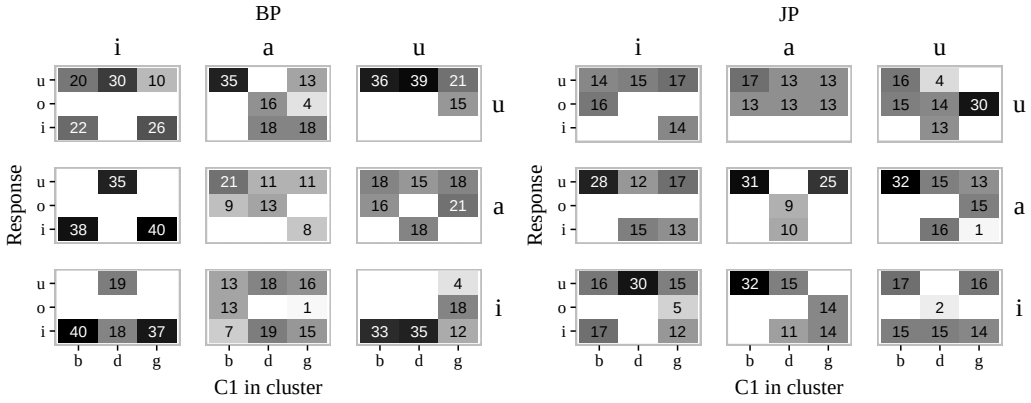

Figure 2.20: *Responses given by exemplar models with added duration-mismatch penalty for BP (left) and JP (right). Numbers indicate trial counts, with darker cell backgrounds representing higher values. Within each 3 x 3 grid, trials are separated according to V_1 (columns) and V_2 (rows). Within each individual rectangle, the cluster C_1 is given by the horizontal axis, while the vertical axis corresponds to response categories. For instance, the JP model yielded /u/-epenthesis on 28 trials involving /ibC_2a/ items.*

/i/-epenthesis The proportion of /i/-epenthesis given by these models is shown in the left panel of Figure 2.21. The main effect of LANGUAGE was not significant (mode: -0.97, HPD: $[-6.02, 2.40]$); the models do not appear to reproduce the higher rates of /i/-epenthesis for BP than JP. Note, however, the change in sign of the posterior mode and the shift in HPD interval towards more negative values, compared to values found for LANGUAGE when there was no duration penalty (mode: 2.85, HPD: $[-1.48, 8.70]$), reflecting the increase in /i/-epenthesis for the BP model (relative to JP) with the addition

of the duration penalty. If we look at the plot in Figure 2.21 it might seem surprising that the main effect of LANGUAGE is not significant. However, note that the interaction LANGUAGE x NSFV was significant (mode: -1.02, HPD: $[-1.50, -0.47]$). The higher rates of /i/-epenthesis for the BP model relative to the JP model was estimated to be due to a greater effect of flanking vowel for the former. Additionally, we found a significant main effect of NSFV (mode: 2.08, HPD: $[1.80, 2.31]$); as for human participants, more /i/ flanking vowels resulted in higher rates of /i/-epenthesis by exemplar models with a duration penalty.

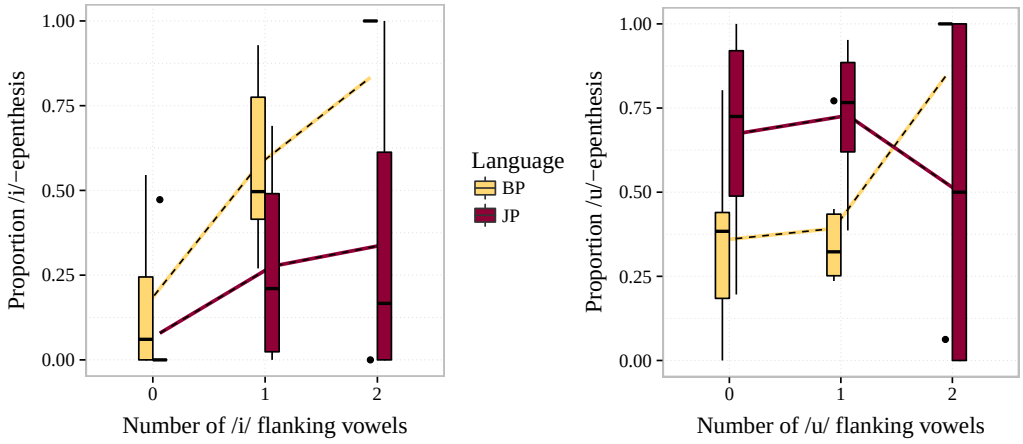

Figure 2.21: *Proportion of /i/-epenthesis (left) and /u/-epenthesis (right) exhibited by BP and JP exemplar models with duration penalty. The box and whiskers plots display the distribution of proportions across C_1C_2 clusters (median, quartiles and extrema). Dashed lines connect mean values.*

/u/-epenthesis The proportion of /u/-epenthesis given by models with duration penalty is shown in the right panel of Figure 2.21. The main effect of LANGUAGE was significant (mode: 2.38, HPD: $[0.20, 4.36]$); models with duration penalty seem to reproduce the higher rates of /u/-epenthesis for JP than BP. The main effect of NSFV was also significant (mode: 0.45, HPD: $[0.23, 0.61]$); models with duration penalties reproduced the tendency to epenthesize /u/ more often with more /u/ flanking vowels. The interaction LANGUAGE x NSFV was also significant (mode: -1.21, HPD: $[-1.61, -0.86]$).

2.4.4 Discussion

In this study we enhanced our production-based exemplar models of nonnative speech perception from section 2.3. Notable improvements include basic speaker adaptation, through speaker-specific standardisation of acoustic features. The newer version of the model, therefore, is not purely acoustic as it previously was. We also included the possibility to add a duration-mismatch penalty to be taken into consideration when computing the similarity between an item to be classified and exemplars used by the model for classification.

Because the statistical analyses were not identical to those in section 2.3, we reanalysed data from the perceptual experiment for a fairer comparison to the most recent version of our models. As expected, we found that BP listeners epenthesise more /i/ than JP listeners, while the opposite pattern is true with /u/-epenthesis. Participant responses are modulated by the quality of flanking vowels; more neighbouring /i/ and /u/ vowels lead to more /i/- and /u/-epenthesis, respectively, by BP and JP participants.

First we simulated the perceptual experiment with models that did not include a duration-mismatch penalty. The models were able to reproduce modulations of epenthetic vowel quality brought by flanking vowels, but they did not reproduce the higher prevalence of default epenthetic vowels. We then examined models with a duration-mismatch penalty. We did find that models with a non-null duration-mismatch penalty gave response patterns closer to that of human participants than when the penalty was absent. We then evaluated the best possible models that did incorporate a duration-mismatch penalty. As models with no duration-mismatch penalty did, these models exhibited higher rates of /i/- and /u/-epenthesis with more /i/ and /u/ flanking vowels, respectively. We also found an overall higher prevalence of /u/ for the JP model relative to the BP model, but the opposite situation with /i/ did not occur. However, the BP model seemed to be more "sensitive" to the effect of /i/ flanking vowels on /i/-epenthesis. Adding a duration-mismatch penalty appears to better approximate response patterns given by human participants in a task probing perceptual epenthesis.

One question that may arise is why adding the duration-mismatch penalty during the parameter optimisation stage (i.e., classification of BP/JP recordings in a leave-one-out method) did not increase classification accuracy if it did help approximating epenthesis patterns. It is difficult to pinpoint the exact reason. One possibility is that this is due to a difference in what "gold standards" for classification labels were in the two cases. During the experiments, a nonnative item was classified into one of three categories (/i, o, u/) by human participants. The classification labels correspond to what participants report hearing. However, for parameter optimisation, we used as labels what had been read by participants when recording the items. While the recordings were transcribed by a trained phonetician, it is not guaranteed that that label corresponds to what BP/JP listeners would report hearing. As a reminder, our items are nonwords, some of them with phonotactically illegal sound sequences (e.g., /aduga/ contains /du/, which is not phonotactically legal in Japanese). It is possible that BP and JP listeners might show variability in their classification of such stimuli.

In contrast, coming back to our experimental results, models that used duration mismatches seem to yield response patterns closer to those given by human participants, with flanking vowel modulations still being reproduced qualitatively. Yet even the best possible models were not able to correctly reproduce the prominence of default vowel epenthesis observed in psycholinguistic experiments. We expected duration mismatch to play an important role in the emergence of the default epenthetic vowel, since we found in section 2.3 that short duration is a characteristic shared by default vowels in JP and BP. We hypothesised that this short duration contributed to the insertion of these vowels being phonetically minimal. It is important to note that our models' processing of duration can be further improved. Indeed, our models assess the percentage of DTW steps involving time dilation or contraction necessary to optimally align two items. Therefore, our models evaluate duration in absolute terms. It would be interesting for duration to be evaluated as something relative instead. For instance, we could take speech rate differences into consideration, or modulate vowel choice by also looking at how probable is the resulting duration of the consonants given the insertion of a vowel (i.e., consonant duration being allocated to said vowel). Modifications such as these would involve a higher level of abstraction than the one provided by our exemplar models, which simply compare two sequences of acoustic frames. Notably, they would require the introduction of discrete units (e.g., phonemes) when processing the acoustic input. Being aware of these limitations, we will later turn towards a very different family of models that allow for these modifications.

2.5 Conclusions

Similarly to what has been previously observed (e.g., [Dupoux et al., 1999, Dehaene-Lambertz et al., 2000, Dupoux et al., 2011, Monahan et al., 2009, Mattingley et al., 2015]), we find that BP and JP participants experience perceptual vowel epenthesis when presented auditory stimuli with illegal consonant clusters. For most cases, BP and JP participants insert what has been called a "default" epenthetic vowel, namely /i/ and /u/, respectively. Confirming intuitions laid out by [Dupoux et al., 2011] and reminiscent of the P-map theory in [Steriade, 2001], acoustic measurements revealed that these vowels were, within their respective languages, acoustically minimal in that they were of shorter duration and closer in formant space to cluster transitions than other candidate vowels.

In section 2.3 we were able to reproduce the effect of coarticulation on epenthetic vowel quality observed in [Dupoux et al., 2011], but this time using naturally produced stimuli. We find that the quality of epenthetic vowel is modulated by the identity of flanking vowels: we see more /u/-epenthesis by BP participants with more /u/ flanking vowels, and more /i/-epenthesis by JP participants with more /i/ flanking vowels. Were these modulations of epenthetic vowel quality due to coarticulation cues contained in the consonant clusters or were they due to a phenomenon of vowel copy based on phonological features as proposed by [Rose and Demuth, 2006, Uffmann, 2006]?

This question was addressed in section 2.2, by assessing the perception of stimuli for which the identity of the flanking vowels was in disagreement with that of the coarticulation cues contained within consonant clusters. It was found that, while both flanking vowel and coarticulation influenced epenthetic vowel quality, it was the latter that was the most determinant. This is reflected by the results of the exemplar models evaluated in sections 2.3 and 2.4. Indeed, these models compared the acoustics of non-native *CC* clusters to native *CVC* exemplars, in order to determine the quality of the vowel to be epenthesized. And while they were unable to mimic default vowel epenthesis, they *were* able to reproduce quality modulations due to neighbouring vowels. Yet, these models could not perform vowel copying in a way other than by exploiting coarticulation remnants within the clusters.

In section 2.3 we provided evidence that default epenthetic vowels are phonetically minimal both spectrally and at the level of duration. Yet we were not able to find evidence that these acoustic cues are sufficient for default epenthetic vowels to emerge, since our models were not able to mimic default epenthetic vowels. As previously stated, our proof of concept exemplar model is very limited, as it performs pure acoustic matching between *CVC* queries and a *CC* template. We showed that the models were not able to reproduce default vowel epenthesis even when taking duration into consideration or/and when adding basic speaker normalisation.

Adding to these concerns, it is important to mention that the model supposes the existence of "multiphonemic" (i.e., sequences of phonemes) exemplars to which the non-native input is compared to. Leaving aside the fact that the use of exemplar representations during speech perception may be controversial, an important side-effect of exemplar-based models is that they are unable to model lack of epenthesis. Yet we saw in both sections 2.2 and 2.3 that participants did choose the "no epenthesis" option in a non-negligible percentage of the trials. As such, in the next chapter we will focus on perception models that are flexible enough to also output illegal structures, as our participants do. Using these models we will continue investigating whether information readily available from the acoustic signal (i.e., phonetics) are sufficient to explain epenthetic vowel quality, or, rather, whether information relative to the frequency of sounds or sound combinations are necessary as well.

As a final reminder, our results are, as those by [Dupoux et al., 2011], better aligned

with one-step theories of non-native speech perception than with two-step theories. Indeed, the influence of acoustic details on epenthetic vowel quality would be lost if epenthesis occurred after an initial categorisation step; computation of the optimal output must therefore incorporate acoustics and phonotactics in a unique step. Following all of these considerations, in the future chapters we will switch from a one-step DTW-based exemplar model of non-native speech perception to more elaborate one-step Hidden Markov Models (HMM).

Chapter 3

Modelling speech perception with ASR systems

3.1 Introduction

3.2 Anatomy of our HMM-based speech recogniser

For the experiments described in this chapter we used Hidden Markov Model (HMM)-based speech recognisers as models of human perception. Speech audio waveforms are transformed into a sequence of **acoustic vectors** $X_{1:T} = x_1, ..., x_T$ through the first step, called feature extraction. In the decoding phase that follows, the trained ASR system attempts to find the sequence of words $w_{1:L} = w_1, ..., w_L$ which is most likely to have produced the sequence of acoustic feature vectors X. Mathematically, this equates to solving the following equation:

$$\widehat{w} = \underset{w}{arg\,max} \{P(w|X)\} \tag{3.1}$$

Put into Bayesian terms, this represents computing the posterior probability $P(w|X)$ of all combinations of words, given the acoustics, and retrieving the index of the most probable one: the decoder selects the sequence of words w with highest posterior probability given the acoustic evidence X. However, due to the fact that there is an infinite number of possible combinations of words, it may be difficult to evaluate all possible combinations in order to find the most probable one. As such, following Bayes's theorem, equation 3.2 can be rearranged as:

$$\widehat{w} = \underset{w}{arg\,max} \{P(X|w)P(w)\}\,^{1} \tag{3.2}$$

The likelihood $P(X|w)$, given by the **acoustic model**, is the probability of the acoustics given the sequence of words. The prior $P(w)$, given by the **language model**, corresponds to the probability of the word sequence, and can be derived from frequency counts. These probabilities can be extracted by training our ASR system using annotated speech corpora.

Nowadays, in the field of ASR, Neural Network (NN)-based speech recognisers are the state-of-the-art. In spite of better performance of NN-based ASR systems, we decided to use HMM-based recognisers, which are better understood than NNs while still offering

[1] In practice $P(X|w)P(w)$ is often computed in the log space as $\alpha\ logP(X|w) + logP(w)$, where α is a scaling factor called acoustic scale (set to 0.1 in our models)

good speech recognition performance. Indeed, before the decoding step, these models offer a clear separation between the acoustic model (AM; i.e., mapping between phoneme categories and acoustics) and the language model (LM; i.e., frequencies of word/phoneme sequences). This allowed us to test ASR systems with different LMs while keeping the AM constant, as well as adapting LMs to mimic the experimental paradigms used when testing human participants. Importantly, unlike NNs which are often qualified as "black boxes", we are able to better understand and analyse how the ASR system processes acoustic input. We will now present the necessary components for building an HMM-based ASR system, namely the speech corpora used to train and test the system and its featural representation, as well as the decoder itself, composed of an acoustic model, a lexicon, and a language model. The interaction of these elements is depicted in Figure 3.1. In the following subsections we will present the components in more detail.

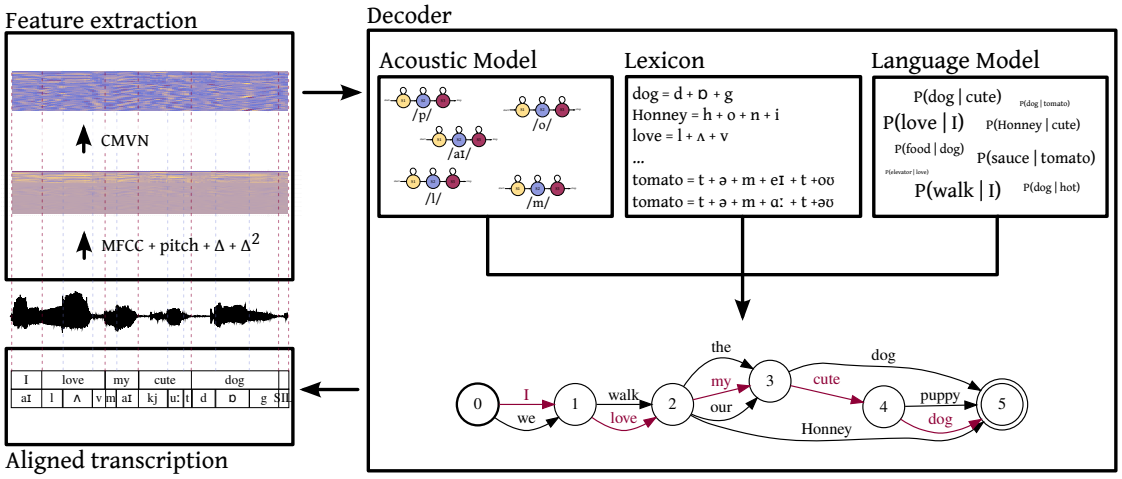

Figure 3.1: *Architecture of our ASR system, including its input (acoustic features) and output (transcription).*

3.2.1 Corpora

In order to train and test our ASR system, we required transcribed speech corpora. These corpora consisted of speech recordings which have been annotated; for each utterance, we have a more or less detailed transcription of what was said. While the ideal annotation is one for which phoneticians have provided phoneme categories (or even phones), as well as their boundaries, often we might only have access to by-utterance annotations where we are only provided with a sequence of words/phonemes for each utterance. In these cases, we rely on forced alignment to automatically find phoneme boundaries.

In the following sections we have trained ASR systems with different "native" languages, namely Japanese (JP) and Korean (KR). These languages were of particular interest because of their relatively restrictive phonotactics with regards to consonant clusters, as well as the availability of corpora of spontaneous speech, which we will now present. We also trained an American English (EN) corpus in order to evaluate our model's performance with respect to state-of-the-art systems.

Corpus of Spontaneous Japanese (CSJ) As the name suggests, the CSJ [Maekawa, 2003] contains recordings of spontaneous Standard Japanese. The corpus is composed of two subparts: (1) academic presentation speech (APS), which consists of live recordings of

academic presentations, and (2) simulated public speech (SPS), where speakers presented everyday topics in front of a small audience. For our models we only kept SPS, which is more representative of everyday conversations at the level of the lexicon, and has a more balanced population than the young, male-dominated APS. Recordings were manually transcribed by native speakers of Japanese using Japanese syllabaries, which meant that the phonetic transcriptions only included phonotactically legal phoneme sequences, even in cases where the actual acoustics might have been closer to illegal sequences. Phoneme boundaries were manually adjusted; however, this alignment was not used when training our models, as it was overwritten by forced alignment due to technical constraints. Our subset of the corpus contained 400,547 utterances produced by 594 speakers[2] (331 female, 263 male), with an average of 674.3 utterances per speaker. The division of the corpus across training, validation, and test set are shown in Table 3.1.

Table 3.1: Datasets used for training and evaluating the Japanese ASR system with the CSJ.

	Proportion	# Utterances	Duration (hh:mm:ss)	# Speakers
train	80%	322,208	152:26:33	475
valid	5%	19,566	9:12:03	119
test	15%	58,773	27:19:14	119

Korean Corpus of Spontaneous Speech (KCSS) The KCSS [Yun et al., 2015] consists of recordings of spontaneous Seoul Korean. Forty speakers aged 10 to 49 (5 female speakers and 5 male speakers per decade) were recorded in a quiet room, for approximately 1 hour each. Speech was ellicited through questions related to the speakers' personal opinions, habits, acquaintances, etc. Recordings were manually transcribed by native speakers of Korean. We used phonetic transcriptions faithful to actual pronunciations which, for instance, include phonetic reduction (akin to *yesterday* being transcribed as /jɛʃeɪ/ instead of the canonical /jɛstɚdeɪ/). The transcription process involved the use of the main writing system of Korean (i.e., hangul) as well as a romanization, meaning that there is a possibility that acoustic sequences closer to phonotactically illegal sequences might have been transcribed as phonotactically legal counterparts. Transcriptions include manually adjusted phoneme boundaries, as well as word syllabification; however this alignment was not used when training our models, as it was overwritten by forced alignment due to technical constraints. The corpus contains 57,504 utterances produced by 40 speakers (as explained above), with an average of 1,437.6 utterances per speaker. The division of the corpus across training, validation, and test sets is shown in Table 3.2.

Wall Street Journal - Read (WSJ) The WSJ [Paul and Baker, 1992] is a corpus of both read and spontaneous American English. For our work, we only kept the read subset of the corpus, which consisted of professionally trained journalists recorded while reading news articles. Contrary to the CSJ and KCSS, the recordings were not phonetically transcribed. However, we had access to the news articles themselves, as well as to

[2]For the CSJ and KCSS, we used utterances from the same speakers for the validation and test sets, but their data was not seen during model training. For the WSJ, data from all speakers was used in the 3 corpus subsets, due to a planned comparison to another corpus not described here. Since the speakers that we used in our experiments are not from any of the corpora, this is not an issue. However, it needs to be kept in mind that error rates (%WER and %PER) for KCSS, CSJ, and WSJ are only comparable within corpus.

Table 3.2: Datasets used for training and evaluating the Korean ASR system with the KCSS.

	Proportion	# Utterances	Duration (hh:mm:ss)	# Speakers
train	80%	46,208	18:58:15	32
valid	5%	2,824	1:16:39	8
test	15%	8,472	3:54:15	8

a dictionary which mapped the standard phonetic pronunciation of words in American English to the words in the articles. In total, 338 speakers (X female, Y male) uttered 71,037 utterances, with an average of 210.2 utterances per speaker. The division of the corpus across training, validation, and test sets is shown in Table 3.3.

Table 3.3: Datasets used for training and evaluating the American English ASR system with the WSJ corpus.

	Proportion	# Utterances	Duration (hh:mm:ss)	# Speakers
train	80%	56,872	115:18:46	338
valid	5%	3,661	7:24:22	338
test	15%	10,504	21:12:19	338

3.2.2 Features

In order for our ASR systems to be able to use speech as input, it is necessary to perform signal analysis. This procedure transforms the continuous raw speech waveform into sequential speech features. This latter form ensures a more more informative representation of speech, with modifications that enhance phonemic contrasts and better approximate how speech is processed by the human cochlea. In this work we used Mel-frequency cepstrum coefficients (MFCC), traditionally used for HMM-based ASR systems.

Speech is recorded with a microphone; the continuous audio signal is digitalized at a sampling rate of 16kHz. The audio is then segmented into frames of 25 ms, with a shift of 10 ms between the beginning of each frame. By using frames, we make the assumption that the signal is stationary within the 25 ms window, and we apply the following proccessing steps to each frame, using Kaldi [Povey et al., 2011]:

1. Pre-processing: The data is extracted and pre-processed (dithering, pre-emphasis, and DC offset removal).
2. Windowing: The data in the 25 ms frame is multiplied by a tapered window (Hamming window), to avoid discontinuities at the edges of the segment.
3. Spectral analysis: By applying a Fast Fourier Transform (FFT), we find out how much energy there is at each frequency band for this frame.
4. Nonlinear frequency scaling: In order to compensate for the fact that human hearing is less sensitive to higher frequencies, frequencies are mapped onto a Mel scale, which is linear until approximately 1000 Hz and logarithmic afterwards. This is done by applying a mel-filter bank with 23 bins, which are equally spaced in the mel-frequency domain. Each filter summarises the amount of energy in a section of the range of frequencies.

5. Cepstral analysis: The log of the energy in each bin is computed, from which we take the cosine transform. We keep 13 MFCCs, including c_0, the zeroth coefficient which represents the average of the log-frequency of the bins [Gales et al., 2008].

6. Cepstral liftering: Coefficients are scaled, ensuring that they have a reasonable range.

We therefore obtain 13 MFCCs that summarise the information at each frame of audio. To these coefficients, we add 3 coefficients carrying information about pitch: normalized-pitch, delta-pitch, voicing-feature[3]. To these 16 static features we add their respective dynamic features (Δ and Δ^2) that describe the evolution of the coefficient values over time. Coefficient values are then standardised using Cepstral Mean Variance Normalisation (CMVN); for each speaker the distribution of each coefficient's values has a mean value of zero and a variance of one.

3.2.3 Acoustic model

Now that we have extracted the acoustic features for the labelled utterances in our corpus, we are able to train the acoustic model (AM). Recall that the AM gives the likelihood $P(X|w)$, which corresponds to the probability of the acoustics given the sequence of words w. In order to simplify things, let's not view an utterance as a sequence of words which are sequences of phonemes themselves, but directly as a sequence of phonemes. Then, we consider the probability of the acoustics X given the sequence of phonemes W. The acoustics corresponding to a given phoneme change during the duration of the phoneme; as such, phones are not static objects but they should be described as having acoustic trajectories. By using Hidden Markov Models (HMM), we can approximate these trajectories as sequences of static states. A priori, the more states, the better the approximation to the real data. However, empirically it has been assessed that having three states is a good compromise for ASR systems. Following this, we chose to model phonemes as three-state HMMs, where the states correspond, respectively, to the beginning, middle, and end portions of the phoneme. This is particularly relevant for phonemes that can be viewed as sequences of discrete articulatory events with distinct acoustic signatures, such as plosives (e.g., /p/) which are often described as an airway closure, followed by a period of occlusion and a possibly audible release. Additionally, the separation into three states allows to account for the fact that the acoustics of the beginning and end of a phoneme may be differently affected by neighbouring phonemes (i.e., coarticulation) in comparison to the medial part.

As their name suggests, HMMs follow a Markovian process; the value of a state only depends on the value of the previous state. The transitions between states are defined by transition probabilities not only between adjacent states, but also within a state itself (i.e., self-loops). These transition probabilities are defined during AM training, based on the transitions between frames in the training corpus. While the duration of phonemes cannot be explicitly learned by the acoustic model, they are implicitly reflected by the transition probabilities in the self-loops: for a given state, the higher the self-loop probability, the longer the model will "remain" at said state and the longer the sequence of acoustic vectors assigned to the corresponding phoneme. A simplified illustration of a phoneme HMM is shown in Figure 3.2.

[3]Information about pitch was added because of its contrastive relevance in Japanese at the lexical level (i.e., pitch accent) and in Korean at the phonemic level (e.g., tonogenesis in the three-way contrasts of plosives). In practice, adding pitch features resulted in a slight improvement of model performance in Japanese (from 41.3% WER to 39.6%; acoustic model with 6000 Gaussians).

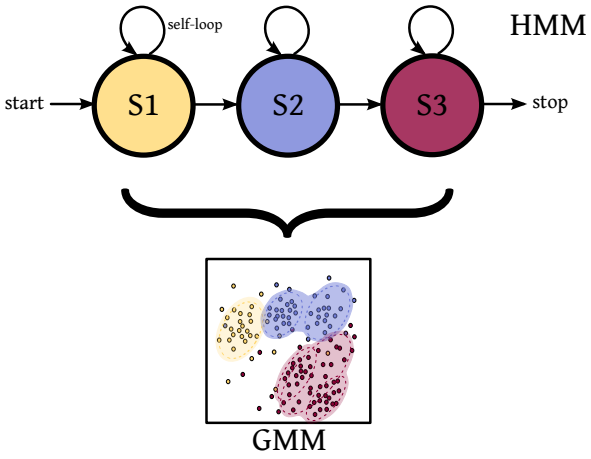

Figure 3.2: *Left-to-right 3-state phoneme HMM with simplified 2-dimensional GMM (one per state). Start and stop states connect the phoneme with the previous and next phoneme, respectively.*

In sum, each phoneme is modelled by a left-to-right 3-state HMM. But what exactly is a state? Our acoustic models are HMM-GMMs, where GMM stands for Gaussian Mixture Models. For each phoneme our 48-dimensional feature vectors define a 48-dimensional space where GMM for the three states are embedded. During training, the acoustic model will have placed individual acoustic frames on this space, based on the values of their feature vector. In other words, acoustic frames from each phoneme portion will occupy a certain part of this space. For each state, we can parametrically define the space covered using mixtures of Gaussian distributions (the aforementioned GMMs). Indeed, GMMs are universal approximators of densities when given enough components. To do this, the model will have fitted a number of diagonal Gaussian distributions to approximate the distribution of datapoints corresponding to each phoneme state. The number of Gaussians allocated to each phoneme state depends on the total number of Gaussians made available to the model, and the complexity of the distribution of the frames in the space. Once that the GMMs are defined, the AM is able to tell us, for any new frame, the likelihood that the frame originated from each GMM (i.e., phoneme state).

3.2.3.1 Why not triphones?

If the reader is already familiar with ASR systems, they may expect us to go a step further and no longer treat phonemes as units for the HMMs (i.e., monophone acoustic models) but, instead, use context-dependent triphones. In this latter representation, an independent three-state HMM is built for each phoneme within a phonemic context. With some simplifications, this equates to no longer having an HMM for the phoneme /p/, but having all context-dependent versions of this phoneme as individual HMMs (e.g., the triphone /p$_{a_i}$/, which is the phone /p/ when preceded by /a/ and followed by /i/). Traditionally, triphone-based HMM-based ASR systems perform better than monophone systems. However, these more complex models are inappropriate for our experiments. Recall that we aim to use these speech recognition systems as models of nonnative speech perception, using tasks analogous to paradigms used in psycholinguistic experiments (namely, identification/forced-choice tasks). Importantly, we are focusing on modelling perceptual vowel epenthesis. This situation excludes the use of triphones because, by definition, our

ASR systems will have to decode speech that does not follow native phonotactics. Decoding such stimuli implies the existence of triphones corresponding to the input, yet the model will have never encountered such triphones in the training data. While this situation might seem analogous to what listeners may experience, one must consider the fact that the ASR system *will attempt to account for said triphones* during decoding in spite of the lack of data. Importantly, poorly estimated, phony triphones (e.g., /h$_{a_p}$/, when decoding /ahpa/) will be put up against well-estimated triphones (e.g., /h$_{a_a}$/) during the forced-choice tasks. The well-estimated triphones might simply be preferred as transcriptions over poorly-estimated ones for this reason alone, irrespective of the actual acoustic match between the stimuli and phoneme models. In order to increase the performance of monophone models at phonetic labelling tasks such as ours, it is possible to increase the number of total Gaussians available to the model [Saraclar, 2001].

3.2.4 Lexicon & language models

As shown in Figure 3.1, the acoustic model is combined with two other components in order to decode speech: the Lexicon and the Language Model (LM).

The lexicon is, put simply, a pronunciation dictionary. It links the acoustic model (i.e., phoneme-level HMMs) with the language model, which is at the word level. For each word, we indicate in the dictionary the sequence of phonemes that constitute it. It is also possible to account for multiple pronunciations of a word due to dialectal differences (e.g., "tomato" pronounced as /təmɑːtəʊ/ or /təmeɪroʊ/), phonological phenomena (e.g., homorganic assimilation: "handbag" /hændbəg/ → /hæmbəg/), or supresegmental information (e.g., stress contrasts: "record" /'rekord/ (noun) vs. /re'kord/ (verb)).

At the word level, the language model specifies $P(W)$, the probability of occurrence of word sequence W. For this we use n-grams: we approximate the probability of a sequence

$$P(W) = P(w_1)P(w_2|w_1)P(w_3|w_1, w_2)...P(w_L|w_1, w_2, ..., w_L) \tag{3.3}$$

by using the product of the probability of the component words, each conditioned on the *n-1* words preceding it. For instance, if $n = 2$, we obtain a bigram model, where the LM specifies the probability of a word depending on a single preceding word. The probability of the word sequence W can then be approximated as:

$$P(W) \approx P(w_1)P(w_2|w_1)P(w_3|w_2)...P(w_L|w_{L-1}) \tag{3.4}$$

In our case, these probabilities are obtained from word counts in the training corpus as follows:

$$P(w_i|w_j) \approx \frac{c(w_i, w_j)}{c(w_i)} \tag{3.5}$$

where $c(w_i, w_j)$ is the number of observations of w_i followed by w_j, and $c(w_i)$ is the total number of occurrences of w_i. Since not all word combinations are bound to appear in the training corpus, smoothing is performed; null probabilities are given a small probability of appearing.

Additionally to the bigram word LM, we computed a unigram phone LM in order to evaluate our models' ability to do phonetic decoding. In this case, the lexicon is identical to the phoneme inventory and the LM consists of phoneme counts.

3.2.5 Decoding

When decoding speech, the ASR system builds a graph containing candidate word sequences that can serve as transcription for the audio input, based on the acoustic model, the lexicon, and the language model. In order to keep the problem computationally

tractable, only the most likely transcription hypotheses are kept; this is known as pruning.

The output of the decoding step is not a single transcription but what is called a lattice. In this graphical representation only the most probable transcriptions are included, with weighted paths connecting words (a minimalistic example without weights can be seen in Figure 3.1). The weight of each path is determined by the product of the acoustic and language model scores (derived from $P(X|w)$ and $P(W)$, respectively). The final score for each possible transcription is obtained by summing all the weights of the path that need to be crossed to reach the sequence of words in the transcription ("I love my cute dog", in the example in Figure 3.1). Having access to lattices means that we are not only able to derive the most probable transcription; we can extract the n-best transcriptions, each with its corresponding alignment, and the total acoustic and language model scores.

3.2.6 Scoring: Assessing native performance

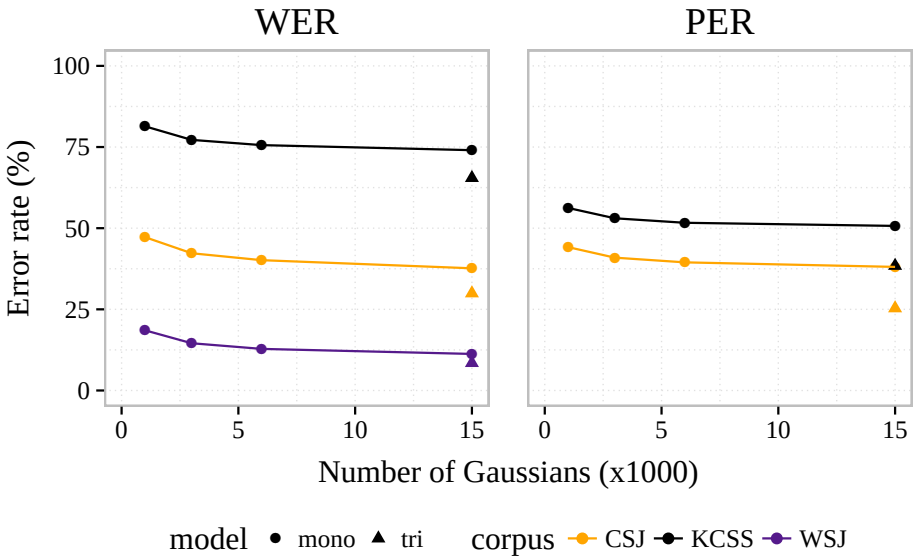

Figure 3.3: *Changes in word error rate (%WER) and phone error rate (%PER) following variation of the number of total Gaussians allocated to monophone acoustic models (circles). The error rates obtained with a triphone model with* 15,000 *Gaussians are included as comparison (triangles). Scores correspond to decoding performed on the validation set (i.e., unseen speakers for the CSJ and KCSS; already seen speakers for the WSJ).*

We tested the decoding performance on the validation set of AMs with total number of Gaussians going from 1,000 to 15,000. These values are used as default total number of Gaussians when training, respectively, monophone and triphone models in Kaldi (without speaker adaptation). In order to do so, we used the language models described in section 3.2.4, namely a word bigram LM and a phone unigram LM, which were used to obtain word error rates (%WER) and phone error rates (%PER), respectively. Note that while we provide word bigram WER% as a reference value to compare with existing speech recognition models[4], our main focus is on PER%. Indeed, we will use our models in

[4]As a reference, in 2015 some state-of-the-art speaker adapted HMM-GMM systems trained on 82 hours of the WSJ achieved 6.3% WER [Panayotov et al., 2015] and 5.4% WER

paradigms involving phonetic decoding of non-native nonwords; the phone unigram PER% evaluation gives us an insight into how well our ASR systems can do phonetic decoding on native (non)words.

As seen in Figure 3.3, we find that the performance of our models increased (i.e., error rates decreased) when increasing the number of total Gaussians from the Kaldi default of 1,000 to 15,000, which would average to approximately 125 Gaussians per state for a language with an inventory of 35 phonemes[5]. Therefore, the acoustic models with the highest amount of Gaussians (i.e., 15,000) give the best performance for monophone models, both at the lexical (%WER) and phonetic (%PER) levels of decoding. We did not pursue increasing the number of Gaussians even further, as performance gain was reaching an asymptote at this point and adding more Gaussians would have increased the computational demands for each experiment. Additionally, we expect that adding "too many" Gaussians might have lead to overfitting of the models to the training set. As expected, triphone models performed better than monophone models at phonetic decoding (%PER), in spite of having the same amount of total Gaussians than our best monophone models (CSJ: 37.96% monophone *vs.* 25.33% triphone; KCSS: 50.70% monophone *vs.* 38.42% triphone). Later in this chapter we will discuss how it might be possible to increase acoustic model performance in future work, without having recourse to triphone HMMs, which as explained previously are not appropriate for our experiments.

Concerning the test set (i.e., 15% of the corpora), we find %WER comparable to those obtained for the validation set (CSJ: 37.48% on test, 37.64% on validation; KCSS: 73.35% on test, 74.03% on validation), and similarly for %PER (CSJ: 37.96% on test, 38.07% on validation; KCSS: 50.70% on test and validation). Since the validation and test sets contain utterances from the same speakers, this information does not allow us to evaluate our models' sensitivity when decoding data from speakers not used in the training data (recall that none of our models have any speaker adaptation; only CMVN is applied when processing the features). However, the fact that validation and test set scores are similar indicates that, while rudimentary, our acoustic models give stable performances when confronted with datasets with structurally different lexical exemplars and acoustically different phonetic exemplars.

[TODO]: Add WSJ

[Chan and Lane, 2015] on the WSJ eval'92 dataset. Contemporary deep neural network-based systems achieved 3.5% WER on the same dataset [Chan and Lane, 2015].

[5]Phoneme counts: CSJ: 37, KCSS: 36, WSJ: 39

3.3 Investigating surface phonotactics

ADD ACKNOWLEDGEMENTS THOMAS + EMMANUEL.

3.3.1 Introduction

[TODO]

3.3.2 Experiment 1

In this experiment we investigated how various versions of our ASR model differing in their language models (LMs) compared to real behavioural data. While we varied the LMs, the acoustic model was kept constant; as stated in the previous section, we used HMM-GMM monophone models with 15000 Gaussians. We used our models to perform simulations of the identification task described in section 2.2, where Japanese listeners were asked to indicate whether they heard an epenthetic vowel within the consonant cluster of $V_1C_1C_2V_1$ items (e.g., /ahpa/). For these items, the quality of the coarticulation cues either matched or mismatched the quality of the flanking vowels. We analysed the results quantitatively in order to assess if injecting additional phonotactic information allowed the model to better approximate human responses. We also performed qualitative analyses in order to see if the best version of the model reproduced the effects observed in section 2.2.

3.3.2.1 Methods

Stimuli We used the same stimuli as in section 2.2. As a reminder, we recorded 3 speakers producing disyllabic $V_1C_1C_2V_1$ and trisyllabic $V_1C_1V_2C_2V_1$, with V_1 a flanking vowel in the set /a, e, i, o, u/, C_1 /h/ or /k/, and C_2 a fixed consonant, /p/ (e.g, /ahpa/, /ahapa/). By cross-splicing the disyllabic natural control items (e.g., /ahpa/), we obtained disyllabic spliced control items (e.g., /ah$_a$pa/), disyllabic spliced test stimuli (e.g., /ah$_u$pa/), and trisyllabic spliced fillers (e.g., /ahapa/), where subscripts indicate the identity of the vowels flanking the clusters in the original recording. Therefore, within each speaker, all stimuli of the same structure (in our example, /ah(V)pa/ items) have acoustically identical flanking vowels.

Language models In order for the decoding task to be analogous to the behavioural experiment described in section 2.2, trial-specific language models were constructed, as shown in Figure 3.4. Thus, when decoding a $V_1C_1(V_2)C_2V_1$ stimulus, the perception model was only given the possibility to transcribe it as $V_1C_1(V_2)(SIL)C_2V_1$, where phones between parentheses are optional, V_2 was from the set of vowels /a, e, i, o, u/, and SIL is an optional silence.

In this section, we investigate the type of phonotactic information that might be used by Japanese listeners when perceiving foreign speech that does not conform to native phonotactics. We test 5 types of language models (LM) when decoding our $V_1C_1(V_2)C_2V_1$ items; these LMs differ only in the weights given to edges between nodes 2 and 3 in the graph shown in Figure 3.4. The weights were obtained by computing frequency counts from the portion of the CSJ used for training the acoustic model. Using the same acoustic model, we compared the following LMs [TODO] Add formulae as annex:

1. A null LM, which implies that listeners base their decoding of consonant clusters on phonetic match alone, without using information on phonotactics.

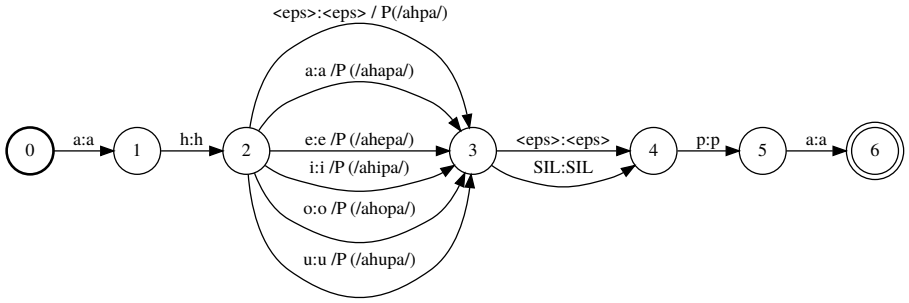

Figure 3.4: *Constrained language model used to test the models (here: LM for /ahpa/ trials). Nodes in the graph represent states, weighted edges represent transitions between states (here: phonemes). When relevant, weighted edges are labeled with the probability to choose that edge when decoding, which affects the final language model score of each possible path. When no weights are shown (e.g., between states 3 and 4), there is no preference between the paths concerned. The language model scores are combined with acoustic scores when decoding experimental items.*

2. A phone-unigram LM, which implies that listeners do not take neighbouring phonemes into consideration when decoding the consonant clusters; only the frequency of the vowel V_2 to be epenthesized (compared to that of C_2) is taken into account when choosing epenthetic vowel quality.

3. An online phone-bigram language model, which implies that listeners decode the clusters as they hear them (i.e., decoding is done from the start of the item), and the choice of (no) vowel is conditioned on the presence of C_1. Therefore, the choice of epenthetic vowel is modulated by C_1V_2 and C_1C_2 diphone frequencies.

4. A retro phone-bigram language model, which implies that listeners decode the clusters based on the most recent information (i.e., decoding is done from the end of the item), and the choice of (no) vowel is conditioned on the presence of C_2. Thus, the choice of epenthetic vowel is modulated by $V_{ep}C_2$ and C_1C_2 diphone frequencies.

5. A batch phone-bigram language model, which implies that listeners decode the item considering the entire structure, taking into consideration the probability of having a vowel V_2 given the presence of C_1 and C_2. Here the choice of epenthetic vowel is modulated by the product of C_1V_2 and V_2C_2 (or by C_1C_2) diphone frequencies.

Identification task simulation After decoding the stimuli, we extracted from the resulting lattice each possible transcription of each item, and the corresponding acoustic and language model scores. An example of how the ASR system decodes the experimental stimuli can be seen in Figure 3.5. From the (scaled) acoustic and language model scores we derived the item posteriorgrams, which indicate how probable a given transcription was given the audio input. We used these probabilities as proxies of the probability that a listener might exploit when performing reverse inference during speech perception, and therefore, the probabilities used when responding in an identification task.

As such, for each item, we obtained a six-dimensional vector $ident_{model} = [p_{none}, p_a, p_e, p_i, p_o, p_u]$, containing a discrete probability distribution, with a probability mass function linking the identification task options 'none', 'a', 'e', 'i', 'o', 'u', to their respective probabilities (i.e., posteriorgrams). We can define the human equivalent $ident_{human} = [p_{none}, p_a, p_e, p_i, p_o, p_u]$, which contains the percentage of responses for each item, after aggregating all participant responses.

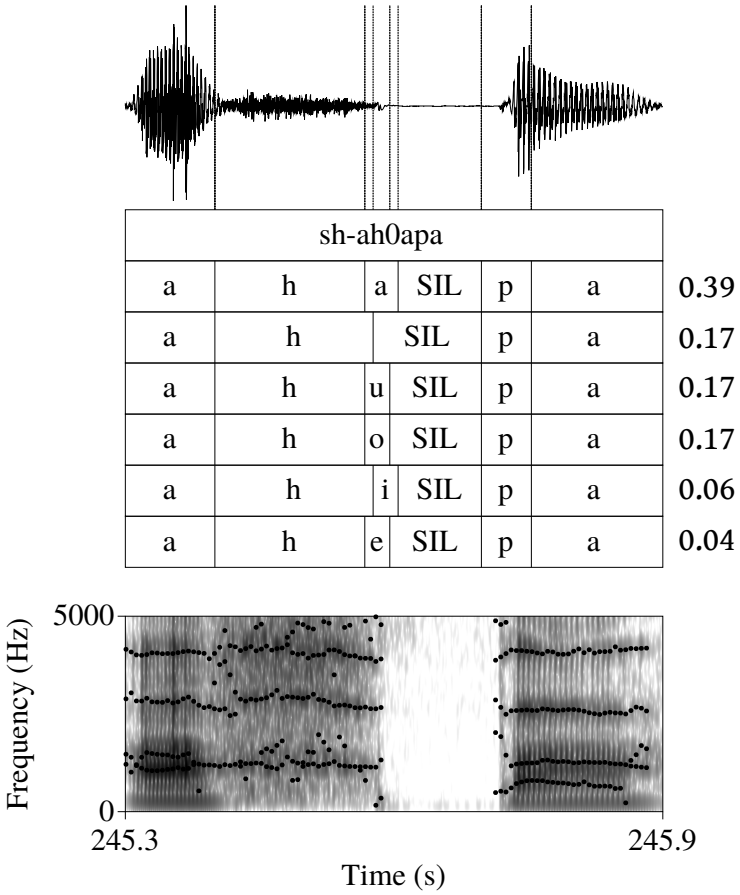

Figure 3.5: *Example of how the ASR system decodes the item /ah$_a$pa/, using the null version of the language model in Figure 3.4. From top to bottom: original waveform, item name, aligned transcriptions given by the model (from the most probable to the least probable, with the corresponding posteriorgrams shown to their right side), and spectrogram with formant contours. SIL = silence. TODO: get updated figure (this is with mono-6K)*

3.3.2.2 Quantitative analysis

In order to perform a global evaluation of the similarity between the behavioural responses and the responses obtained with the five LMs described above, we computed the Pearson's product-moment correlation coefficient between the human and model posteriorgrams. The model with the highest correlation to the human data was the NULL LM ($r = 0.76$), followed by the UNIGRAM, BIGRAM ONLINE, and BIGRAM RETRO LMs ($r = 0.65$), and lastly, the BIGRAM BATCH LM ($r = 0.62$). Numerically, the NULL LM better approximated the human data.

In order to assess if the correlation differences between the NULL LM and other LMs were significant, we computed these differences and their corresponding 95% confidence intervals (CIs), using bootstrapping with 1000 samples[6]. As can be seen in Table 3.4, the correlation between the human data and the output of the NULL LM was significantly higher than those of other LMs.

Contrary to the null and unigram LMs, the bigram models were subject to an arbitrarily set smoothing parameter, which determined the probability of choosing a sequence

[6]Sampling was done by item.

Table 3.4: *Difference in correlation with human data between the null LM and other LMs. The lower and upper bounds of the 95% confidence intervals are given between brackets. Positive values indicate higher correlation between human data and null model output than between human data and other LM output.*

	Correlations	Difference	Significant?
null vs. unigram	$0.76 - 0.65$	$0.10\ [0.07, 0.14]$	Yes
null vs. bigram online	$0.76 - 0.65$	$0.11\ [0.08, 0.14]$	Yes
null vs. bigram retro	$0.76 - 0.65$	$0.10\ [0.07, 0.13]$	Yes
null vs. bigram batch	$0.76 - 0.62$	$0.13\ [0.10, 0.17]$	Yes

of phonemes that had never been observed in the training data. We set this smoothing parameter to 10^{-8}, which is a strict value, as it is relatively close to zero. This was done in order to evaluate whether the acoustic match could rescue decoding options which are not supported by the language's phonotactics. In order to evaluate the similarity between models' outputs and human data without the influence of the value of the smoothing parameter, we computed the correlation between the human data and models' posteriorgrams after excluding the posteriorgrams for "none" responses and re-normalising the remaining posteriorgrams. As such, we are focusing on the correlation related to epenthetic vowel quality. Here, the highest correlation still corresponded to the NULL LM ($r = 0.77$), followed by the BIGRAM RETRO ($r = 0.74$), the BIGRAM ONLINE ($r = 0.73$), and finally the UNIGRAM and the BIGRAM BATCH LMs ($r = 0.71$). As shown in Table 3.5, while the difference between the correlations diminished relative to what is shown in Table 3.4, the CIs still did not overlap with zero, meaning that the correlation between the human data and the output of the NULL LM was significantly higher than those of other LMs.

Table 3.5: *Difference in correlation with human data between the null LM and other LMs, after removing the "none" responses. The lower and upper bounds of the 95% confidence intervals are given between brackets. Positive values indicate higher correlation between human data and null model output than between human data and other LM output.*

	Correlations	Difference	Significant?
null vs. unigram	$0.77 - 0.71$	$0.05\ [0.03, 0.08]$	Yes
null vs. bigram online	$0.77 - 0.73$	$0.04\ [0.02, 0.06]$	Yes
null vs. bigram retro	$0.77 - 0.74$	$0.02\ [0.01, 0.04]$	Yes
null vs. bigram batch	$0.77 - 0.71$	$0.06\ [0.03, 0.08]$	Yes

3.3.2.3 Qualitative analyses

Identification accuracy Using the set of filler items such as /ahapa/ and /okipo/ (i.e., spliced items with a full vowel between C_1 and C_2), we can assess identification accuracy relative to our item labels. Indeed, recall that while our phonetically-trained speakers were instructed to read items following "standard" IPA pronunciations, it is possible for our human participants to not perceive the intended vowel categories due to adaptation processes (e.g., misperceiving /u/, which is not realised as [u] but as [ɯ] in Japanese), and/or due to speaker idiosyncrasies.

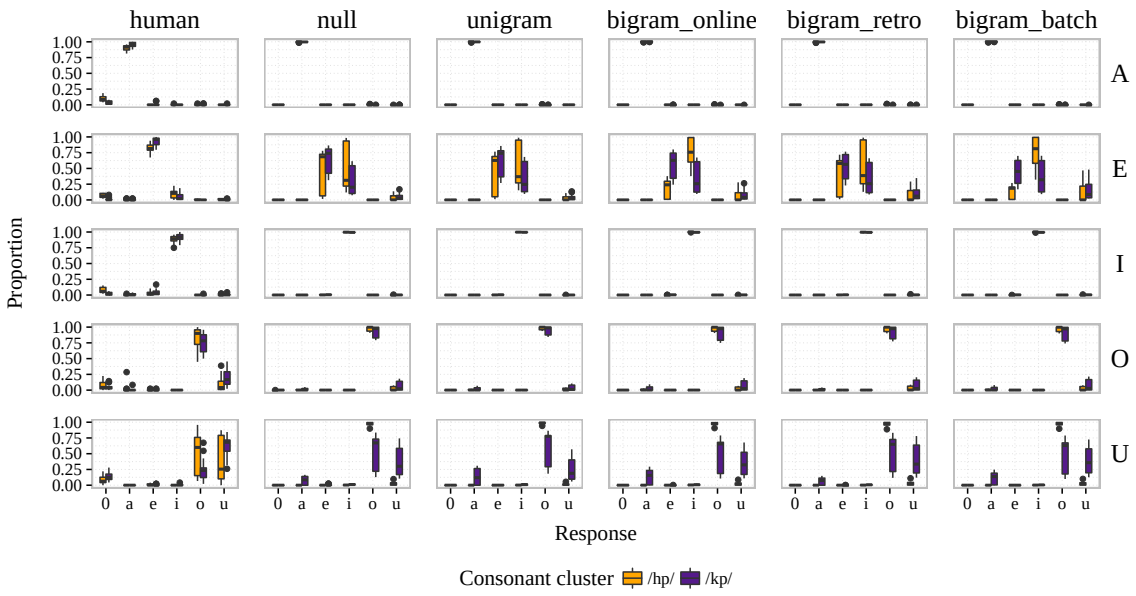

Figure 3.6: *Response patterns for the identification task on full vowel stimuli (filler items). Human and models responses are separated by columns; intended identity of the medial vowel is given by rows. Within each rectangle, the horizontal axis corresponds to possible responses from the set {"none", "a", "e", "i", "o" , "u"}. The vertical axis corresponds to proportion of responses (i.e., posteriorgrams, in the case of models). The box and whiskers plots display the distribution of the proportions across items (median, quartiles and extrema). For instance, we can see from the first row that, similar to humans, all models correctly classified most /VCapV/ items as containing a medial /a/ vowel.*

Overall, human participants identified the correct intended vowel category in 79.5% of the trials. As can be seen in Figure 3.6, this was mostly due to confusions between the intended /o/ and /u/ categories, with most errors being /u/ being identified as /o/. Consistent with our correlation analyses, the NULL LM gave the highest accuracy out of all models (accuracy: 74%), followed by the BIGRAM RETRO (accuracy: 72.6%), the UNIGRAM (accuracy: 72.5%), the BIGRAM ONLINE (accuracy: 72.6%), and finally the BIGRAM BATCH LM (accuracy: 68.7%). As seen in Figure 3.6, like human participants, the models showed difficulty categorising /u/ items as such; in particular, these were almost always classified as exemplars of /o/ when $C_2 = /h/$. However, unlike human participants, models misperceived /e/ as /i/. This misperception appears to be worse for the BIGRAM ONLINE and BIGRAM BATCH language models. In sum, while the accuracy rates for the models were close to that of humans, we saw that the models showed not only quantitative but also qualitative differences in their identification of non-native full vowels, in comparison with human participants. Below we continue our qualitative analyses on what was determined to be the best model according to the correlation analyses and the filler item accuracy, namely the ASR system with a NULL LM.

Control items Human participants experienced vowel epenthesis in 56% (/hp/: 52%; /kp/: 61%)[7] of control items in which the flanking vowel and coarticulation are of the same

[7]Note that since posteriorgrams are computed by weighting items from all three speakers are equally, values reported in this section might differ slightly from those in section 2.2. Indeed, due to how data was cleaned in section 2.2, some trials were removed and the number of trials per item

quality. The NULL LM gave an output of 68% epenthesis, with 72% and 65% epenthesis for /hp/ and /kp/, respectively. As such, the model gave a higher percentage of epenthesis for /hp/ clusters compared to /kp/clusters, while the opposite was true for humans.

Now we focus on epenthetic vowel quality, meaning that we perform analyses after removing "none" responses and re-normalising posteriorgrams. We find that, like humans (/hp/: 44%; /kp/: 87%; total: 66%), the model gave lower percentages of default /u/-epenthesis for /hp/ (47%) than /kp/ (65%) clusters (total: 56%). However, this difference is not as marked as it is for humans. Recall that in section 2.2 we found that humans experiences significantly more default /u/-epenthesis for /kp/ clusters, while experiencing significantly more vowel copy epenthesis for /hp/ clusters. These patterns of responses mirrored loanword data. Do we find these effects in the output of our model?

We first examined possible effects of consonant cluster on default /u/-epenthesis by using the R statistical software [R Core Team, 2016], using Markov chain Monte Carlo generalised linear mixed-models [Hadfield, 2010, Plummer et al., 2006]. These Bayesian models sample coefficients from the posterior probability distribution conditioned on the data and given priors. We used priors that are standard for linear models. Model convergence was assessed by visual inspection of trace plots and the Gelman–Rubin convergence diagnostic [Gelman and Rubin, 1992], using eight chains with different initialisations. Effects were considered statistically significant if the 95% highest posterior density (HPD) interval estimated for the coefficient of interest did not include zero. We report both the posterior mode and the 95% HPD interval.

The left panel of Figure 3.7 shows the posteriograms of /u/-epenthesis for humans and all models. For the ASR system with the null LM, we assessed the variation of the continuous response variable "u" response POSTERIORGRAM that was caused by the fixed effect CONSONANT cluster (/kp/ *vs.* /hp/; contrast coded with deviance coding). We initially included random intercepts for SPEAKER and ITEM, as well as a random slope for SPEAKER on CONSONANT. However, these were removed as their addition caused the models to be singular (estimated null variances), with consequently poor trace plots. We found the main effect of CONSONANT to be significant (mode: -0.19, HPD: $[-0.29, -0.07]$), meaning that as for humans (mode: -0.42, HPD: $[-0.61, -0.26]$), the model gave significantly more /u/-epenthesis for /hp/- than /kp/-clusters. However, as evidenced by the statistical model coefficients, the magnitude of the effect is larger for humans than for the model.

Turning to vowel copy epenthesis in control items for which the flanking vowel was not /u/, we used the same statistical models but with copy vowel POSTERIORGRAM as the continuous response variable. For instance, for the item /ek$_e$pe/, this was the posteriorgram for the "e" response. The distribution of posteriorgrams for humans and all models is shown in the right panel of Figure 3.7. While there was a trend in the same direction for the null LM, namely higher percentages of vowel copy for /hp/- than /kp/-clusters, we did not find a significant main effect of CONSONANT for the model (mode: 0.11, HPD: $[-0.02, 0.24]$) as we did for humans (mode: 0.39, HPD: $[0.20, 0.58]$).

Test items Next we examine the identification task response patterns for test items. As a reminder, for these spliced items, the vowel coarticulation was different from the flanking vowels.

As shown in Figure 3.8, reponses that were represented the most in the null model posteriorgrams were "none" (%), "i" (%), and "u" (%). These were also the responses that human participants gave the most ("none" (%), "i" (%), and "u" (%)).

per speaker might have differed in some cases. In order to ensure that human and model data are comparable, we re-do statistical analyses of human data when necessary and report the resulting coefficients.

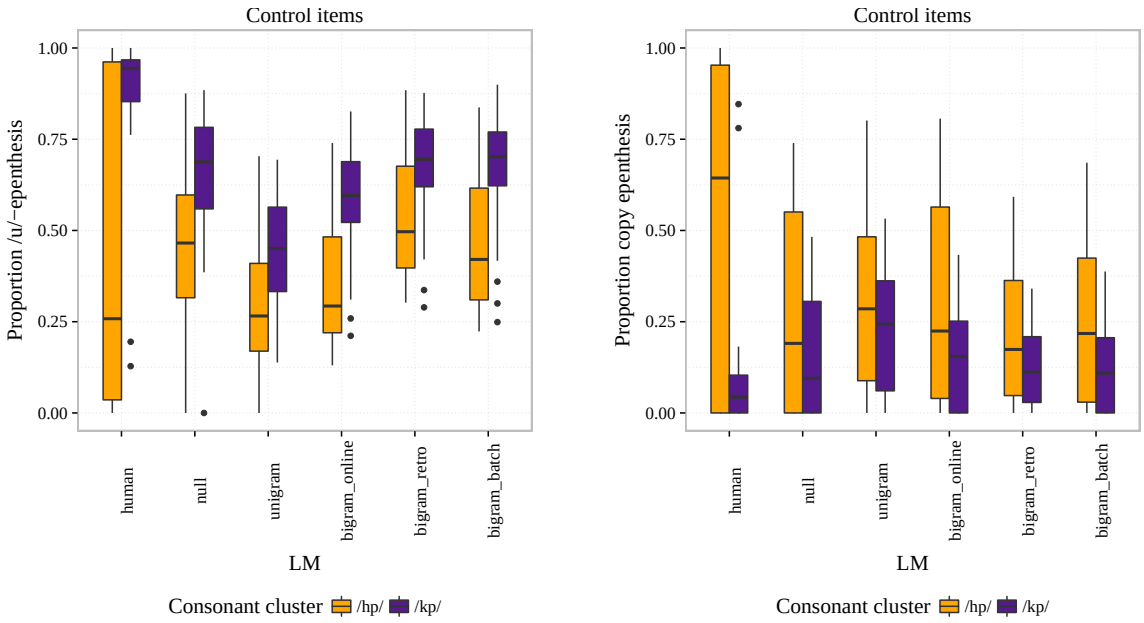

Figure 3.7: *Proportion of default /u/-epenthesis (left) and vowel copy epenthesis (right) given by human participants and models. The box and whiskers plots display the distribution of the proportions across items (median, quartiles and extrema).*

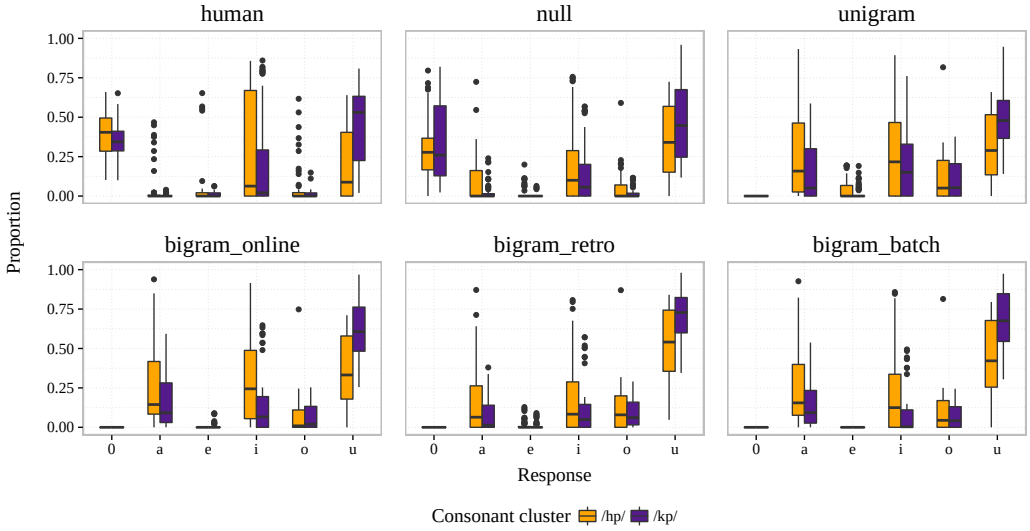

Figure 3.8: *Response patterns for the identification task on spliced test stimuli. Horizontal axes correspond to possible responses from the set {"none", "a", "e", "i", "o" , "u"}. Vertical axes correspond to proportion of responses (i.e., posteriorgrams, in the case of models). The box and whiskers plots display the distribution of the proportions across items (median, quartiles and extrema).*

We saw in section 2.2 that, for human participants, responses were mainly determined by the quality of the vowel coarticulation within the consonant cluster. This manifested itself in the appearance of horizontal bars, and some very faint vertical bars, in the top panels of Figure 3.9. Do we observe something similar in the output of the model with null LM?

When examining the bottom panels in Figure 3.9, we see that response patterns are

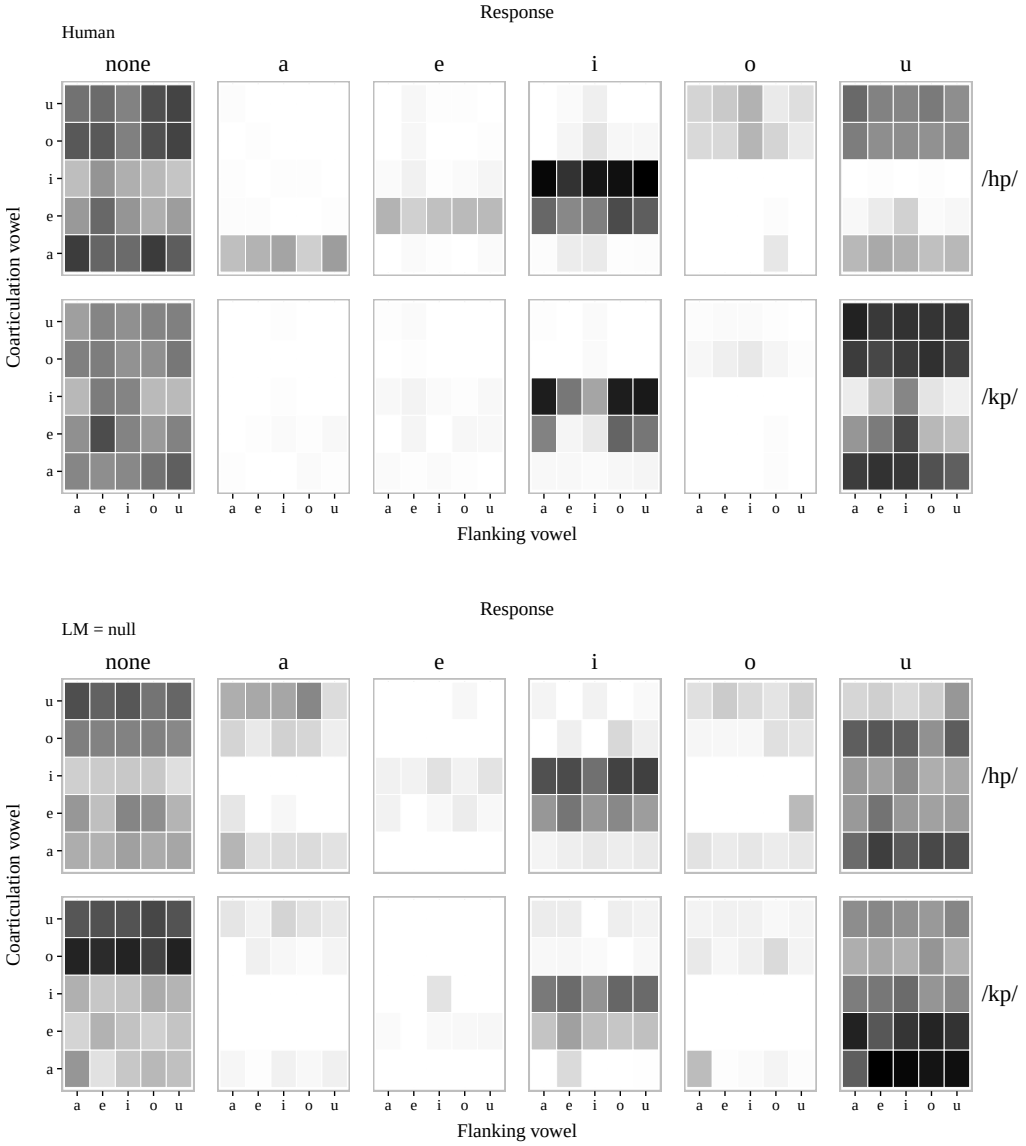

Figure 3.9: *Counts of responses for the test items for human participants (top panel) and the ASR model with a null LM (bottom panel). Within each panel: Top: /hp/-items; bottom: /kp/-items. Within each rectangle, flanking vowels and vowel coarticulation are given in the horizontal and vertical axes, respectively. Darker colours indicate higher counts.*

more noisy than for human participants. In spite of that, we can notice several similarities to human responses. We generally see that responses are mostly organised in horizontal lines, with "none" and "u" responses spread relatively uniformly across all combinations of vowel coarticulations and flanking vowels. This spread was even more uniform than for human responses, where less "u" responses were given for items with front vowel coarticulation (i.e., [i, e]). In spite of this increase in "u" responses for items with front vowel coarticulation, the model responses show that, as for humans, most "i" responses were triggered by front vowel coarticulation.

When focusing on /hp items, we see that for human responses there is a correspondence between the quality of the vowel coarticulation and the response (e.g., most "a" responses come from items with [a] coarticulation). This correspondence is blurred in model responses:

- "a" responses: A horizontal line is visible corresponding to [a] vowel coarticulation as it is for human responses. However, additional horizontal lines corresponding to back vowel coarticulation (i.e., [o, u]) are also visible. The source of most "a" responses are items with [u] coarticulation.

- "e" responses: These were triggered by front vowel coarticulation for the model, while for humans they were triggered specifically by [e] vowel coarticulation (horizontal line) and /e/ flanking vowel (fainter vertical line).

- "i" responses: Similar to humans, the majority of "i" responses given by the model were triggered by front vowel coarticulation. However, instead of seeing fainter vertical lines corresponding to front vowel flanking vowels as in human responses, for the model, "i" responses were also triggered to a minor extent by non-front vowel coarticulation, even when the flanking vowel was not a front vowel.

- "o" responses: For humans, this response was triggered by back vowel coarticulation. For the model, [a] vowel coarticulation also triggered "o" responses. In other words, "o" responses were mostly triggered by non-front vowel coarticulation.

Additionally, we see that differences between model responses for /hp/-items and /kp/-items is not as apparent as it is for human responses. In the latter (top panels of Figure 3.9), we see that participants barely responded {"a", "e", "o"} for /kp/-items. Meanwhile, for model responses, the rectangles for {"a", "e", "o"} responses for /kp/-items are fainter versions of their /hp/ counterparts.

Results from section 2.2 led us to conclude that vowel coarticulation, which was less present in /kp/ clusters, influenced response patterns less for /kp/-items than for /hp/-items. Coherent with qualitative analyses on vowel copy epenthesis in control items, the difference of the effect of vowel coarticulation on /hp/- and /kp/-items is not as marked for the model as it is for human participants.

3.3.2.4 Summary

In summary, through quantitative analyses, we found that the responses from the ASR model with a null LM better approximated human responses, out of the different LM tested. Qualitative analyses showed that the null LM model responses where generally similar to human responses, but with some added noise.

For the identification of full vowels, like humans, the model was accurate at identifying /a, i, o/. Like humans, the model identified /u/ as "o" in many instances; however, the specific patterns were not exactly mirroring the confusions observed in human responses. Also unlike humans, the model identified instances of /e/ as "i".

For the identification of control items (i.e., spliced items with matching vowel coarticulation and flanking vowel), the model numerically mirrored the two effects observed in human responses: more default /u/ epenthesis for /kp/-items than /hp/-items, and more copy vowel epenthesis for /hp/-items than /kp/-items. However, the effects are damped for the model; the latter difference was not significant for the model, while the former was significant but of lower magnitude than in human responses. Also unlike humans, the model epenthesized vowels more often for /hp/-items than for /kp/-items, as the opposite was true for humans.

Concerning the test items (i.e., spliced items with mismatching vowel coarticulation and flanking vowel), like humans, most of the null LM model responses are in the set {"none", "u", "i"}. When examining model responses more in detail, we found that vowel coarticulation was driving responses as for humans, but the influence was less specific. One thing to note is that the noise observed in the model responses does not appear to

be random; it is in line with the acoustics of the stimuli. As seen in the annex acoustic analyses in section 2.2, vowel coarticulations in /hp/-items can be clustered as follows, based on their formant values: [[[a,u],o][e,i]]. There is a separation of front and non-front vowel coarticulations, which is also seen in the model responses. Since humans also seem to be sensitive to this acoustic proximity (e.g., "i" responses mostly triggered by [i,e] coarticulation; "o" responses mostly triggered by [o,u] coarticulation), a question that arises is if the noise observed in the model might be reduced when using a more performant acoustic model in the ASR system.

3.3.3 Experiment 2

As in the previous experiment, here we investigated how various versions of our ASR model differing in their language models (LMs) compared to real behavioural data. The models were used to simulate the identification task described in sections 2.3 and 2.4, where Japanese listeners were asked to indicate whether they heard an epenthetic vowel within the consonant cluster of $V_1C_1C_2V_2$ items (e.g., /abgi/). For human participants, we saw that (1) they mostly experienced default /u/-epenthesis, and (2) the quality of the flanking vowels V_1 and V_2 modulated their responses due to coarticulation. Does the output of the ASR model that best approximated human responses reflected the two aforementioned effects?

3.3.3.1 Methods

Stimuli We used the same stimuli as in sections 2.3 and 2.4. As a reminder, a native French speaker recorded 54 items with the structure $V_1C_1C_2V_2$, with V_1 and V_2 vowels from the set {/a/, /i/, /u/}, and C_1C_2 a cluster from the set {/bg/, /bn/, /db/, /dg/, /gb/, /gn/} (e.g. /abgi/).

Language models In order for the decoding task to be analogous to the behavioural experiment described in section 2.3.2, trial-specific language models were constructed, as shown in Figure 3.10. Thus, when decoding a $V_1C_1C_2V_2$ stimulus, the perception model was only given the possibility to transcribe it as $V_1C_1(V_{ep})(SIL)C_2V_2$, where phones between parentheses are optional and V_{ep} was from the set of vowels /a, e, i, o, u/, and SIL is an optional silence. Concerning the weights between states 2 and 3, we created language models in a way analogous to the LMs in Experiment 1, adapted to the $V_1C_1C_2V_2$ items used in this experiment.

Identification task simulation We used the same procedure as in Experiment 1. An example of how the ASR system decodes the experimental stimuli can be seen in Figure 3.11.

3.3.3.2 Results: Quantitative analysis

As in Experiment 1, we computed the Pearson's product-moment correlation coefficient between the human and model posteriorgrams in order to assess a global measure of the ressemblance between models' outputs and human data from section 2.3.2. The model with the highest correlation to the human data was the BIGRAM RETRO LM ($r = 0.43$), followed by the NULL ($r = 0.40$), the UNIGRAM ($r = 0.30$), the BIGRAM ONLINE ($r = 0.23$) and lastly, the BIGRAM BATCH LM ($r = 0.19$). Numerically, the BIGRAM RETRO LM better approximated the human data.

In order to assess if the correlation differences between the NULL LM and other LMs were significant, we computed these differences and their corresponding 95% confidence

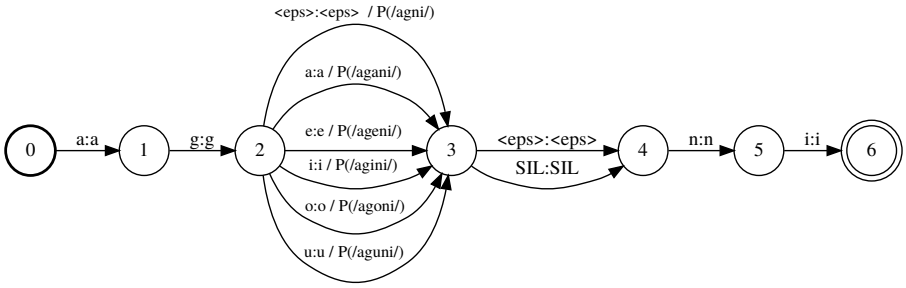

Figure 3.10: *Constrained language model used to test the models (here: LM for decoding /agni/). Nodes in the graph represent states, weighted edges represent transitions between states (here: phonemes). When relevant, weighted edges are labeled with the probability to choose that edge when decoding, which affects the final language model score of each possible path. When no weights are shown (e.g., between states 3 and 4), there is no preference between the paths concerned. The language model scores are combined with acoustic scores when decoding experimental items.*

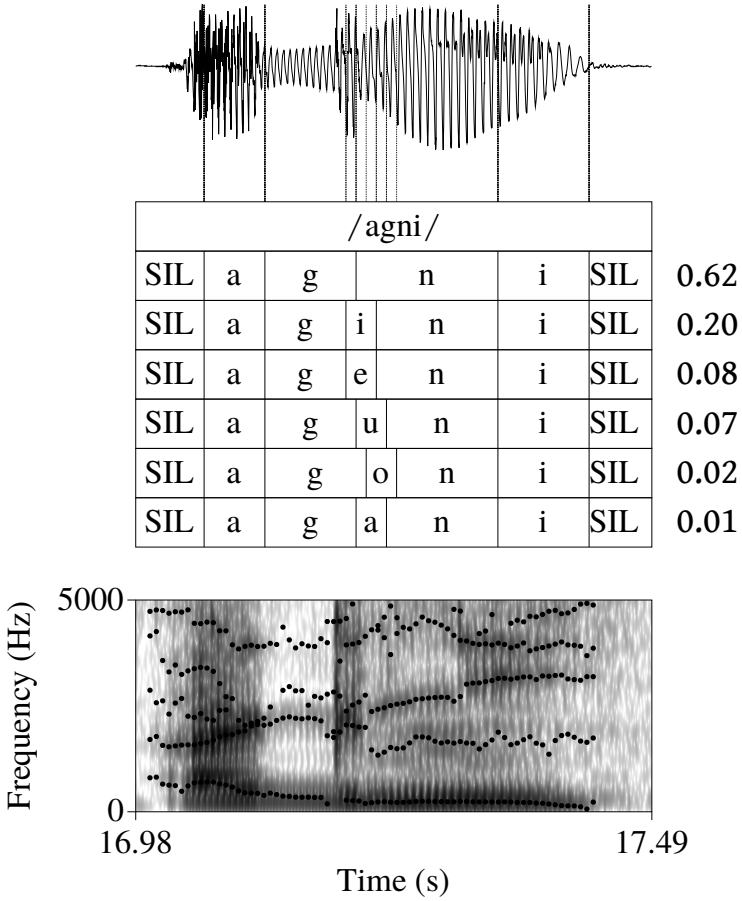

Figure 3.11: *Example of how the ASR system decodes the item /agni/, using the null version of the language model in Figure 3.10. From top to bottom: original waveform, item name, aligned transcriptions given by the model (from the most probable to the least probable, with the corresponding posteriorgrams shown to their right side), and spectrogram with formant contours. SIL = silence. TODO: get updated figure (this is with mono-6K)*

intervals (CIs), using bootstrapping with 1000 samples. As can be seen in Table 3.6, the correlation between the human data and the output of the NULL LM was significantly higher than those of the UNIGRAM, BIGRAM ONLINE and BIGRAM BATCH. While the correlation to the human data for the BIGRAM RETRO LM was numerically higher than for the NULL LM, we did not find evidence of a significant difference between the two as the CIs of their difference overlaps zero.

Table 3.6: *Difference in correlation with human data between the null LM and other LMs. The lower and upper bounds of the 95% confidence intervals are given between brackets. Positive values indicate higher correlation between human data and null model output than between human data and other LM output.*

	Correlations	Difference	Significant?
null vs. unigram	$0.40 - 0.30$	$0.11\ [0.03, 0.18]$	Yes
null vs. bigram online	$0.40 - 0.23$	$0.18\ [0.03, 0.31]$	Yes
null vs. bigram retro	$0.40 - 0.43$	$-0.03\ [-0.13, 0.08]$	No
null vs. bigram batch	$0.40 - 0.19$	$0.21\ [0.06, 0.35]$	Yes

Similarly to how we did in Experiment 1, we evaluated the similarity between models' outputs and human data without focusing on percentage of vowel epenthesis. For this we computed the correlation between the human data and models' posteriorgrams after excluding the posteriorgrams for "none" responses and re-normalising the remaining posteriorgrams. Recall that, as a consequence, we are focusing on the correlation related to epenthetic vowel quality. The highest correlation corresponded to the NULL LM ($r = 0.53$), followed by BIGRAM RETRO ($r = 0.46$), UNIGRAM ($r = 0.33$), BIGRAM ONLINE ($r = 0.21$), and BIGRAM BATCH ($r = 0.17$). As can be seen in Table 3.7, the correlation between the human data and the output of the NULL LM was significantly higher than those between the human data and other LMs.

Table 3.7: *Difference in correlation with human data between the null LM and other LMs, after removing the "none" responses. The lower and upper bounds of the 95% confidence intervals are given between brackets. Positive values indicate higher correlation between human data and null model output than between human data and other LM output.*

	Correlations	Difference	Significant?
null vs. unigram	$0.53 - 0.33$	$0.21\ [0.16, 0.26]$	Yes
null vs. bigram online	$0.53 - 0.21$	$0.33\ [0.18, 0.46]$	Yes
null vs. bigram retro	$0.53 - 0.46$	$0.08\ [0.04, 0.12]$	Yes
null vs. bigram batch	$0.53 - 0.17$	$0.36\ [0.21, 0.51]$	Yes

3.3.3.3 Results: Qualitative analysis

Default vowel Figure 3.12 shows response patterns from the behavioural experiment and model simulations. The most frequent responses given by Japanese listeners were "u" (63%), "i" (15%), and "none" (13%), with "o", "e", and "a" being infrequent responses ($< 5\%$ each). The model with the null LM also shares the same three most frequent reponses, ordered as follows based on their posteriorgrams: "none" (28%), "u" (25%), "i"

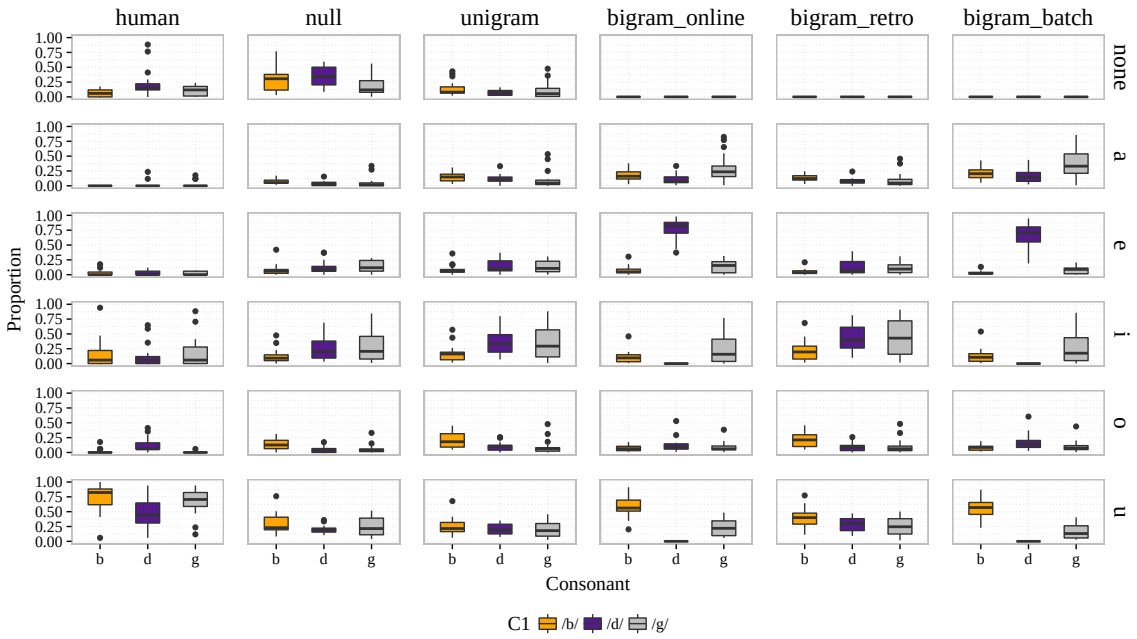

Figure 3.12: *Response patterns for the identification task. Human and models responses are separated by columns; responses are given by rows. Within each rectangle, the horizontal axis separates proportions according to C_1. The vertical axis corresponds to proportion of responses (i.e., posteriorgrams, in the case of models). The box and whiskers plots display the distribution of the proportions across items (median, quartiles and extrema).*

(23%); other responses ("e", "o", "a") had posteriorgrams below 12%. While human and model responses assign most of the responses to the same three options, we saw that the model's preferred response is not "u" but "none". Yet, humans experienced default /u/-epenthesis in more than half of the trials. As such, the model was not able to reproduce the default epenthetic vowel preference.

Effect of coarticulation Next we examined if the coarticulation effect observed in human responses also appeared in model responses. Statistical analyses were performed with the R statistical software [R Core Team, 2016], using Markov chain Monte Carlo generalised linear mixed-models [Hadfield, 2010, Plummer et al., 2006]. These Bayesian models sample coefficients from the posterior probability distribution conditioned on the data and given priors. We used priors that are standard for linear models. Model convergence was assessed by visual inspection of trace plots and the Gelman–Rubin convergence diagnostic [Gelman and Rubin, 1992], using eight chains with different initialisations. Effects were considered statistically significant if the 95% highest posterior density (HPD) interval estimated for the coefficient of interest did not include zero. We report both the posterior mode and the 95% HPD interval.

In order to assess the influence of V_1 and V_2 (henceforth: flanking vowels) on epenthetic vowel quality (/i/ or /u/), we chose as fixed effect for our statistical models Number of Same Flanking Vowels (NSFV; considered as a continuous variable with values 0, 1, or 2 instead of a factor with 3 levels, in order to reduce the number of model parameters and promote convergence). Due to the almost null variance and the consequent poor trace plot for the random intercept Cluster, we did not include it in the statistical models.

Our response variable was the continuous variable POSTERIORGRAM.[8]

The left panel of Figure 3.13 shows the posteriorgrams for /i/-epenthesis given by our ASR-based model with a "null" language model. The main effect of NSFV was significant (mode: 0.14, HPD: $[0.06, 0.22]$). An increased number of /i/ flanking vowels resulted in higher posteriorgrams for stimuli transcriptions with /i/ epenthesis.

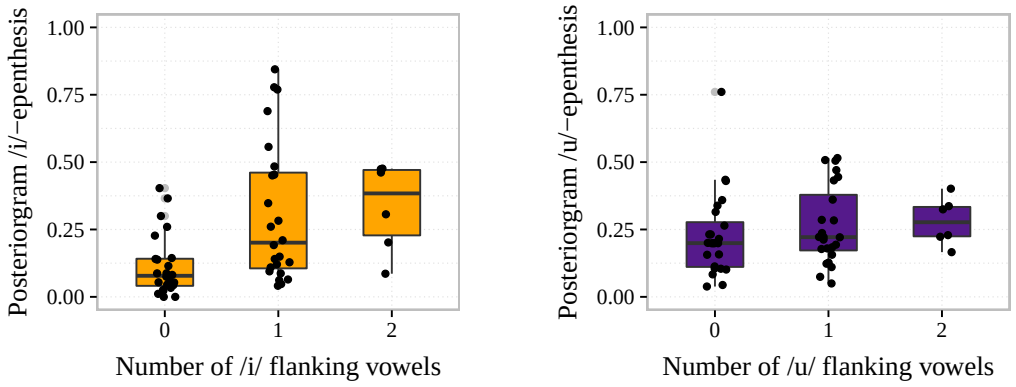

Figure 3.13: *Posteriorgrams for /i/-epenthesis (left) and /u/-epenthesis (right) obtained when decoding with a "null" language model. The box and whiskers plots display the distribution of posteriorgrams across experimental items, represented by individual dots.*

The right panel of Figure 3.13 shows the posteriorgrams for /u/-epenthesis given by our ASR-based model with a "null" language model. The main effect of NSFV was not significant (mode: 0.03, HPD: $[-0.03, 0.09]$). Therefore, an increased number of /u/ flanking vowels did not result in significantly higher posteriorgrams for stimuli transcriptions with /u/ epenthesis.

3.3.3.4 Summary

In summary, we compared the output of our various ASR models to responses given by Japanese listeners in the experiment described in sections 2.3 and 2.4. Quantitative analyses revealed that the ASR model using a null LM during decoding was better approximating human responses, in particular when examining epenthetic vowel quality. Focusing on the null model, it was able to capture Japanese listeners' preference for responding "none", "u", and "i" during the identification task. However, while humans responded "u" in more than half of the experimental trials, the model posteriorgrams for these three options were numerically very close. As such, the model was unable to capture the "default" status of /u/-epenthesis in Japanese.

Now we turn to coarticulation effects observed in the behavioural task. We saw in sections 2.3 and 2.4 that Japanese listeners were more prone to epenthesizing vowels /i/ and /u/ when more flanking vowels were of the same quality. The model was able to reflect this coarticulation effect partially: more /i/ flanking vowels resulted in significantly more /i/-epenthesis; however, we did not find evidence for the analogous situation for /u/-epenthesis.

3.3.4 Discussion

[TODO]

[8] Responses by human participants and exemplar models were given by trial; therefore in previous analyses the response variable was binomial.

3.4 Investigating acoustic/phonetic match

3.4.1 Introduction

3.4.2 Experiment 1: Phonological alternations

3.4.2.1 Methods

Stimuli The stimuli, which have been previously used in [Durvasula and Kahng, 2015], were kindly provided by the authors from said paper. They consist of 12 items of the form /eC(V)ma/, with C either an alveolar consonant from the set {/t^h, s/} or a palatal consonant from the set {/c^h, ʃ/}, and V a vowel from the set {/ɨ, i/}[9]. Each item was recorded twice by a male trained phonetician. The speaker is a native speaker of Indian English and Telugu, also a near-native speaker of standard Hindi. The clusters present in the items are phonotactically legal in these two latter languages. All items were produced with stress on the first syllable. The organisation of the stimuli, based on place of articulation of C, is shown on Table 3.8.

Table 3.8: *Experimental items. Reproduced from [Durvasula and Kahng, 2015] (Table I).*

	vowels		
	[ɨ]	[i]	none
alveolar	et^{h}ɨma	et^{h}ima	et^{h}ma
	esɨma	esima	esma
palatal	ec^{h}ɨma	ec^{h}ima	ec^{h}ma
	eʃɨma	eʃima	eʃma

ASR system Two populations of listeners were simulated in this experiment, based on [Durvasula and Kahng, 2015]: we simulated an English-listening control group using the acoustic model trained on English data (WSJ corpus), and a Korean-listening target group using the acoustic model trained on Korean data (KCSS). As a reminder, we selected the HMM-GMM monophone models with the best performance, namely the models with 15000 Gaussians.

Concerning the language models used during the decoding, in order for the decoding task to be analogous to the behavioural experiment described in [Durvasula and Kahng, 2015], trial-specific language models were constructed, as shown in Figure 3.14. Thus, when decoding the stimulus /eC_1(V_2)ma/, the perception model was only given the possibility to transcribe it as /eC_1(V_2)(SIL)ma/, where phones between parentheses are optional, V_2 was from the set of vowels /i, ɨ/, and SIL is an optional silence.

Figure 3.14: [TODO]

Figure 3.15: [TODO]

[9]Following the original article, we use [ɨ] to denote the close back unrounded vowel found in the Korean vowel inventory. However, the notation [ɯ] has also previously been used (e.g., in [Kabak and Idsardi, 2007])

Identification task simulation After decoding the stimuli with the ASR models, we extracted from the resulting lattices each possible transcription of each item, and the corresponding acoustic and language model scores. From the (scaled) acoustic and language model scores we derived the item posteriorgrams, which indicate how probable a given transcription was given the audio input. We used these probabilities as proxies of the probability that a listener might exploit when performing reverse inference during speech perception, and therefore, the probabilities used when responding in an identification task.

As such, for each item, we obtained a three-dimensional vector $ident_{model} = [p_{none}, p_i, p_ɪ]$, containing a discrete probability distribution, with a probability mass function linking the identification task options {'none', 'i', 'ɪ', to their respective probabilities (i.e., posteriorgrams). We can define the human equivalent $ident_{human} = [p_{none}, p_i, p_ɪ]$, which contains the percentage of responses for each item, after aggregating all participant responses.

3.4.2.2 Results

English model (control)

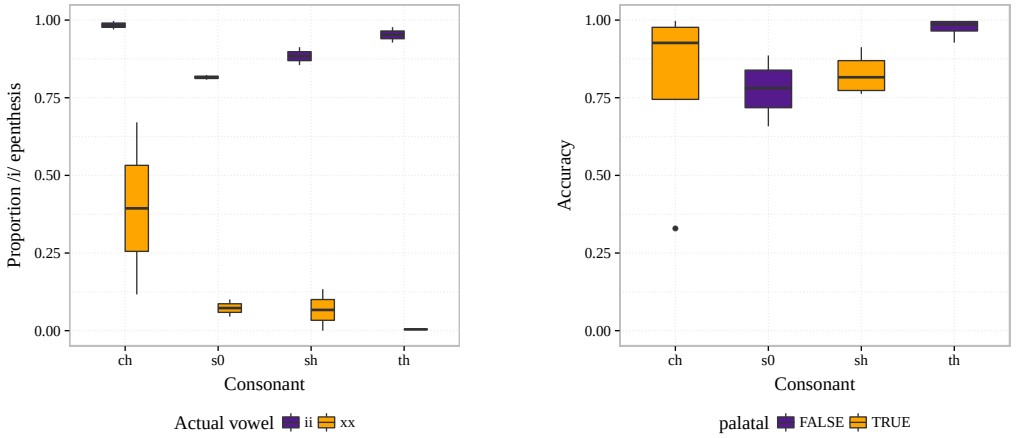

Figure 3.16: *[TODO]*

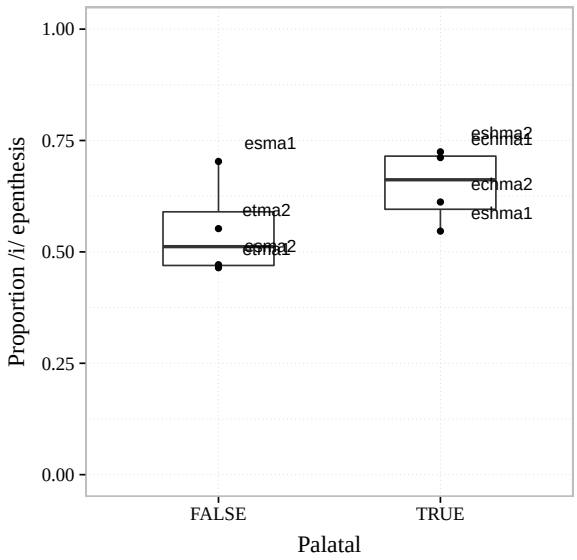

Figure 3.17: *[TODO]*

Korean model (test)

3.4.2.3 Discussion

3.4.3 Experiment 2: Allophony

3.4.3.1 Methods

Stimuli We recorded a female native speaker of Serbo-Croatian in a sound proof room reading a list of *X* items containing a C_1C_2 cluster either in syllable-initial position ($C_1C_2V_1C_3V_1$, e.g., /znapa/) or intra-syllabic position ($V_1C_1C_2V_1$, e.g., /azna/). V_1 and C_3 were always set to /a/ and /p/, respectively. We also recorded the "epenthesized" equivalents of said stimuli, namely $C_1V_{ep}C_2V_1C_3V_1$ (/zənapa/) and $V_1C_1V_{ep}C_2V_1$ (/azəna/) with V_{ep} set as [ə]. For all stimuli stress fell on the first V_1. The list of stimuli is given by Table X.

Behavioural experiment XX monolingual native listeners of American English were recruited through the online platform Amazon's Mechanical Turk. An additional XX participants were also tested, but they were excluded from the analyses if they met at least one of the following conditions: extensive exposure to languages other than English, auditory problems, dyslexia, unable to use headphones or earbuds during the experiment. This information was retrieved from pre-test and post-test questionnaires.

After audio setup[10] and *XX* training trials, in each experimental trial participants heard an item (e.g., /azna/). Since the grapheme-to-phoneme mapping is not as transparent in English as it is in Japanese or Korean, and because the position of the cluster in the item was not fixed, the task was slightly altered compared to other experiments described in previous sections. Participants were not asked if they had heard a vowel between the consonants; instead, they were asked to provide the orthographic form of what they had heard: if the auditory stimulus was /azna/, participants would be asked to click on the nonword that they heard, between options *"azna"* and *"azana"*. Since online participantsare often not as immersed in the experiment as participants tested in a laboratory setting, the experiment was self-paced and participants were able to listen to the stimuli as many times as necessary.

Each participant completed *XX* trials. For each item, participants heard either the cluster version (e.g., /azna/) or the "epenthesized" version (e.g., /azəna/). Presentation of trials was counterbalanced between participants.

ASR system We simulated perception of nonnative nonwords by English listeners using acoustic models trained on American English data (WSJ corpus). As in previous experiments, we used HMM-GMM monophone models with 15000 Gaussians. However, we tested two types of acoustic models:

1. WPD-False (non-word position-dependent) acoustic models: These are the type of models that have been used in all previous sections. For these models, all acoustic realisations of a phoneme are grouped together in a unique HMM-GMM, no matter what the position in a word of the phones were the acoustic tokens originate from. Therefore, for instance, in these models there is only one HMM-GMM corresponding to the phoneme /a/.

2. WPD-True (word position-dependent) acoustic models: In these models, different HMM-GMMs are built for phonemes, according to their position in a word (initial,

[10] Participants were given the opportunity to setup the audio to comfortable hearing levels.

medial, final, isolate). For instance, there will be four HMM-GMMs for the phoneme /a/.

WPD-False and WPD-True acoustic models are allocated the same number of Gaussians, even though the latter have more phones (up to four times more than WPD-False models). This means that it is almost certain that the average number of Gaussians per phone HMM-GMM is lower in WPD-True than in WPD-False acoustic models. Also, since now acoustic realisations are separated according to their position in a word, we expect Gaussians to be distributed differently amongst HMM-GMMs.

Concerning the language model used for decoding stimuli, item-specific language models were constructed, as shown in Figure 3.18. Thus, when decoding a $C_1(V_{ep})C_2V_1C_3V_1$ stimulus, the perception model was only given the possibility to transcribe it as $C_1(V_{ep})C_2V_1C_3V_2$, where phones between parentheses are optional and $V_{ep} = [ə]$. And similarly for $V_1C_1(V_{ep}C_2V_1$ items. While V_1 was intended to be /a/ phonologically, we allowed the model to transcribe V_1 as any phoneme associated with the grapheme [a]. This allowed to account for phonetic reduction in our stimuli, but also to foresee for the possibility that these transcriptions might be considered by English-speaking participants from the psycholinguistic experiment, due to item transcriptions being presented orthographically on-screen.

We use a null/uniform language model, which implies that listeners base their decoding of consonant clusters on phonetics alone, without using information on phonotactics.

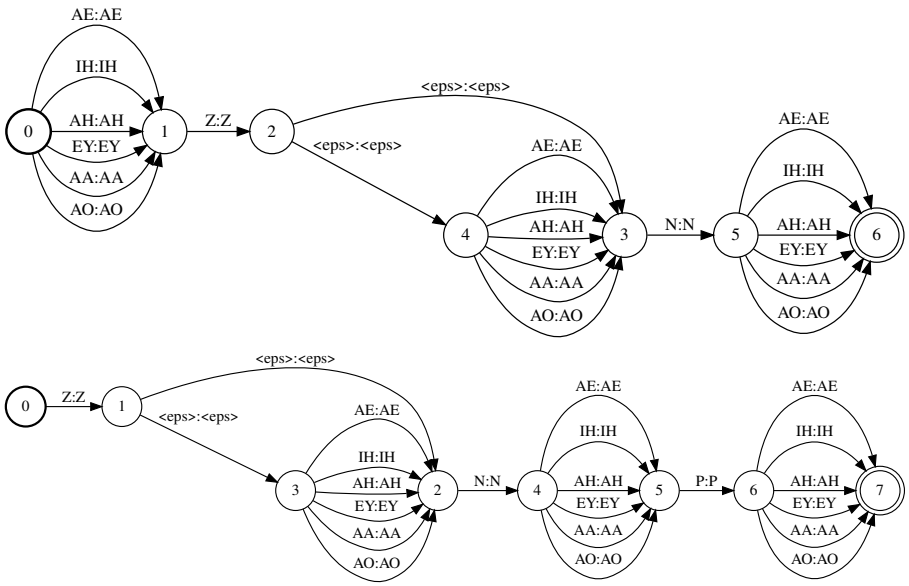

Figure 3.18: Constrained language model used to test the models (here: LMs for /azna/ (top) and /znəpa/ (bottom) trials). Nodes in the graph represent states, edges represent transitions between states (here: phonemes, transcribed in WSJ notation). Models were given the choice to transcribe the phoneme /a/ with any of the phonemes linked to the grapheme ⟨*a*⟩, as English listeners might have also done so during the task. The LMs are "null", as they only constrain the possible decoding outputs without assigning higher or lower probabilities to certain edges. The optimal decoding path is therefore only dependent on the acoustic scores.

Identification task simulation After decoding the stimuli, we obtained for each possible transcription of each item the corresponding acoustic and language model scores.

From these we derived the item posteriorgrams; we collapsed together reponses with and without epenthesis, respectively. As such, posteriorgrams indicated the probability of epenthesizing [ə] given the acoustic input. We used these probabilities as proxies of the probability that a listener might exploit when performing reverse inference during speech perception, and therefore, the probabilities used when responding in an identification task. In other words, for each item, we obtained a percentage of vowel epenthesis.

3.4.3.2 Results

Comparing models

Comparing models to humans

3.4.3.3 Discussion

3.4.4 General discussion

3.5 Conclusions

Chapter 4

General Discussion

Paradigm

- Identification task → metalinguistic. How to model A(B)X task with our models?

ASR system

- Native performance not state-of-the-art (accuracy in control items not at 100%). Influence on results?
- Influence of using forced-alignment instead of manual alignments?
- Input features?
- Corpus quality? Transcriptions are not phonetic (e.g., katakana, building from gold lexicon, ...)

Quantitative approach

- This thesis == POC: introduce methodology for testing theories, but need more power
- Automatisation of the process of making items from corpora (either for identification task or for ABX task) → test subjects in mass (e.g., Mechanical Turk)?
- cf Appendix for example of corpus ABX task

Bibliography

[Bates et al., 2015] Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

[Berent et al., 2008] Berent, I., Lennertz, T., Jun, J., Moreno, M. A., and Smolensky, P. (2008). Language universals in human brains. *Proceedings of the National Academy of Sciences*, 105(14):5321–5325.

[Berent et al., 2007] Berent, I., Steriade, D., Lennertz, T., and Vaknin, V. (2007). What we know about what we have never heard: Evidence from perceptual illusions. *Cognition*, 104(3):591–630.

[Boersma et al., 2002] Boersma, P. et al. (2002). Praat, a system for doing phonetics by computer. *Glot international*, 5(9/10):341–345.

[Bombien et al., 2016] Bombien, L., Winkelmann, R., and Scheffers, M. (2016). *wrassp: an R wrapper to the ASSP Library*. R package version 0.1.4.

[Chan and Lane, 2015] Chan, W. and Lane, I. (2015). Deep recurrent neural networks for acoustic modelling. *arXiv preprint arXiv:1504.01482*.

[Chi et al., 2005] Chi, T., Ru, P., and Shamma, S. A. (2005). Multiresolution spectrotemporal analysis of complex sounds. *The Journal of the Acoustical Society of America*, 118(2):887–906.

[Daland et al., 2015] Daland, R., Oh, M., and Kim, S. (2015). When in doubt, read the instructions: Orthographic effects in loanword adaptation. *Lingua*, 159:70–92.

[Davidson and Shaw, 2012] Davidson, L. and Shaw, J. A. (2012). Sources of illusion in consonant cluster perception. *Journal of Phonetics*, 40(2):234–248.

[de Jong and Park, 2012] de Jong, K. and Park, H. (2012). Vowel epenthesis and segment identity in korean learners of english. *Studies in Second Language Acquisition*, 34(01):127–155.

[Dehaene-Lambertz et al., 2000] Dehaene-Lambertz, G., Dupoux, E., and Gout, A. (2000). Electrophysiological correlates of phonological processing: a cross-linguistic study. *Journal of Cognitive Neuroscience*, 12(4):635–647.

[Dupoux et al., 1999] Dupoux, E., Kakehi, K., Hirose, Y., Pallier, C., and Mehler, J. (1999). Epenthetic vowels in Japanese: A perceptual illusion? *Journal of experimental psychology: human perception and performance*, 25(6):1568.

[Dupoux et al., 2001] Dupoux, E., Pallier, C., Kakehi, K., and Mehler, J. (2001). New evidence for prelexical phonological processing in word recognition. *Language and cognitive processes*, 16(5-6):491–505.

[Dupoux et al., 2011] Dupoux, E., Parlato, E., Frota, S., Hirose, Y., and Peperkamp, S. (2011). Where do illusory vowels come from? *Journal of Memory and Language*, 64(3):199–210.

[Durvasula and Kahng, 2015] Durvasula, K. and Kahng, J. (2015). Illusory vowels in perceptual epenthesis: the role of phonological alternations. *Phonology*, 32(03):385–416.

[Escudero et al., 2009] Escudero, P., Boersma, P., Rauber, A. S., and Bion, R. A. (2009). A cross-dialect acoustic description of vowels: Brazilian and European Portuguese. *The Journal of the Acoustical Society of America*, 126(3):1379–1393.

[Gales et al., 2008] Gales, M., Young, S., et al. (2008). The application of hidden markov models in speech recognition. *Foundations and Trends® in Signal Processing*, 1(3):195–304.

[Gelman and Rubin, 1992] Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, pages 457–472.

[Giorgino, 2009] Giorgino, T. (2009). Computing and visualizing dynamic time warping alignments in R: The dtw package. *Journal of Statistical Software*, 31(7):1–24.

[Guevara-Rukoz et al., 2017] Guevara-Rukoz, A., Parlato-Oliveira, E., Yu, S., Hirose, Y., Peperkamp, S., and Dupoux, E. (2017). Predicting epenthetic vowel quality from acoustics. In *INTERSPEECH*.

[Hadfield, 2010] Hadfield, J. D. (2010). MCMC methods for multi-response generalized linear mixed models: The MCMCglmm R package. *Journal of Statistical Software*, 33(2):1–22.

[Hallé et al., 2014] Hallé, P., Seguí, J., Domínguez, A., and Cuetos, F. (2014). Special is especial but stuto is not astuto: Perception of prothetic /e/ in speech and print by speakers of spanish. In Jaichenco, V. and Sevilla, Y., editors, *Psicolingüística en Español. Homenaje a Juan Seguí.*, pages 31–47. Buenos Aires: Secretariá de Publicaciones, Facultad de Filosofía y Letras, Universidad de Buenos Aires.

[Han, 1962] Han, M. S. (1962). Unvoicing of vowels in Japanese. *Onsei no kenkyuu*, 10:81–100.

[Hyafil et al., 2015] Hyafil, A., Fontolan, L., Kabdebon, C., Gutkin, B., and Giraud, A.-L. (2015). Speech encoding by coupled cortical theta and gamma oscillations. *Elife*, 4:e06213.

[Kabak and Idsardi, 2007] Kabak, B. and Idsardi, W. J. (2007). Perceptual distortions in the adaptation of English consonant clusters: Syllable structure or consonantal contact constraints? *Language and Speech*, 50(1):23–52.

[Keating, 1988] Keating, P. A. (1988). Underspecification in phonetics. *Phonology*, 5(02):275–292.

[Lenth, 2016] Lenth, R. V. (2016). Least-squares means: The R package lsmeans. *Journal of Statistical Software*, 69(1):1–33.

[Maekawa, 2003] Maekawa, K. (2003). Corpus of spontaneous japanese: Its design and evaluation. In *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*.

[Mattingley et al., 2015] Mattingley, W., Hume, E., and Hall, K. C. (2015). The influence of preceding consonant on perceptual epenthesis in Japanese. In *Proceedings of the 18th International Congress of Phonetic Sciences*, pages 888:1–5.

[Monahan et al., 2009] Monahan, P. J., Takahashi, E., Nakao, C., and Idsardi, W. J. (2009). Not all epenthetic contexts are equal: Differential effects in Japanese illusory vowel perception. *Japanese/Korean linguistics*, 17:391–405.

[Nosofsky, 1992] Nosofsky, R. M. (1992). Similarity scaling and cognitive process models. *Annual review of Psychology*, 43(1):25–53.

[Panayotov et al., 2015] Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). Librispeech: an asr corpus based on public domain audio books. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 5206–5210. IEEE.

[Paul and Baker, 1992] Paul, D. B. and Baker, J. M. (1992). The design for the wall street journal-based csr corpus. In *Proceedings of the workshop on Speech and Natural Language*, pages 357–362. Association for Computational Linguistics.

[Peperkamp and Dupoux, 2003] Peperkamp, S. and Dupoux, E. (2003). Reinterpreting loanword adaptations: the role of perception. In *Proceedings of the 15th international congress of phonetic sciences*, volume 367, page 370.

[Plummer et al., 2006] Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). CODA: Convergence diagnosis and output analysis for MCMC. *R News*, 6(1):7–11.

[Povey et al., 2011] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., and Vesely, K. (2011). The kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society. IEEE Catalog No.: CFP11SRW-USB.

[R Core Team, 2016] R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[Rose and Demuth, 2006] Rose, Y. and Demuth, K. (2006). Vowel epenthesis in loanword adaptation: Representational and phonetic considerations. *Lingua*, 116(7):1112–1139.

[Sakoe and Chiba, 1978] Sakoe, H. and Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing*, 26(1):43–49.

[Saraclar, 2001] Saraclar, M. (2001). *Pronunciation modeling for conversational speech recognition*. The Johns Hopkins University.

[Schatz, 2016] Schatz, T. (2016). *ABX-Discriminability Measures and Applications*. PhD thesis, Ecole Normale Supérieure, Paris.

[Sebastián-Gallés, 2005] Sebastián-Gallés, N. (2005). Cross-language speech perception. *The handbook of speech perception*, pages 546–566.

[Steriade, 2001] Steriade, D. (2001). The phonology of perceptibility effects: the p-map and its consequences for constraint organization. *Ms., UCLA*.

[Uffmann, 2006] Uffmann, C. (2006). Epenthetic vowel quality in loanwords: Empirical and formal issues. *Lingua*, 116(7):1079–1111.

[Vance, 1987] Vance, T. J. (1987). *Introduction to Japanese phonology.* Albany, N.Y.: State University of New York Press.

[Wilson and Davidson, 2013] Wilson, C. and Davidson, L. (2013). Bayesian analysis of non-native cluster production. In *Proceedings of NELS*, volume 40.

[Wilson et al., 2014] Wilson, C., Davidson, L., and Martin, S. (2014). Effects of acoustic–phonetic detail on cross-language speech production. *Journal of Memory and Language*, 77:1–24.

[Yun et al., 2015] Yun, W., Yoon, K., Park, S., Lee, J., Cho, S., Kang, D., Byun, K., Hahn, H., and Kim, J. (2015). The korean corpus of spontaneous speech. *Daegu: Industry-Academic Cooperation Foundation, Keimyung University (Distributor)*.

COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science (2018) 1–32
2335
ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/cogs.12616

Are Words Easier to Learn From Infant- Than Adult-Directed Speech? A Quantitative Corpus-Based Investigation

Adriana Guevara-Rukoz,[a] Alejandrina Cristia,[a] Bogdan Ludusan,[a,b] Roland Thiollière,[a] Andrew Martin,[c] Reiko Mazuka,[b,d] Emmanuel Dupoux[a]

[a]*Laboratoire de Sciences Cognitives et Psycholinguistique, ENS/EHESS/CNRS/PSL*
[b]*Laboratory for Language Development, RIKEN Brain Science Institute*
[c]*Faculty of Letters, Department of English Literature and Language, Konan University*
[d]*Department of Psychology and Neuroscience, Duke University*

Received 15 July 2016; received in revised form 4 December 2017; accepted 26 February 2018

Abstract

We investigate whether infant-directed speech (IDS) could facilitate word form learning when compared to adult-directed speech (ADS). To study this, we examine the distribution of word forms at two levels, acoustic and phonological, using a large database of spontaneous speech in Japanese. At the acoustic level we show that, as has been documented before for phonemes, the realizations of words are more variable and less discriminable in IDS than in ADS. At the phonological level, we find an effect in the opposite direction: The IDS lexicon contains more distinctive words (such as onomatopoeias) than the ADS counterpart. Combining the acoustic and phonological metrics together in a global discriminability score reveals that the bigger separation of lexical categories in the phonological space does not compensate for the opposite effect observed at the acoustic level. As a result, IDS word forms are still globally less discriminable than ADS word forms, even though the effect is numerically small. We discuss the implication of these findings for the view that the functional role of IDS is to improve language learnability.

Keywords: Speech perception; Psycholinguistics; Language development; Word learning; Infant-directed speech; Hyperspeech

1. Introduction

Infants' language acquisition proceeds at an amazing speed despite the inherent difficulties in discovering linguistic units such as phonemes and words from continuous

Correspondence should be sent to Adriana Guevara-Rukoz, Laboratoire de Sciences Cognitives et Psycholinguistique, 29 rue d'Ulm, 75005 Paris, France. E-mail: adriana.guevara.rukoz@ens.fr

speech. A popular view holds that part of the problem may be alleviated by the infants' caregivers, who may simplify the learning task when they speak to their infants in a particular register called infant-directed speech (IDS). In this paper, we compare IDS and adult-directed speech (ADS) in terms of dimensions that are relevant to the learnability of sound categories. We first review alternative hypotheses about a possible facilitatory role of IDS.

1.1. IDS-ADS differences in the context of learnability

The notion that particular speech registers may have articulatory and acoustic properties that enhance speech perception may have been first introduced by Lindblom in the context of his Hyper and Hypo-articulation (H&H) theory (1990). In the case of hyper-articulation, the resulting listener-oriented modifications are referred to as 'hyperspeech'. Here, the priority is to enhance differences among contrasting elements, and it runs counter the speaker-oriented tendency to produce more economical articulatory sequences.

Fernald (2000) proposed a more general definition of hyperspeech in the context of language acquisition. The idea is that parents may manipulate linguistic levels other than articulatory ones, such as information relating to word frequency or neighborhood density, resulting in facilitated perception:

> [T]he hyperspeech notion should not be confined to articulatory factors at the segmental level, but should be extended to a wider range of factors in speech that facilitate comprehension by the infant.

While the hyperspeech notion initially refers to a modification of language as to enhance perception, Kuhl et al. (1997) go one step further, positing that IDS register-specific modifications may also enhance *learning*:

> Our findings demonstrate that language input to infants has culturally universal characteristics designed to promote language learning.

We call this last hypothesis the *Hyper Learnability* Hypothesis (HLH). It goes beyond the hyperspeech hypothesis in that it refers not to perception but to the language learning processes operating in the infant. Importantly, these two notions may not necessarily be aligned. In some instances, both hyperspeech and HLH are congruent with the usually reported properties of IDS: exaggerated prosody and articulation (Fernald et al., 1989; Soderstrom, 2007), shorter sentences (Fernald et al., 1989; Newport, Gleitman, & Gleitman, 1977; Phillips, 1973), simpler syntax (Newport et al., 1977; Phillips, 1973), and slower speech rate (Englund & Behne, 2005; Fernald et al., 1989) (see Golinkoff, Can, Soderstrom, & Hirsh-Pasek, 2015; Soderstrom, 2007, for more comprehensive reviews). All of these properties are plausible candidates for facilitating both language perception and language learning at the relevant linguistic levels—namely phonetic, prosodic, lexical and

syntactic—by making these features more salient or more contrastive to the infant. Yet, in other instances, perception and learning may diverge. As Kuhl (2000) notes:

> Mothers addressing infants also increase the variety of exemplars they use, behaving in a way that makes mothers resemble many different talkers, a feature shown to assist category learning in second-language learners.

In this case, increase in variability, which is known to negatively affect speech perception in both adults and children (see Bergmann, Cristia, & Dupoux, 2016; Mullennix, Pisoni, & Martin, 1989; Ryalls & Pisoni, 1997) is nevertheless hypothesized to positively affect learning in infants. Work by Rost and McMurray (2009) suggests that this might be the case for 14-month-old infants learning novel word-object mappings. However, it appears that not any kind of variability will do; only increased variability in certain cues —specifically those irrelevant to the contrasts of interest— promoted learning of word-object mappings (Rost & McMurray, 2010). This illustrates the very important point that HLH cannot be empirically tested independently of a specific hypothesis or theory of the learning process in infants. Ideally, the hypothesis or theory should be explicit enough that it could be implemented as an algorithm, which derives numerical predictions on learning outcomes when run on speech corpora of ADS and IDS (Dupoux, 2016). Unfortunately, as of today, such algorithms are not yet available for modeling early language acquisition in infants. Yet a reasonable alternative is to resort to measurements that act as a *proxy* for learning outcomes within a given theory.

In the following, we focus on a component of language processing which has been particularly well studied: speech categories. For this component, a variety of theories have been proposed, which can be separated in two types: bottom-up theories and top-down theories. We review these two types in the following sections and discuss possible proxies for them.

1.2. Bottom-up theories: Discriminability as a proxy

Bottom-up theories propose that phonetic categories emerge from the speech signal; they are extracted by attending to certain phonetic dimensions (Jusczyk, Bertoncini, Bijeljac-Babic, Kennedy, & Mehler, 1990), or by identifying category prototypes (Kuhl, 1993). More explicitly, Maye, Werker, and Gerken (2002) proposed that infants construct categories by tracking statistical modes in phonetic space. This idea can be made even more computationally explicit by using unsupervised clustering algorithms, such as Gaussian mixture estimation (De Boer & Kuhl, 2003; Lake, Vallabha, & McClelland, 2009; McMurray, Aslin, & Toscano, 2009; Vallabha, McClelland, Pons, Werker, & Amano, 2007), or self-organizing neural maps (Guenther & Gjaja, 1996; Kohonen, 1988; Vallabha et al., 2007). Given the existence of such computational algorithms, it would seem easy to test if IDS enhances learning by running them on IDS and ADS data, and then evaluating the quality of the resulting clusters.

However, this is not so simple for two reasons. First, each of the above-mentioned algorithms makes different assumptions about the number, granularity, and shape of

phonetic categories, parameters which could potentially lead to different outcomes. Even more problematic is that this subset of algorithms does not exhaust the space of possible clustering algorithms.

Since we do not know which of these assumptions and algorithms are those that best approximate computational mechanisms used by infants, applying these algorithms to data may not get us any closer to a definitive answer. Second, these particular algorithms have only been validated on artificially simplified data (e.g., representing categories as formant measurements extracted from hand-segmented data) and not on a corpus of realistic speech. In fact, when similar algorithms are run on real speech, they fail to learn phonetic categories; instead, they learn smaller and more context-dependent units (e.g., Varadarajan, Khudanpur, & Dupoux, 2008; see also Antetomaso et al., 2016). The unsupervised discovery of phonetic units is currently an unsolved problem which gives rise to a variety of approaches (see Versteegh, Anguera, Jansen, & Dupoux, 2016, for a review).

Given the unavailability of effective phoneme discovery algorithms that could test the bottom-up version of HLH, many researchers have adopted a more indirect approach using descriptive measures of phonetic category distributions as a *proxy* for learnability. Here, we review two such proxies: category *separation* and category *discriminability*.

Category separation corresponds to the distance between the center of these categories in phonetic space. Kuhl et al. (1997) measured the center of the 'point' vowels /a/, /i/, and /u/ in formant space, in ADS and IDS, across three languages (American English, Russian, and Swedish). Results revealed that the spatial separation between the center of these vowels was increased in IDS compared to ADS. This observation has been replicated in several studies (Andruski, Kuhl, & Hayashi, 1999; Bernstein Ratner, 1984; Burnham, Kitamura, & Vollmer-Conna, 2002; Cristia & Seidl, 2014; Liu, Kuhl, & Tsao, 2003; McMurray, Kovack-Lesh, Goodwin, & McEchron, 2013; Uther, Knoll, & Burnham, 2007; although see Benders, 2013). However, it is less clear that separation generalizes to other segments beyond the three point vowels. For instance, Cristia and Seidl (2014) attested increased separation of the point vowels in speech spoken to 4- and 11-month-old learners of American English, but not for other vowel contrasts (e.g., [i-I]). The between-category distance among the latter vowel categories was not larger in IDS than in ADS (see also McMurray et al., 2013, for similar results). This is problematic for learnability because one might argue on computational grounds that the vowels that are difficult to learn are probably not the point vowels which are situated at the extreme of the vocal space, but rather the ones that are in the middle and have several competitors with which they can be confused.

There is another reason to doubt that separation is a very good proxy in the first place. As shown in Fig. 1, categories are defined not only by their center, but also by their variability. If, for instance, IDS not only increases the separation between category centers compared to ADS, but also increases within-category variability, the two effects could cancel each other out or even wind up making IDS more difficult to learn. In fact, as we mentioned above, Kuhl et al. (1997) reported that parents tend to be more variable in their vowel productions in IDS than ADS. This was confirmed in later studies (Cristia & Seidl, 2014; Kirchhoff & Schimmel, 2005; McMurray et al., 2013). If so, what is the net effect of these two opposing tendencies on category learnability?

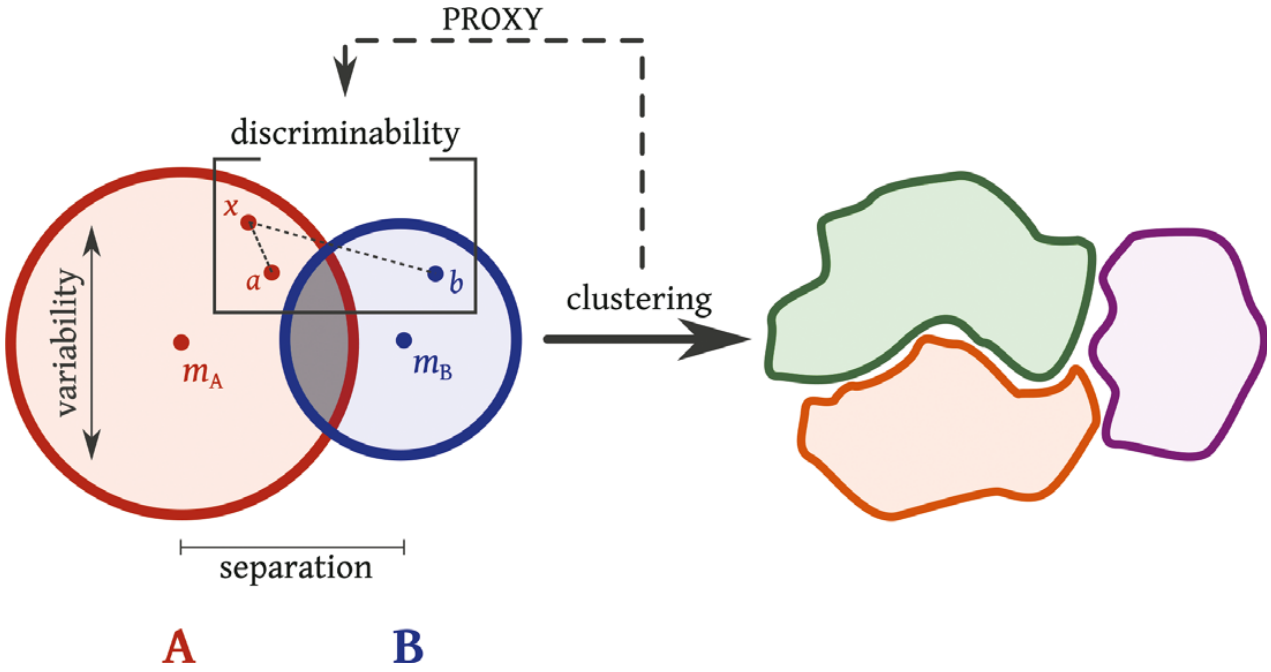

Fig. 1. Schematic view of separation, variability, and discriminability between two categories A and B (left), and a possible clustering obtained from the distributions (right). *Separation* measures the distance between the center of categories A and B; it is computed as the distance between the medoids m_A and m_B. *Variability* measures the spatial spread of tokens within a given category; it is computed as the average distance between tokens in a category. *Discriminability* depends on both variability and separation; it is quantified with an ABX score as the probability that a given token x (say, of A) is less distant to another token a of A than to a token b of B.

Previous work by Schatz (2016) has shown that the performance of unsupervised clustering algorithms can be predicted by a psychophysically inspired measure: the ABX *discrimination score*. The intuition behind this measure is illustrated in Fig. 1: it is defined as the probability that tokens within a category are closer to one another than between categories. If the two categories are completely overlapping, the ABX score is 0.5. If, on the other hand, the two categories are well segregated, the score can reach 1.[1] This work has demonstrated that the ABX score tends to be more statistically stable than standard clustering algorithms (*k*-nearest neighbors, spectral clustering, hierarchical clustering, *k*-means, etc.) while predicting their outcomes better than they predict each other's outcomes. All in all, this method is independent of specific learning algorithms, is non-parametric (i.e., it does not assume particular shapes of distributions) and can operate on any featural representation including raw acoustic features. It can therefore be used as a stable proxy of unsupervised clustering and, therefore, of bottom-up learnability.

Using this measure, Martin et al. (2015) systematically studied the discriminability of 46 phonemic contrasts of Japanese by running the ABX discriminability test on a speech corpus with features derived from an auditory model, namely mel spectral features. The outcome was that, on average, phonemic categories were actually *less discriminable* in IDS than in ADS. While most contrasts did not differ between the two registers, the few

that systematically differed pointed rather toward a decrease in acoustic contrastiveness in IDS at the phonemic level.

To sum up, if one uses ABX-discriminability as a proxy for bottom-up learnability, we can conclude that the HLH is not supported by the data available. However, bottom-up learning is not the only theoretical option available to account for phonetic learning in infants. Next, we examine top-down theories.

1.3. Top-down theories: Three learnability subproblems

Top-down theories of phonetic category learning share with linguists the intuition that phonemes are defined, not so much through their acoustic properties, but rather through their function. The function of phonemes is to carry meaning contrasts at the lexical level. Top-down theories therefore posit that phonemes emerge from the lexicon. As stated by Werker and Curtin (2005) (see also Beckman & Edwards, 2000):

> As the vocabulary expands and more words with overlapping features are added, higher order regularities emerge from the multidimensional clusters. These higher order regularities gradually coalesce into a system of contrastive phonemes. (p. 217)

There are many ways to flesh out these ideas in terms of computational mechanisms. All of them involve at least the requirement that (some) word forms are learned and that these forms constrain the acquisition of phonetic categories. This can be summarized in terms of three subproblems (Fig. 2B): (a) segmenting word tokens from continuous speech, (b) clustering said word tokens into types, and (c) using said types to learn phonetic categories via a contrastive mechanism. Arguably, these three subproblems are interdependent (in fact, some models address several of them jointly, for example, Feldman, Griffiths, & Morgan, 2009, or iteratively, for example, Versteegh, Anguera, Jansen, & Dupoux, 2016), and only a fully specified model would enable to fully test the functional impact of IDS for learnability under such a theory. Yet, as above, we claim that one can develop measures that can act as proxies for learnability, even in the absence of a full model.

In what follows, we focus on the second subproblem, that is, the clustering of word types, which we take to be of central importance for phonetic category learning. Indeed, in case of a failure to solve subproblem 1 (e.g., infants undersegment "*the dog*" into "*thedog*," or oversegment "*butterfly*" into "*butter fly*"), it is still possible to use contrastive learning with badly segmented proto-words to learn phonetic categories (Fourtassi & Dupoux, 2014). In contrast, in case of a failure to solve subproblem 2 (e.g., infants merge "*cat*" and "*dog*" into a signal word type, or split "*tomato*" into many context or speaker dependant variants), then it is much more dubious that contrastive learning can be of any help to establish phonetic categories. Our experiments therefore only address subproblem 2, and we come back to the other two subproblems in the General Discussion.

2341

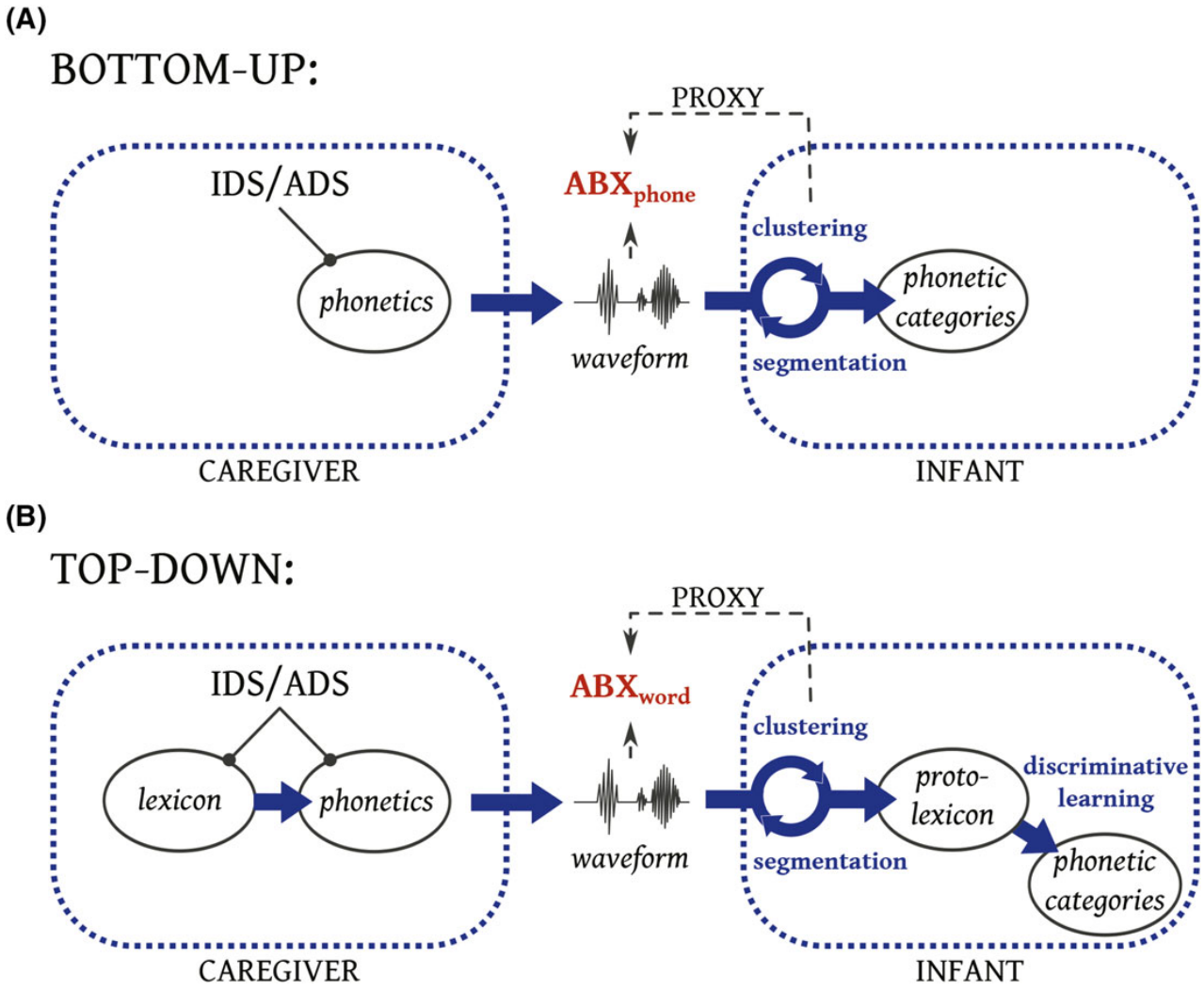

Fig. 2. Schematic view of (A) bottom-up and (B) top-down models of phonetic learning, together with ABX discriminability as a proxy for measuring the effect of adult-directed speech (ADS) versus infant-directed speech (IDS) on learnability.

1.4. The present study: Word form discriminability

The construction of word form categories is a similar computational problem to the problem of constructing phonetic categories discussed above. Both can be formulated as unsupervised clustering problems, the only difference being the granularity and number of categories being formed. Instead of sorting out instances of 'i', 'a', and 'o' into clusters, the problem is to sort out instances of 'cat', 'dog', and 'tomato' into clusters. Therefore, in both instances, it is possible to use ABX discriminability as a proxy for the (bottom-up) learnability of these categories. Of course, words being composed of phonemes, one would expect a correlation between ABX discriminability on phonemes and on words. However, the word form level introduces two specific types of effects making such a correlation far from trivially true.

First, the word level typically introduces specific patterns of phonetic variability. For instance, the word 'tomato' can be produced in a variety of ways: /tʰəˈmeɪroʊ/, /təˈmeɪtə/,

/təˈmɑtəʊ/, etc. Some of these variations are dependent on the dialect but others can surface freely within speaker, or depending on context, speaking style, or speaking rate. Such phonetic effects translate into distinct acoustic realizations of the word forms, potentially complicating the task of word form category learning. Could it be that IDS limits this source of variation, thereby helping infants to construct word form categories? Some studies have shown the use of more canonical forms in IDS than ADS (e.g., Dilley, Millett, McAuley, & Bergeson, 2014), while others have not (e.g., Fais, Kajikawa, Amano, & Werker, 2010; Lahey & Ernestus, 2014), but to our knowledge no study has looked at the global effect of these variations on word discriminability, and done so systematically. This is what we will examine in Experiment 1.

Second, and setting aside phonetic realization to focus on abstract phonological characteristics, words tend to occupy sparse regions of phonological space. Put differently, there are many more unused possible word forms than actual ones. This results in minimal pairs being generally rare. For instance, a corpus analysis reveals that, in English, Dutch, French, and German, minimal pairs will concern less than 0.1% of all pairs (Dautriche, Mahowald, Gibson, Christophe, & Piantadosi, 2017); in fact, two words selected at random will differ in more than 90% of their phonemes on average. This should make word form clustering an easier task than phonetic clustering, a welcome result for top-down theories. However, it could be that IDS modulates this effect by containing a different set of words than the vocabulary directed to adults. Corpora descriptions of IDS suggest that this is the case: Caregivers use a reduced vocabulary (Henning, Striano, & Lieven, 2005; Kaye, 1980; Phillips, 1973), which often includes a set of lexical items with special characteristics, such as syllabic reduplications and mimetics (Ferguson, 1964; Fernald & Morikawa, 1993; Mazuka, Kondo, & Hayashi, 2008). May IDS boost learning by containing more phonologically distinct word forms than ADS? This is what we will examine in Experiment 2.

The overall learnability of word forms, as far as clustering is concerned, is the combined effect of phonetic/acoustic discriminability (isolated in Experiment 1) and phonological discriminability (isolated in Experiment 2). As these two factors may go in different directions, we study the global discriminability of IDS versus ADS word form lexicons in Experiment 3.

1.5. Japanese IDS

Like other variants of IDS around the globe (Ferguson, 1964), Japanese IDS is characterized by the presence of Infant-Directed Vocabulary (IDV), 'babytalk' specifically used when interacting with infants. According to a survey and corpora studies by Mazuka et al. (2008), these words are mostly phonologically unrelated to words in the ADS lexicon. In particular, IDV presents many instances of reduplications (around 65%) and onomatopoeias/mimetic words (around 40%).[2] Phonological structures found in IDV are, in fact, more similar to phonological patterns produced by Japanese infants earlier in development than to patterns found in the adult lexicon (Tsuji, Nishikawa, & Mazuka, 2014; a list of 50 earlier produced words is given by Iba, 2000). In addition to pattern repetition

within words, IDS also presents more content word repetition, as well as more frequent and longer pauses, making utterances in IDS shorter than in ADS (Martin, Igarashi, Jincho, & Mazuka, 2016).

Regarding the phonetics of Japanese IDS, it presents pitch-range expansion (Igarashi, Nishikawa, Tanaka, & Mazuka, 2013), but it is not slower than ADS when taking into account local speech rate (Martin et al., 2016). More related to our question of phonetic categories, vowel space expansion in F1 x F2 space has been attested in Japanese IDS (Andruski et al., 1999; Miyazawa, Shinya, Martin, Kikuchi, & Mazuka, 2017); however, IDS categories presented higher variability and overlap (Miyazawa et al., 2017), consistent with the decrease in acoustic discriminability observed by Martin et al. (2015). In fact, contrary to intuition, IDS appears to present more devoicing of non-high vowels than ADS (i.e., less canonical and identifiable tokens), due to breathiness (Martin, Utsugi, & Mazuka, 2014). This paralinguistic modification of speech, which is thought to convey affect, is more prevalent in IDS than ADS (Miyazawa et al., 2017).

1.6. Corpus

Most of the Japanese studies cited above, as well as the work described in this paper, have used data from the RIKEN Japanese Mother-Infant Conversation Corpus, R-JMICC (Mazuka, Igarashi, & Nishikawa, 2006), a corpus of spoken Japanese produced by 22 mothers in two listener-dependent registers: IDS and ADS (Igarashi et al., 2013).

For our study, a word was defined as a set of co-occurring phonemes with word boundaries following the gold standard for words in Japanese, roughly corresponding to dictionary entries. Lexical derivations were considered to belong to a separate type category with respect to their corresponding lemmas. For instance, */nai/* and */aru/*, inflections of the verb ある */aru/* (English: *to be*), were evaluated as separate words. Homophones were collapsed into the same word category in the analyses.

Because of the emphasis given to phonological structure when defining word categories, devoiced vowels were considered to be phonologically identical to their voiced counterparts, and similarly for abnormally elongated vowels or consonants that did not result in lexical modifications (i.e., use of gemination for emphasis). Additionally, fragmented, mispronounced, and unintelligible words were not included in our analyses (approximately 5% out of the initial corpus). The resulting corpus is henceforth referred to as the *base corpus*; information about its content can be found in Table 1.

Table 1
Description of the base corpora for adult-directed speech (ADS) and infant-directed speech (IDS)

	ADS	IDS
Duration	3 h	11 h
Types	1,382	1,765
Tokens	12,248	34,253

2. Experiment 1: Acoustic distribution of word tokens

In this experiment, we ask whether caregivers articulate words in a more or less 'distinctive' manner when addressing their infants. Our aim is to answer this question at a purely acoustic level, that is, taking into account phonetic and acoustic variability, after removing influences from other aspects that vary across registers (e.g., lexical structure). Therefore, the following analyses have been restricted to the lexicon of words that are *common* to IDS and ADS for each parent.

Our main measure is ABX discriminability applied to entire words. As in Martin et al. (2015), we use the ABX_{score} which shows classification at chance with a value of 0.5, while perfect discrimination yields a score of 1. As such, a higher ABX_{score} for IDS than ADS would mean that, on average, parents make their word categories more acoustically discriminable when addressing their infants, making these words easier to learn according to top-down theories.

The ABX discriminability measure implies computing the acoustic distance between word tokens, and computing the probability that two tokens belonging to the same word type are closer to one another than two tokens belonging to two distinct word types.

Since it is the first time that such a discriminability measure is used at the word level, we validate it in a control condition in which there are a priori reasons to expect differences in discriminability between two speech registers. Namely, we assess the discrimination of words common to ADS and read speech (RS). This register is typically articulated in a slower, clearer, and more canonical fashion than spontaneous speech. Knowing this, we expect the ABX_{score} to be higher in read speech (RS) than in spontaneous speech (ADS).

Moreover, in order to further validate the application of our method to word units, two additional submeasures are explored, following the distinctions introduced in Fig. 1: between-category *separation* and within-category *variability*.

2.1. Methods

2.1.1. Control corpus

The Read Speech (RS) subsection of the RIKEN corpus consists of recordings from a subset of 20 out of the 22 parents which had also previously been recorded in the ADS and IDS registers. Participants read 115 sentences containing phonemes in frequencies similar to those of typical adult-directed speech (Sagisaka et al., 1990). We extracted the words that were common to the read and the ADS subcorpora for each individual parent. We obtained between 19 and 32 words, each of them having between 2 and 49 occurrences. All of these word tokens were selected for subsequent analysis in the control ADS versus RS comparison.

2.1.2. Experimental corpus

All 22 participants had data in the IDS and ADS registers. For each participant, we selected the words that were common to the two registers. We obtained between 43 and

64 word types (individual numbers can be seen in the Appendix Table A1). All of the word tokens for these types were selected for subsequent analyses in the experimental condition comparing ADS versus IDS. We did not match IDS and ADS on number of tokens per type to maximize the reliability of the metrics. Since ABX is an unbiased metric of discriminability, the size of a corpus will only modulate the standard error, not the average of the metric. It therefore cannot bias the discriminability score in IDS versus ADS; simply the fact that the ADS scores are estimated from a smaller corpus means that they will be noisier than the IDS scores. Matching the IDS corpus size to that of ADS would result in increasing the noise in the IDS scores. Number of total tokens per speaker are shown in Fig. 3.

2.1.3. Acoustic distance

The three acoustic measures that were computed, namely separation, variability, and discriminability (ABX_{score}), all depend on a common core function which provides the measure of acoustic distance between two word tokens.

As in Martin et al. (2015), we represented word tokens using compressed Mel filterbanks, which corresponds to the first stage of an auditory model (Moore, 1997; Schatz, 2016).

Specifically, the audio file of each token was converted into a sequence of auditory spectral frames sampled 100 times per second, obtained by running speech through a bank of 13 band-pass filters centered on frequencies spread according to a Mel scale between 100 and 6855 Hz (Schatz et al., 2013). The energy of the output of each of the 13 filters was computed and their dynamic range was compressed by applying a cubic root. In summary, word tokens were represented as sequences of frames, which are vectors with 13-dimensions (i.e., 1 value per filter).

The distance between a pair of tokens was computed as follows. First, the two tokens of interest were realigned in the time domain by performing dynamic time warping (DTW; Sakoe & Chiba, 1978): This algorithm searches the optimal alignment path between the sequences of frames of the two tokens that are being compared. The distance between two aligned frames being compared was set to be the angle between the two 13-dimensional feature vectors representing said frames. Secondly, the average of the frame-wise distances along the optimal alignment path was set as the distance between that pair of tokens.

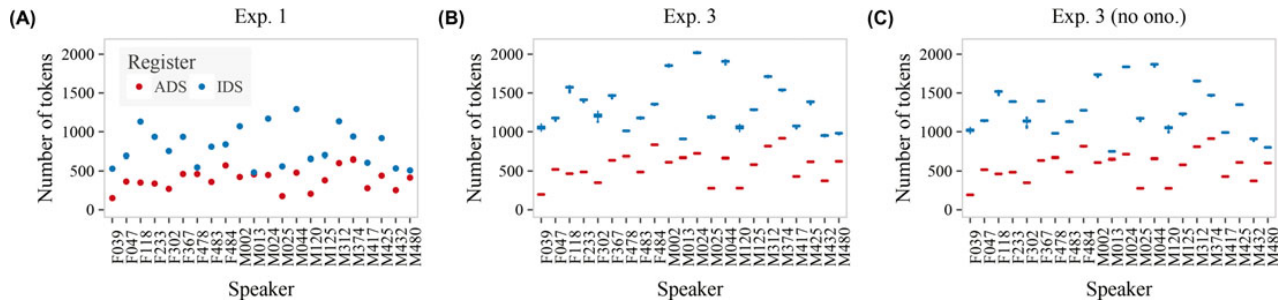

Fig. 3. Number of tokens used in Exp. 1 (A) and Exp. 3, with (B) and without (C) onomatopoeias, per speaker. For Exp. 3, boxplots show the distribution of number of tokens within the 100 sampled lexicons.

Each of the three measures was computed separately for each speaker, both for IDS and for ADS.

2.1.4. Discriminability

Discriminability calculations were performed as in Martin et al. (2015) by estimating the probability that two tokens within a category are less distant than two tokens in two different categories. This score is computed for each pair of word types, and then aggregated by averaging across all of these pairs (ABX_{score}). The calculations were done using the ABXpy package available on https://github.com/bootphon/ABXpy.

More specifically, for each pair of word types A and B, we compiled the list of all possible *(a,b,x)* triplets where *a* was a token of category *A*, *b* a token of category *B* and *x* a token of either *A* or *B*. For instance, for word types *A* = */nai/* and *B* = */aru/*, there could be a triplet with tokens $a = [nai]_1$, $b = [aru]_1$, and $x = [nai]_2$. The distance $d(a, x)$ between tokens *a* and *x* was compared to the distance $d(b, x)$ between tokens *b* and *x*. In this example, since both *a* and *x* are tokens of category *A*, we expect the acoustic distance between them to be smaller than their distance to a token belonging to a different category (i.e., token *b* of type *B*).

As such, if $d(a, x) > d(b, x)$ (i.e., $[nai]_2$ more similar to $[aru]_1$ than to $[nai]_1$), the response given by the algorithm was deemed to be incorrect and an ABX_{score} of 0 was assigned to that specific triplet. On the other hand, if as expected $d(a, x) < d(b, x)$, the algorithm returned a response deemed as correct and a score of 1 was given to the triplet. A final mean ABX_{score} for all triplets was then computed for each speaker, separately for IDS and ADS, only taking into account word pairs that were observed in both speech registers.

2.1.5. Separation

For each pair of word types, we computed the distance between their medoids. A medoid is defined as the word token which minimizes the average distance to all of the other tokens in that word type. In case of ties, we used a set of medoids, and their scores were averaged. Separation can be viewed as a generalization of the notion of phonetic expansion, except that it applies to entire word forms instead of particular segments (e.g., vowels).

2.1.6. Variability

For each word type, variability was computed as the average distance between each token and every other token within the same word type. By definition, only word types with more than one token were included in the calculation. One can view this measure as analogous to the standard deviation in univariate distributions.

2.2. Results and discussion

Regarding the control condition, we compared the acoustic discriminability of the word types common to ADS and RS. We obtained an average ABX discriminability

score per speaker per register (ADS or RS). A paired Student's *t*-test revealed that words were significantly more discriminable in RS than in ADS ($t(19) = 8.74$; $p < .0001$; Cohen's $d = 2.68$), with RS having an ABX_{score} 0.09 points higher than ADS, on average (ABX_{score} of 92% vs. 83%, respectively). As shown in Fig. 4 (panels D and H), all 20 parents showed this effect; individual scores can be found in the Appendix Table A1. In other words, on average the algorithm made twice as many errors classifying word tokens into categories in ADS compared to RS. This confirms that the ABX measure is able to capture the expected effects of read versus spontaneous speech on acoustic discriminability.

Focusing on the experimental condition, for each of the three measures (discriminability, separation, variability), we computed an aggregate score across word types separately for each parent and register (individual scores can be found in the Appendix Table A1). We then analyzed the effect of register by running a paired Student's *t*-test across parents.

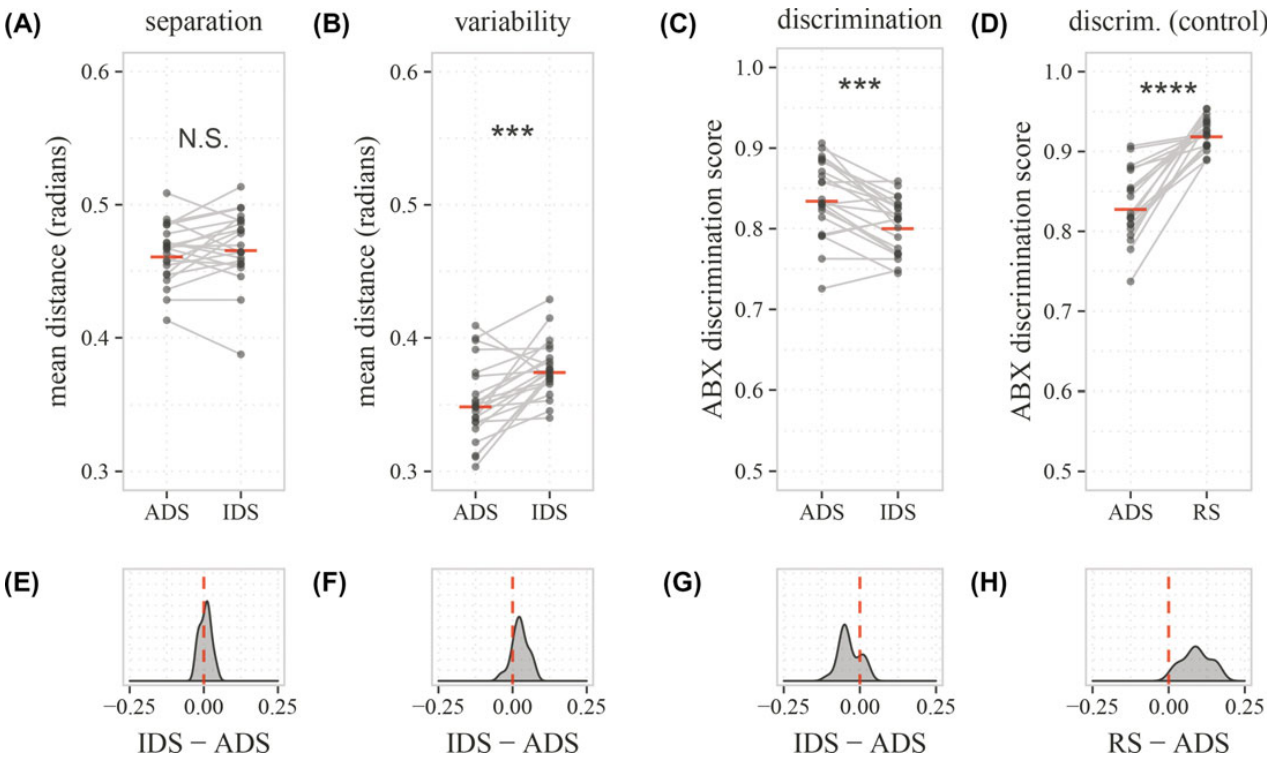

Fig. 4. Acoustic distinctiveness scores computed on word types common to infant-directed speech (IDS) and adult-directed speech (ADS) (panels A, B, C, E, F, G), or computed on word types common to ADS and RS (control condition; panels D and H). Upper panels display the distribution of the scores across speakers, as well as means within a speech register (red horizontal lines). Gray lines connect data points corresponding to the same caregiver in both registers (either ADS-IDS or ADS-RS). Bottom panels show the distribution of IDS minus ADS (or RS minus ADS) score differences. Densities to the right of the red zero line denote higher scores for IDS (or RS). A, E: Mean between-category separation (ADS vs. IDS). B, F: Mean within-category variability (ADS vs. IDS). C, G: Mean ABX discrimination score (ADS vs. IDS). D, H: Mean ABX discrimination score (ADS vs. RS; control condition). N.S., Non-significant difference. ***$p < .001$. ****$p < .0001$.

The results are visually represented in Fig. 4. First, the analysis revealed a numerically small but statistically reliable degradation in acoustic discriminability of words in IDS compared to ADS (ABX_{score} IDS: 80% vs. ADS: 84%; $t(21) = -4.73$; $p < .001$; Cohen's $d = -0.84$). This is consistent with the degradation in discriminability previously observed at the level of individual phonemes (Martin et al., 2015; McMurray et al., 2013). Second, the trend for greater separation of word categories in IDS compared to ADS was not statistically significant (IDS: 0.47 rad vs. ADS: 0.46 rad; $t(21) = 1.23$; $p > .05$; Cohen's $d = 0.21$). Finally, there was a reliable increase in variability in IDS relative to ADS (IDS: 0.38 rad vs. ADS: 0.35 rad; $t(21) = 4.28$; $p < .001$; Cohen's $d = 1.0$). This increased variability is consistent with what has been observed at the level of individual phonemes (Cristia & Seidl, 2014; McMurray et al., 2013).

In sum, we found that word discrimination is more easily achieved in ADS than in IDS. This can be analyzed as being due to a large increase in variability in IDS which is not being compensated for by a necessary increase in separation. This is in contrast to predictions posited by the *HLH*, but consistent with previous work at the phonemic level (Martin et al., 2015). In a way, this is not a totally surprising result, since by virtue of matching word types across registers, the effect of register on phoneme variability and discriminability is passed on to the level of words. What is new, however, is that the IDS register does not compensate for the phonetic variability by producing more canonical word forms. Next, we examine the content of the lexicon in the two registers.

3. Experiment 2: Phonological density

In this experiment, we focus on the phonological structure of the IDS and ADS lexicons. The core question is whether parents would select a set of words that are somewhat more 'distinctive' in IDS, yielding a sparser lexicon. Such a sparse lexicon could compensate for the increased phonetic variability measured in Experiment 1, thereby helping infants to cluster word forms into types.

We use normalized edit distance (NED) as our main measure of the sparseness of the IDS and ADS lexicons. Normalized edit distance is defined as the proportion of changes (i.e., segmental additions, deletions, and substitutions) to be performed in order to transform one word into another. The smaller the edit distance between two words, the more structurally similar they are.

NED takes into consideration not only phonological neighbors (i.e., words that differ by one phoneme), but also higher order neighbors when evaluating variation in the phonological structure of the lexicon in a psychologically relevant way. It is the direct phonological equivalent of the *separation* metric used in Experiment 1. Indeed, both metrics measure the average distance between word categories: *separation* measures acoustic distance, while *NED* measures phonological distance. Experiment 1 showed that parents do not reliably expand the acoustic space when using IDS; Experiment 2 asks: Are they expanding the phonological space when using this register?

Before moving on to the analysis, we point out that mean NED may vary with lexicon size. Indeed, as more and more words are added to a lexicon, changes in the neighborhood structure are to be expected. Typically, short words tend to have denser neighborhoods as the lexicon size increases (as the combinatorial possibilities for constructing distinct short words quickly saturate). At the same time, the ratio between short and long words tends to decrease with lexicon size, because most new additions in a lexicon tend to be long, and long words tend to have sparser neighborhoods than short words. In order to limit the influence of such properties on our results, IDS and ADS corpora were *matched in lexicon size* before any comparison was performed.

3.1. Methods

3.1.1. Sampling

As can be seen in Table 1, the volume of data available for both speaking registers in the *base corpus* was imbalanced; the IDS subset of the corpus contains more words (types and tokens) than its ADS counterpart. In order to account for this mismatch, we performed a frequency-dependent sampling of word types that matched their number in both speech registers. Types which were more frequently uttered by a speaker had a higher probability of being included in a sample than rarer ones. Moreover, since the measurement used in this section heavily relies on the nature of the words sampled, and as a way to increase estimation reliability, sampling was performed 100 times per speaker per register. For instance, if a speaker uttered 82 word types in ADS and 237 in IDS, we created 100 subsets of the IDS lexicon by sampling 82 types from the 237 available 100 times. The final metric for said speaker in a given speech register was the mean NED obtained from the corresponding 100 samples. On average, a sample contained 179.64 $\pm$ 49 word types (see Table A2 of the Appendix for more information).

3.1.2. Normalized edit distance

For each parent, within each speech register, we computed the edit distance (ED) between every possible pair of types in the sampled lexicons. ED, also called the Levenstein distance, is defined as the minimal number of additions, deletions or substitutions needed to transform one string into another. It is computed using an algorithm very similar to the Dynamic Time Warping (DTW) algorithm used in Experiment 1; the algorithm finds a path that minimizes the total number of edits (insertions, deletions and substitutions, all of them equally weighted). The maximal number of changes $max(x, y)$ is defined as the maximum length of the two types X and Y under comparison. Normalized edit distances (NEDs) were therefore derived as follows:

$$\mathrm{NED}_{XY} = \frac{\mathrm{ED}_{XY}}{\max(x, y)}$$

where x and y correspond to the phonemic lengths of two distinct words X and Y. For instance, the ED between 'tall' /tɔl/ and 'ball' /bɔl/ is 1 (one substitution: /t/ $\Rightarrow$ /b/). Both

words are 3 phonemes long, so max(x, y) = 3. Therefore, the NED between these types is $\frac{1}{3}$. The more structurally similar two types are, the closer their NED will be to zero.

3.2. Results and discussion

The distribution of the difference in mean NEDs for IDS and ADS across parents is shown on panels A and C of Fig. 5. Individual scores can be found in the Appendix Table A2. A pair-wise Student's t-test showed a systematic pattern of larger normalized edit distances in IDS than ADS (IDS: 0.877 vs. ADS: 0.871; $t(21) = 5.00$; $p < .0001$; Cohen's $d = 1.38$). This difference shows that, overall, the IDS lexicon contains words that are phonologically more distinctive than those in the ADS lexicon. In hindsight, a difference of this sort may have been expected as IDS has been found to contain "babytalk" or infant-directed vocabulary, that is, a special vocabulary which includes onomatopoeias and phonological reduplications (Ferguson, 1964; Fernald & Morikawa, 1993). This hypothesis was verified in our dataset; we found that onomatopoeias and mimetic words (hereafter referred to solely as "onomatopoeias") constituted approximately 30% of the average sample of IDS word types used in this experiment, whereas they represented less than 2% of an average ADS sample (*cf.* Appendix Table A2), this latter frequency being consistent with the use of mimetic words in Japanese observed in previous work (Saji & Imai, 2013).

In order to study the effect of onomatopoeias on phonological discriminability, we performed a post hoc analysis by resampling words after removing all onomatopoeias from the base corpus. We then re-computed the mean NED for ADS and IDS. Individual scores can be found in the right side of the Appendix Table A2. A paired Student's t-test revealed that the previously noted difference between IDS and ADS mean NED scores was no longer significant after onomatopoeia removal (IDS: 0.872 vs. ADS: 0.870; $t(21) = 1.14$; $p > 0.05$; Cohen's $d = 0.31$, visual representation on panels B and D of Fig. 5). Therefore, the IDS lexicon was found to be globally sparser than the ADS lexicon, and this effect seems to be principally driven by the unequal presence of onomatopoetic sounds in both speech registers.

Infant-directed words may facilitate lexical development not only by decreasing the overall phonological density of the lexicon, which directly impacts the clustering subproblem detailed in the introduction, but also in virtue of other intrinsic learning properties that would be relevant to a more complete model of early word learning. In the introduction, we focused on the three key word learning subproblems of segmentation, word clustering, and phonetic categorization. At this point, it is imperative to point out that there are other factors that impact word learning in infancy above and beyond these particular processes.

When asked about vocabulary specifically used when addressing infants, Japanese women report a set of words of which 40% of the items are sound-symbolic (Mazuka et al., 2008). An iconic relationship between an acoustic form and the semantics of the referent (Imai & Kita, 2014) has been shown to help 14-months-old infants finding a word's referent (Miyazaki et al., 2013), and it also facilitates the identification by pre-

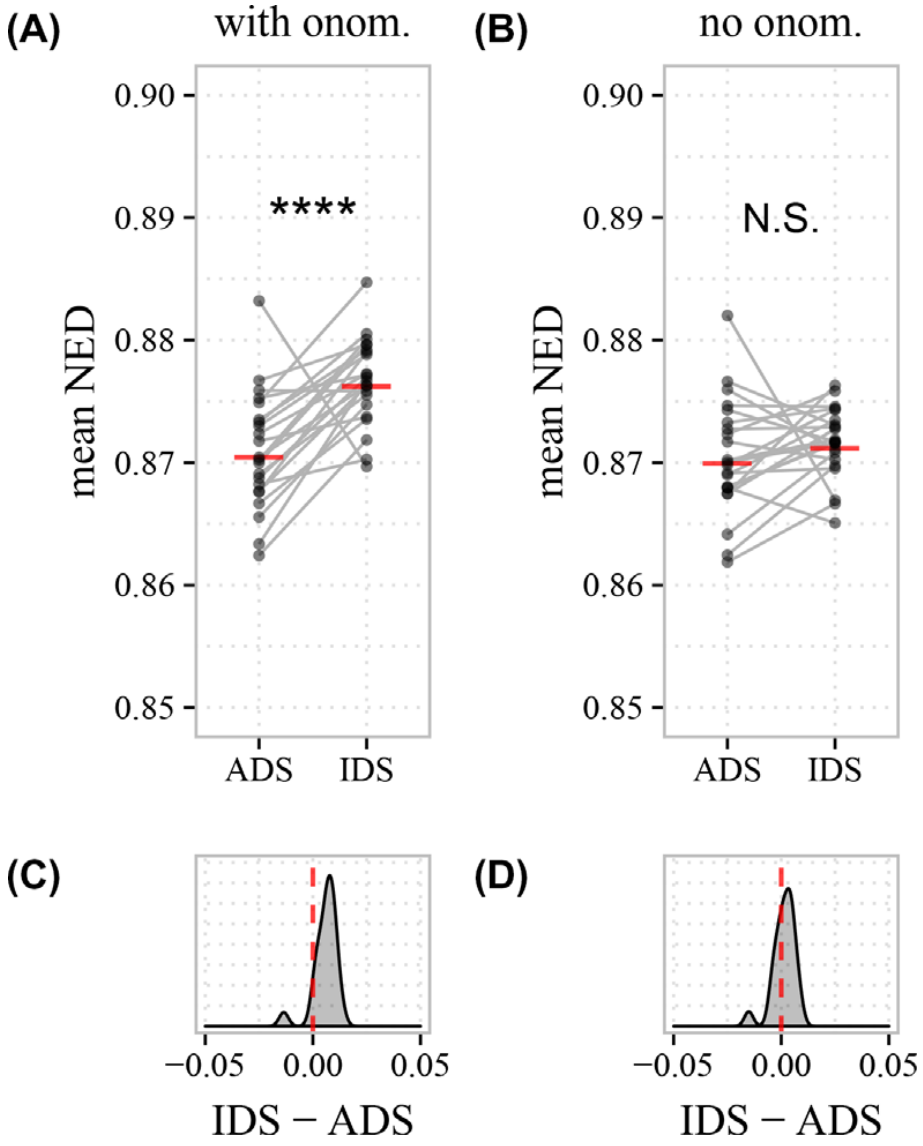

Fig. 5. Global phonological density scores (mean normalized edit distance) for adult-directed speech (ADS) and infant-directed speech (IDS), computed on lexicons matched for number of types across the two registers. Upper panels display the distribution of the scores across individual speakers, as well as means within a speech register (red horizontal lines). Gray lines connect data points corresponding to the same caregiver in both registers. Bottom panels show the distribution of IDS minus ADS score differences. Densities to the right of the red zero line denote higher scores for IDS. A, C: Samples from base corpus. B, D: Samples from base corpus after onomatopoeia removal. N.S., Non-significant difference. ****$p < .0001$.

school children of the specific features of an action a verbal word form is referring to (Imai, Kita, Nagumo, & Okada, 2008; Kantartzis, Imai, & Kita, 2011). Additionally, around 65% of the reported items contain reduplication of phonological patterns (Mazuka et al., 2008), which may impact learning at a range of levels. Repetitive patterns may be more salient and generalizable than other equally complex patterns (Endress, Dehaene-Lambertz, & Mehler, 2007; Endress, Nespor, & Mehler, 2009), and this salience could facilitate lexical acquisition in infants. This is supported by recent data showing that 9-month-old English-learning infants segment words containing reduplications (e.g., *neenee*)

from running speech more easily than words without reduplications (e.g., *neefoo*) (Ota & Skarabela, 2018). Furthermore, English-learning 18-month-old infants appear to better learn novel object labels when these contain reduplications (Ota & Skarabela, 2016). In fact, reduplication has been found to be a characteristic shared by many items from the specialized set of "babytalk" words in various languages (Ferguson, 1964), in spite of the tendency to avoid such repetitive patterns in adult language (Leben, 1973).

Similarly to what was observed in the survey by Mazuka et al. (2008), the majority of the word types tagged as onomatopoeias in our IDS corpus (i.e., around 30% of the types) present reduplication and/or sound symbolism (e.g., わんわん /waNwaN/ *dog*; ころころ /korokoro/ *light object rolling repeatedly*). Since infants seem to have a learning bias for words with these phonological characteristics, the higher proportion of onomatopoeias in IDS compared to ADS may provide an additional anchor for infant word learning.

As a reviewer pointed out, it may seem counterintuitive at first to focus on the enhanced learnability of IDS-specific words, since children are expected to eventually master all words, whether they are specific to IDS or present in both IDS and ADS. However, we are not concerned here with all of language acquisition, but only with the possibility that top-down cues affecting sound category learning are more helpful in IDS compared to ADS. Thus, even if the words that are learned are not part of a general target lexicon, they might nonetheless present an easier word clustering subproblem, and in that way lead to a lexicon that can be used as seed for subsequent sound category extraction routines.

In sum, we have found that IDS contains a higher proportion of onomatopoeias and mimetic words than ADS. Aside from their remarkable distinctiveness and salience, these items seem to contribute to decreasing the global density of the IDS lexicon. While words in IDS seem to be more spread in phonological space than words in ADS, phoneme-like representations may not yet be available to infants until a larger vocabulary is amassed (Beckman, Munson, & Edwards, 2007; Lindblom, 1992; Metsala & Walley, 1998; Pierrehumbert, 2003). As such, one may wonder if, similarly, words may be more distant in the acoustic space when taking the structural differences into account. Indeed, we notice that the effect size is almost twice as large for the phonological NED (Cohen's $d = -1.38$) than for the acoustic discriminability (Cohen's $d = -0.84$). However, given that they are not based on exactly the same tokens, it remains possible that the phonological advantage does not compensate for the acoustic disadvantage. Indeed, the difference in mean NED between IDS and ADS, while statistically significant, is numerically very small, representing a difference of less than one percent of a word. The following experiment examines the question of the effect of phonological structure on acoustic discriminability, by integrating both factors in one global discriminability measure.

4. Experiment 3: Net discriminability

In Experiment 1, we found that when we looked at the exact same word types in both registers, the IDS tokens were acoustically more confusable than the ADS tokens, due to

the increased variability in IDS word categories in the acoustic space. In other words, when removing the influence of structural peculiarities of the lexicons, IDS does not present an advantage over ADS in acoustic discriminability. We then saw in Experiment 2 that the lexicons of IDS and ADS differed structurally. Words from the IDS lexicon were phonologically more distinct than those in the ADS lexicon, in part due to onomatopoeias and mimetic words.

Here, we put these two previous results together and ask the following question: When accounting for register-specific lexical structure, is the IDS lexicon acoustically clearer than the ADS lexicon? In other words, if we take a random pair of word tokens from two different word types found in the IDS recordings, are these tokens more or less acoustically distinct than a like-built pair in the ADS recordings?

4.1. Method

4.1.1. Sampling

In order to observe the combined effects of the differences in phonological structure on acoustic discriminability, the same sampled lexicons used for Experiment 2 were used for this section, that is, 100 lexicon subsets per register per speaker, matched in number of word types across speech registers.

As it was done in Experiment 1, number of tokens per type were not matched in order to maximize the reliability of the ABX metric. Individual number of types can be seen in Table A2 of the Appendix, with total number of tokens shown in Fig. 3.

4.1.2. Computing acoustic discriminability

Acoustic discriminability was computed as described in Experiment 1. A mean ABX score was computed per sampled lexicon subset. ABX scores were collapsed by computing the mean ABX score per speaker per register.

4.2. Results and Discussion

We compared the mean ABX scores for ADS and IDS obtained on the sampled lexicons used in Experiment 2 (Fig. 6). Individual scores can be found in the Appendix Table A2. A paired Student's *t*-test revealed that mean ABX_{score} were significantly larger for ADS than for IDS, whether onomatopoeias were included in the lexicon subsets (ABX_{score} IDS: 86% vs. ADS: 87%; $t(21) = -2.37$, $p < .05$; Cohen's $d = -0.41$) or not (ABX_{score} IDS: 85% vs. ADS: 87%; $t(21) = -2.57$, $p < .05$; Cohen's $d = -0.43$). As such, similar to what was found in Experiment 1, words are less discriminable in IDS than in ADS even after taking into account the phonological specificities of the infant-directed lexicon.

This result underlines the importance of assessing effects of language acquisition enhancers not only in terms of their statistical significance across parents (*p* values, Cohen's *d*), but also quantitatively, that is, in terms of their numerical strength when combined together. To see this more clearly, we computed the increase or decrease in the score under study as a percentage relative to the ADS score taken as a baseline.

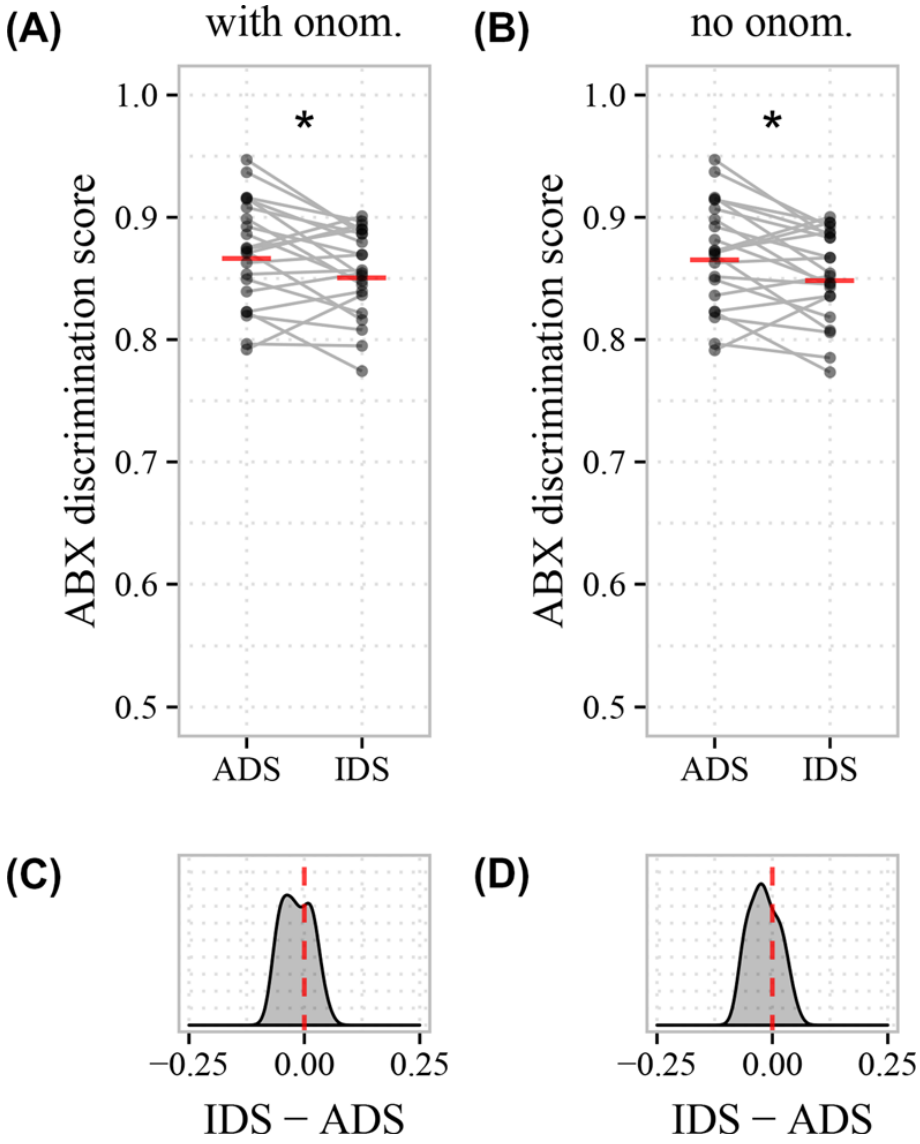

Fig. 6. Acoustic-based ABX word discrimination error in adult-directed speech (ADS) and infant-directed speech (IDS) computed on lexicons matched for number of word types across the two registers. Upper panels display the distribution of the scores across speakers, as well as means within a speech register (red horizontal lines). Gray lines connect data points corresponding to the same caregiver in both registers. Bottom panels show the distribution of IDS minus ADS score differences. Densities to the right of the red zero line denote higher error rates for IDS. A, C: Samples from base corpus. B, D: Samples from base corpus after onomatopoeia removal. *$p < .05$.

In Experiment 1, the decrement in discriminability in IDS was 4% relative to ADS, and this effect was robust across participants (Cohen's $d = -0.84$). In Experiment 2, the increase in NED represented a numerically smaller effect of less than 1% for IDS relative to ADS. This effect was actually even more robust across participants (Cohen's $d = 1.38$). Interestingly, when the two effects are combined (Experiment 3), the outcome is not determined by which effect was more statistically robust across participants, but by which one was numerically larger. Indeed, the outcome yields a numerically small (1%

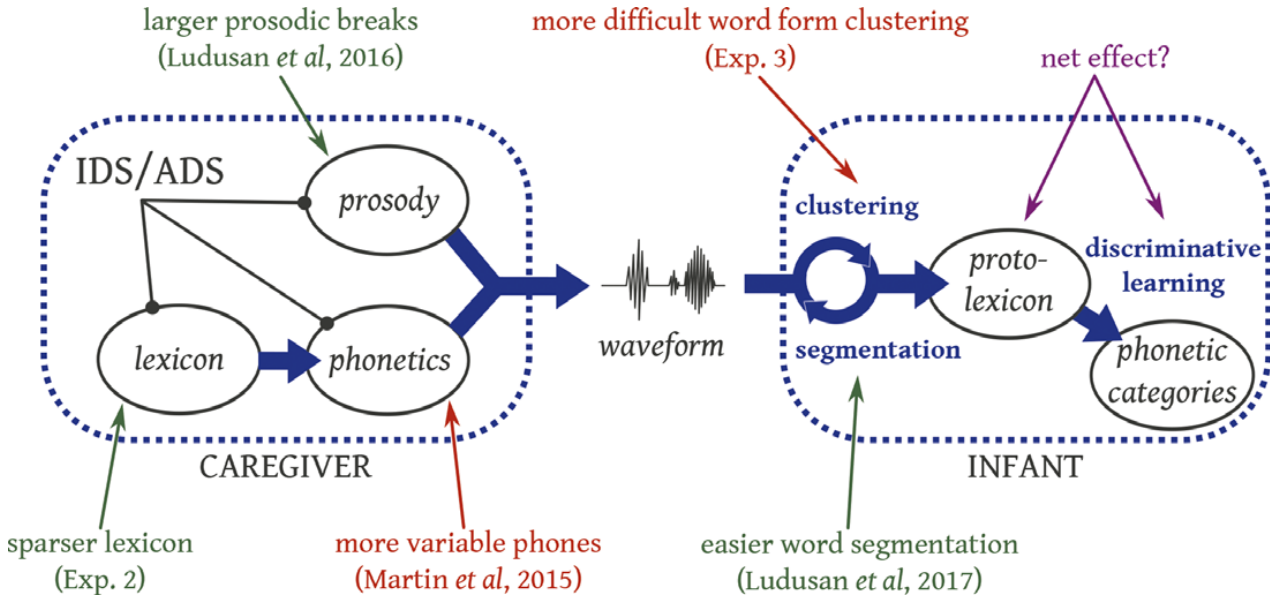

Fig. 7. Summary of infant-directed speech (IDS) characteristics relative to adult-directed speech (ADS) in a top-down model of phonetic category learning for the RIKEN corpus. Enhanced characteristics of IDS relative to ADS are shown in green, while those for which the opposite trend is observed are shown in red.

relative) decrement in discriminability, which is also much weaker across participants (Cohen's $d = -0.41$).

5. General discussion

The Hyper Learnability Hypothesis (HLH) states that when talking to their infants, parents modify the linguistic properties of their speech in order to facilitate the learning process. In this paper, we focused on the learning of phonetic categories and reviewed two classes of theories in order to quantitatively assess the HLH: (a) bottom-up theories assume that phonetic categories emerge through the unsupervised clustering of acoustic information, (b) top-down theories assume that phonetic categories emerge through contrastive feedback from learned word types. Previous work has already addressed bottom-up theories: Martin et al. (2015) examined phonemes in a corpus of Japanese laboratory recordings and found that phonemes produced by caregivers addressing their 18- to 24-month old infants were less discriminable than ADS phonemes. This rules out the HLH for that corpus and bottom-up theories. In this study, we focused on top-down theories using the same corpus and investigated the acoustic discriminability of word types.

In Experiment 1, we compared the acoustic discriminability of words that are common to both speech registers, and found that words are *less* discriminable in IDS than in ADS (an absolute decrease in ABX_{score} of 4%), likely because of increased within-category variability. This result parallels the increase in phonetic variability found in previous studies (Cristia & Seidl, 2014; Kirchhoff & Schimmel, 2005; McMurray et al., 2013), and it is consistent with the decreased phoneme discriminability measured by Martin

et al. (2015). It is not consistent, however, with the claim that words in IDS are uttered in a more canonical way than in ADS (Dilley et al., 2014; but see Fais et al., 2010; Lahey & Ernestus, 2014). In Experiment 2, we turned to the structure of the phonological lexicon. We found that the IDS lexicon was globally more spread out than that of ADS, as shown by a larger normalized edit distance between words for the former. Interestingly, this effect was attributable mostly to a higher prevalence of onomatopoeias and mimetic words in IDS. These words have idiosyncratic phonological properties, such as reduplications, which are likely responsible for the increase in global distinctiveness found in the IDS lexicon, compared to the ADS lexicon. In Experiment 3, a final analysis measured the net effect of the opposite trends found in Experiments 1 and 2, and found that, on average, words were still less acoustically discriminable in IDS than in ADS, although the effect was now considerably reduced (an absolute decrease in ABX_{score} of 1%).

Overall, then, the word form clustering subproblem is not easier to solve by using IDS input than with ADS input; quite to the contrary, there is a numerically small but consistent trend in the opposite direction. Does this undermine the HLH for top-down theories of phonetic learning as a whole? Clearly, the answer is "no," since – as explained in the Introduction – HLH actually encompasses two other learning subproblems (cf. Fig. 7). We discuss relevant evidence on IDS-ADS differences bearing on each subproblem in turn.

Regarding the problem of finding word token boundaries, Ludusan and colleagues have started studying word form segmentation using either raw acoustics or text-like phonological representations as input. Ludusan, Seidl, Dupoux, and Cristia (2015) studied the performance of acoustic word form discovery systems on a corpus of American English addressed to 4- or 11-month-olds versus adults. The overall results are similar to those of Experiment 3; that is, the two registers give similar outcomes, if anything, with a very small difference in favor of ADS, rather than the expected IDS. Computational models of word segmentation from running speech represented via acoustics are, however, well-known to underperform compared to models that represent speech via textual representations (Versteegh et al., 2016). Thus, in Ludusan, Mazuka, Bernard, Cristia, and Dupoux (2017), we studied word form segmentation from text-like representations using the same RIKEN corpus as input, and a selection of state-of-the-art cognitively based models of infant word segmentation. Results showed an advantage of IDS over ADS for most algorithms and settings.

Beyond the question of whether segmentation is easier in IDS versus ADS, we cannot move on to the next learning subproblem without pointing out that, for future work to assess the net effect of register on word segmentation, one would need to know more about the size and composition of infants' early lexicon. In fact, most accounts propose that the phonological system is extracted from the long-term lexicon, rather than on the fly from experience with the running spoken input (discussed in Bergmann, Tsuji, & Cristia, 2017). In the present paper, we have done a systematic study of word discriminability across the whole set of words present in the corpus, as if infants could segment the corpus exactly as adults do. This is, of course, unlikely. In fact, recent evidence suggests that infants may be using a suboptimal segmentation algorithm (Larsen, Dupoux, & Cristia,

2017), which leads them to accumulate a "protolexicon" containing not only words, but also over- or under-segmented tokens that do not belong to the adult-like lexicon (Ngon et al., 2013). Such protowords can nonetheless help with contrastive learning (Fourtassi & Dupoux, 2014; Martin, Peperkamp, & Dupoux, 2013).

Regarding contrastive learning of phonetic categories, it is too early to know whether the net effect of register will be beneficial or detrimental. For instance, a detrimental effect of phonetic variability in a bottom-up setting can become beneficial in a top-down setting, by presenting infants with more varied input, and therefore preparing them for future between-speaker variability. This is illustrated in the supervised learning of phonetic categories in adults (Lively, Logan, & Pisoni, 1993). However, as suggested by Rost and McMurray (2010), variability should be limited to acoustic cues that are not relevant to phonetic contrasts in order to promote learning. In order to fully assess the net effect of register, two important elements have to be clarified. First, one would need to have a fully specified model of contrastive learning itself. Candidate computational models have been proposed (e.g., Feldman et al., 2009; Fourtassi & Dupoux, 2014), but not fully validated with realistic infant-directed speech corpora (but see Versteegh et al., 2016, for an application to ADS corpora).

Throughout the above discussion, an important take-home message is that it is essential to posit well-defined, testable theories of infant learning, which can be evaluated using quantitative measures, even when fully specified computational models are not yet available. Individual studies focus only on a few pieces of the puzzle and the magnitude of each evaluated effect must be observed relative to other effects. For instance, in our study, even the relatively large effect of IDS versus ADS on the discriminability of word forms found in Experiment 1 has to be compared to the much larger effect (by a factor of 2) of read versus spontaneous speech found within the ADS register. What we propose as a methodology is to break down theories of language acquisition into component parts, and to derive proxy measures for each component to derive a more systematic grasp of the quantitative effects of register. Before closing, we would like to discuss two limitations of this study, one regarding the corpus and the other regarding the theory tested (the HLH).

The main limitation of the RIKEN corpus is that it was recorded in the laboratory and did not include naturalistic interactions between adults as they may occur in the home environment. The presence of an experimenter and props (toys, etc.) in the laboratory setting may induce some degree of non-naturalness in the interaction, both with the infant, and with the adult. Johnson, Lahey, Ernestus, and Cutler (2013) found that in Dutch, ADS is not a homogeneous register, and that it bears similarities with IDS when the addressed adult is familiar as opposed to unfamiliar.[3] It remains to be assessed whether similar results are obtained in more ecological and representative IDS and ADS samples. In addition, this study is limited by the relatively small size of the corpus. Because we analyzed each parent separately, the size of the analyzed lexicons was between 82 and 260 words, which may under-represent the range of words heard in a home setting. Finally, our analysis is limited to Japanese. There is evidence that vowel hyperarticulation varies across languages (Benders, 2013; Englund & Behne, 2005; Kuhl et al., 1997), and

more generally that the specifics of the IDS register varies across culture (e.g., Fernald & Morikawa, 1993; Igarashi et al., 2013). It would therefore be important to replicate our methods in more ecological, cross-linguistic corpora. Fortunately, the availability of wearable recording systems such as the LENA© device (Greenwood, Thiemann-Bourque, Walker, Buzhardt, & Gilkerson, 2011) increases the prospects of automatizing the collection and analysis of naturalistic speech (Soderstrom & Wittebolle, 2013).

The second limitation of this study is that we restricted our quantitative analysis to the testing of the HLH. However, the HLH is not the only hypothesis that can be addressed. Other theories have been proposed regarding the etiology and role of IDS in the linguistic development of infants (i.e., why caregivers use it, and what are the actual effects on the child). Some modifications of the input may indeed have pedagogical functions (enhancing learnability), while other modifications may decrease learnability while increasing some other factor in the parent–infant interaction. For instance, it has been documented that mothers sometimes violate the grammar of their language when teaching new words, probably in order to place the novel word in a sentence-final position (Aslin, Woodward, LaMendola, & Bever, 1996), which is salient because of properties of short-term memory. Similarly, it has sometimes been suggested that caregivers inadvertently sacrifice phonetic precision in order to make infants more comfortable and/or more receptive to the input (Papoušek & Hwang, 1991; Reilly & Bellugi, 1996). Increased phonetic variability in IDS at the phonemic level may stem from a slower speaking rate (McMurray et al., 2013), or from exaggerated prosodic variations (Fernald et al., 1989; Martin et al., 2016; Soderstrom, 2007), or possibly from gestural modifications that convey a positive affect, such as smiling (Benders, 2013), increased breathiness (Miyazawa et al., 2017) or even a vocal tract that is shortened to resemble the child's own (Kalashnikova, Carignan, & Burnham, 2017). According to a study by Trueswell et al. (2016), successful word learning interactions tend to be those in which actions performed by both caregivers and infants are precisely synchronized, with time-locking of gaze, speech and gestures. By focusing on efficiently capturing the infant's attention, caregivers could create an optimal learning environment, in spite of potential degradations brought upon lexical acoustic clarity. A similar interpretation is held by authors such as Csibra and Gergely (2006), who argue that one of the main roles of IDS is to inform the infant that speech is being directed to her, thus highlighting the pedagogical nature of the interaction as a whole. In this view, the goal of caregivers would not be to provide clearer input, but to make language interactions and their attached learning situations more exciting and attractive to infants.

Another direction entirely, is to propose that IDS may help infants to *produce* language. Ferguson (1964) describes "babytalk" as a subset of phonologically-simplified words due to reduced consonant clusters, use of coronals instead of velars, word shortening, etc. These adaptations would make it easier for developing infants to imitate the words, and/or they may be inspired by previous generations' production errors. In fact, previous work performed on our corpus shows that, if anything, the structural properties of words in our IDS sample better fit early patterns of Japanese infant speech production than those of words in ADS (Tsuji et al., 2014). While the causal relationship between babytalk use and infant word production should be further assessed experimentally, the

phonological properties of our IDS corpus suggest that, to some extent, parental input may be encouraging infant word production.

In brief, while the HLH focuses on the change in informational content of IDS which may boost (or hinder) the learnability of particular linguistic structures, IDS could have a beneficial effect on completely different grounds: enhancing overall attention or positive emotions which would increase depth of processing and retention, or facilitating production, thereby counteracting the inadvertent acoustic degradation of local units of speech such as words and phonemes. For these alternative theories of HLH to be testable within our quantitative approach, we would need to formulate these theories with enough precision that they can either be implemented, or proxies can be derived to analyze realistic corpora of caregivers/infants interactions.

To conclude, the last 50 years we have learned a great deal about how IDS and ADS differ, yet much remains to be understood. We believe it is crucial in this quest to bear in mind a detailed model of early language acquisition, and to submit predictions of this model to systematic, quantitative tests.

Acknowledgments

This work was supported by the European Research Council (Grant ERC-2011-AdG-295810 BOOTPHON), the Agence Nationale de la Recherche (Grants ANR-2010-BLAN-1901-1 BOOTLANG, ANR-14-CE30-0003 MechELex, ANR-10-IDEX-0001-02 PSL*, and ANR-10-LABX-0087 IEC), the James S. McDonnell Foundation, the Fondation de France, the Japan Society for the Promotion of Science (Kakenhi Grant 24520446, to A. Martin), and the Canon Foundation in Europe. We thank Bob McMurray and two anonymous reviewers for helpful feedback.

Author contributions

R. Mazuka oversaw the collection and coding of the corpus. A. Martin wrote the algorithms for extracting words and their phonological structure. R. Thiollière provided coding support with the ABX task. A. Cristia directed the literature review. B. Ludusan assisted with preparation of the ADS-RS comparison. A. Guevara-Rukoz and E. Dupoux carried out the acoustical and phonological analyses and, along with A. Cristia, produced the first draft. All authors contributed to the writing of this manuscript.

Notes

1. Schatz (2016) has shown that an ABX score of 1 between categories A and B implies that the two categories can be discovered without error by the clustering algorithm *k*-means.

2. In a study by Fernald and Morikawa (1993), Japanese mothers used onomatopoetic words more readily than American mothers.
3. In addition to these effects, Japanese and many other languages have a set of specialized morphemes that depend on familiarity between the talkers; this could have artificially increased the difference between IDS and ADS in the present corpus.

References

Andruski, J. E., Kuhl, P. K., & Hayashi, A. (1999). The acoustics of vowels in Japanese women's speech to infants and adults. In J. J. Ohala, Y. Hasegawa, M. Ohala, D. Granville & A. C. Bailey (Eds.), *Proceedings of the 14th International Congress on Phonetic Sciences* (Vol. *3*, pp. 2177–2179). San Francisco, CA.

Antetomaso, S., Miyazawa, K., Feldman, N., Elsner, M., Hitczenko, K., & Mazuka, R. (2016). Modeling phonetic category learning from natural acoustic data. In M. LaMendola & J. Scott (Eds.), *Proceedings of the 41th Annual Boston University Conference on Language Development* (pp. 32–45). Somerville, MA: Cascadilla Press.

Aslin, R. N., Woodward, J. Z., LaMendola, N. P., & Bever, T. G. (1996). Models of word segmentation in fluent maternal speech to infants. In J. L. Morgan & K. Demuth (Eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition* (pp. 117–134). Hillsdale, NJ: Erlbaum.

Beckman, M. E., & Edwards, J. (2000). The ontogeny of phonological categories and the primacy of lexical learning in linguistic development. *Child Development*, *71*(1), 240–249.

Beckman, M. E., Munson, B., & Edwards, J. (2007). Vocabulary growth and the developmental expansion of types of phonological knowledge. *Laboratory Phonology*, *9*, 241–264.

Benders, T. (2013). Mommy is only happy! Dutch mothers' realisation of speech sounds in infant-directed speech expresses emotion, not didactic intent. *Infant Behavior and Development*, *36*(4), 847–862.

Bergmann, C., Cristia, A., & Dupoux, E. (2016). Discriminability of sound contrasts in the face of speaker variation quantified. In A. Papafragou, D. Grodner, D. Mirman & J. C. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 1331–1336). Austin, TX: Cognitive Science Society.

Bergmann, C., Tsuji, S., & Cristia, A. (2017). Top-down versus bottom-up theories of phonological acquisition: A big data approach. In *Proceedings of Interspeech 2017* (pp. 2103–2107). Available at https://osf.io/vypwu/ https://doi.org/10.21437/interspeech.2017-1443.

Bernstein Ratner, N. (1984). Patterns of vowel modification in mother–child speech. *Journal of Child Language*, *11*(03), 557–578.

Burnham, D., Kitamura, C., & Vollmer-Conna, U. (2002). What's new, pussycat? On talking to babies and animals. *Science*, *296*(5572), 1435.

Cristia, A., & Seidl, A. (2014). The hyperarticulation hypothesis of infant-directed speech. *Journal of Child Language*, *41*, 913–934.

Csibra, G., & Gergely, G. (2006). Social learning and social cognition: The case for pedagogy. *Processes of Change in Brain and Cognitive Development. Attention and Performance XXI*, *21*, 249–274.

Dautriche, I., Mahowald, K., Gibson, E., Christophe, A., & Piantadosi, S. T. (2017). Words cluster phonetically beyond phonotactic regularities. *Cognition*, *163*, 128–145.

De Boer, B., & Kuhl, P. K. (2003). Investigating the role of infant-directed speech with a computer model. *Acoustics Research Letters Online*, *4*(4), 129–134.

Dilley, L. C., Millett, A. L., McAuley, J. D., & Bergeson, T. R. (2014). Phonetic variation in consonants in infant-directed and adult-directed speech: The case of regressive place assimilation in word-final alveolar stops. *Journal of Child Language*, *41*(01), 155–175.

Dupoux, E. (2016). Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. *Cognition*, *173*, 43–59.

Endress, A. D., Dehaene-Lambertz, G., & Mehler, J. (2007). Perceptual constraints and the learnability of simple grammars. *Cognition*, *105*(3), 577–614.

Endress, A. D., Nespor, M., & Mehler, J. (2009). Perceptual and memory constraints on language acquisition. *Trends in Cognitive Sciences*, *13*(8), 348–353.

Englund, K. T., & Behne, D. M. (2005). Infant directed speech in natural interaction — Norwegian vowel quantity and quality. *Journal of Psycholinguistic Research*, *34*(3), 259–280.

Fais, L., Kajikawa, S., Amano, S., & Werker, J. F. (2010). Now you hear it, now you don't: Vowel devoicing in Japanese infant-directed speech. *Journal of Child Language*, *37*(02), 319–340.

Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009). Learning phonetic categories by learning a lexicon. In N. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 2208–2213).

Ferguson, C. A. (1964). Baby talk in six languages. *American Anthropologist*, *66*, 103–114.

Fernald, A. (2000). Speech to infants as hyperspeech: Knowledge-driven processes in early word recognition. *Phonetica*, *57*(2–4), 242–254.

Fernald, A., & Morikawa, H. (1993). Common themes and cultural variations in Japanese and American mothers' speech to infants. *Child Development*, *64*(3), 637–656.

Fernald, A., Taeschner, T., Dunn, J., Papousek, M., de Boysson-Bardies, B., & Fukui, I. (1989). A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of Child Language*, *16*(03), 477–501.

Fourtassi, A., & Dupoux, E. (2014). A rudimentary lexicon and semantics help bootstrap phoneme acquisition. In R. Morante & W. Yih (Eds.), *Proceedings of the 18th Conference on Computational Natural Language Learning* (pp. 191–200). Association for Computational Linguistics.

Golinkoff, R. M., Can, D. D., Soderstrom, M., & Hirsh-Pasek, K. (2015). (Baby) talk to me: The social context of infant-directed speech and its effects on early language acquisition. *Current Directions in Psychological Science*, *24*(5), 339–344.

Greenwood, C. R., Thiemann-Bourque, K., Walker, D., Buzhardt, J., & Gilkerson, J. (2011). Assessing children's home language environments using automatic speech recognition technology. *Communication Disorders Quarterly*, *32*(2), 83–92.

Guenther, F. H., & Gjaja, M. N. (1996). The perceptual magnet effect as an emergent property of neural map formation. *Journal of the Acoustical Society of America*, *100*, 1111–1121.

Henning, A., Striano, T., & Lieven, E. V. (2005). Maternal speech to infants at 1 and 3 months of age. *Infant Behavior and Development*, *28*(4), 519–536.

Iba, M. (2000). An analysis of the first 50 words acquired by young Japanese children. *Language and Culture: The Journal of the Institute for Language and Culture*, *4*, 45–56.

Igarashi, Y., Nishikawa, K., Tanaka, K., & Mazuka, R. (2013). Phonological theory informs the analysis of intonational exaggeration in Japanese infant-directed speech. *The Journal of the Acoustical Society of America*, *134*(2), 1283–1294.

Imai, M., & Kita, S. (2014). The sound symbolism bootstrapping hypothesis for language acquisition and language evolution. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *369*(1651), 20130298.

Imai, M., Kita, S., Nagumo, M., & Okada, H. (2008). Sound symbolism facilitates early verb learning. *Cognition*, *109*(1), 54–65.

Johnson, E. K., Lahey, M., Ernestus, M., & Cutler, A. (2013). A multimodal corpus of speech to infant and adult listeners. *The Journal of the Acoustical Society of America*, *134*, EL534–EL540.

Jusczyk, P. W., Bertoncini, J., Bijeljac-Babic, R., Kennedy, L. J., & Mehler, J. (1990). The role of attention in speech perception by young infants. *Cognitive Development*, *5*(3), 265–286.

Kalashnikova, M., Carignan, C., & Burnham, D. (2017). The origins of babytalk: Smiling, teaching or social convergence? *Royal Society Open Science*, *4*(8), 170306.

Kantartzis, K., Imai, M., & Kita, S. (2011). Japanese sound-symbolism facilitates word learning in English-speaking children. *Cognitive Science*, *35*(3), 575–586.

Kaye, K. (1980). Why we don't talk 'baby talk' to babies. *Journal of Child Language*, *7*(03), 489–507.

Kirchhoff, K., & Schimmel, S. (2005). Statistical properties of infant-directed versus adult-directed speech: Insights from speech recognition. *The Journal of the Acoustical Society of America*, *117*(4), 2238–2246.

Kohonen, T. (1988). The 'neural' phonetic typewriter. *Computer*, *21*(3), 11–22.

Kuhl, P. K. (1993). Early linguistic experience and phonetic perception: Implications for theories of developmental speech perception. *Journal of Phonetics*, *21*, 125–139.

Kuhl, P. K. (2000). A new view of language acquisition. *Proceedings of the National Academy of Sciences*, *97*(22), 11850–11857.

Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Chistovich, L. A., Kozhevnikova, E. V., Ryskina, V. L., & Lacerda, F. (1997). Cross-language analysis of phonetic units in language addressed to infants. *Science*, *277*(5326), 684–686.

Lahey, M., & Ernestus, M. (2014). Pronunciation variation in infant-directed speech: Phonetic reduction of two highly frequent words. *Language Learning and Development*, *10*(4), 308–327.

Lake, B., Vallabha, G., & McClelland, J. (2009). Modeling unsupervised perceptual category learning. *IEEE Transactions on Autonomous Mental Development*, *1*(1), 35–43. https://doi.org/10.1109/TAMD.2009.2021703.

Larsen, E., Dupoux, E., & Cristia, A. (2017). Relating unsupervised word segmentation to reported vocabulary acquisition. In *Proceedings of Interspeech* (pp. 2198–2202).

Leben, W. R. (1973). *Suprasegmental phonology*. (Unpublished doctoral dissertation). Cambridge, MA: Massachusetts Institute of Technology.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In Hardcastle & Marchal (Eds.), *Speech production and speech modelling* (pp. 403–439). Dordrecht, the Netherlands: Springer.

Lindblom, B. (1992). Phonological units as adaptive emergents of lexical development. *Phonological Development: Models, Research, Implications*, *131*, 163.

Liu, H.-M., Kuhl, P. K., & Tsao, F.-M. (2003). An association between mothers' speech clarity and infants' speech discrimination skills. *Developmental Science*, *6*(3), F1–F10.

Lively, S. E., Logan, J. S., & Pisoni, D. B. (1993). Training Japanese listeners to identify English/r/and/l/. II: The role of phonetic environment and talker variability in learning new perceptual categories. *The Journal of the Acoustical Society of America*, *94*(3), 1242–1255.

Ludusan, B., Cristia, A., Martin, A., Mazuka, R., & Dupoux, E. (2016). Learnability of prosodic boundaries: Is infant-directed speech easier? *The Journal of the Acoustical Society of America*, *140*(2), 1239–1250.

Ludusan, B., Mazuka, R., Bernard, M., Cristia, A., & Dupoux, E. (2017). 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 (Volume 2: Short Papers)* (pp. 178–183).

Ludusan, B., Seidl, A., Dupoux, E., & Cristia, A. (2015). Motif discovery in infant-and adult-directed speech. In *Conference on Empirical Methods in Natural Language Processing* (p. 93–102). Proceedings of CogACLL2015, (pp. 93–102)

Martin, A., Igarashi, Y., Jincho, N., & Mazuka, R. (2016). Utterances in infant-directed speech are shorter, not slower. *Cognition*, *156*, 52–59.

Martin, A., Peperkamp, S., & Dupoux, E. (2013). Learning phonemes with a proto-lexicon. *Cognitive Science*, *37*(1), 103–124.

Martin, A., Schatz, T., Versteegh, M., Miyazawa, K., Mazuka, R., Dupoux, E., & Cristia, A. (2015). Mothers speak less clearly to infants than to adults: A comprehensive test of the hyperarticulation hypothesis. *Psychological Science*, *26*(3), 341–347.

Martin, A., Utsugi, A., & Mazuka, R. (2014). The multidimensional nature of hyperspeech: Evidence from Japanese vowel devoicing. *Cognition*, *132*(2), 216–228.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*(3), B101–B111.

Mazuka, R., Igarashi, Y., & Nishikawa, K. (2006). Input for Learning Japanese: RIKEN Japanese Mother-Infant Conversation Corpus. Technical report of IEICE, TL2006-16, 106 (165), 11–15.

Mazuka, R., Kondo, T., & Hayashi, A. (2008). Japanese mothers' use of specialized vocabulary in infant-directed speech: Infant-directed vocabulary in Japanese. In N. Matasaka (Ed.), *The origins of language* (pp. 39–58). New York: Springer.

McMurray, B., Aslin, R. N., & Toscano, J. C. (2009). Statistical learning of phonetic categories: Insights from a computational approach. *Developmental Science*, *12*(3), 369–378. https://doi.org/10.1111/j.1467-7687.2009.00822.x.

McMurray, B., Kovack-Lesh, K. A., Goodwin, D., & McEchron, W. (2013). Infant directed speech and the development of speech perception: Enhancing development or an unintended consequence? *Cognition*, *129* (2), 362–378.

Metsala, J. L., & Walley, A. C. (1998). Spoken vocabulary growth and the segmental restructuring of lexical representations: Precursors to phonemic awareness and early reading ability. In J. Metsala, & L. Ehri (Eds.), *Word recognition in beginning literacy* (pp. 89–120). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Miyazaki, M., Hidaka, S., Imai, M., Yeung, H. H., Kantartzis, K., Okada, H., & Kita, S. (2013). The facilitatory role of sound symbolism in infant word learning. In M. Knauff, M. Pauen, N. Sebanz & N. Matasaka (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (Vol. *1*, pp. 3080–3085). Austin, TX: Cognitive Science Society.

Miyazawa, K., Shinya, T., Martin, A., Kikuchi, H., & Mazuka, R. (2017). Vowels in infant-directed speech: More breathy and more variable, but not clearer. *Cognition*, *166*, 84–93.

Moore, B. C. J. (2012). *An introduction to the psychology of hearing* (6th ed.). Leiden: Brill.

Mullennix, J. W., Pisoni, D. B., & Martin, C. S. (1989). Some effects of talker variability on spoken word recognition. *The Journal of the Acoustical Society of America*, *85*(1), 365–378.

Newport, E., Gleitman, H., & Gleitman, L. (1977). Mother, I'd rather do it myself: Some effects and non-effects of maternal speech style. In C. Snow & C. Ferguson (Eds.), *Talking to children: Language input and acquisition*. New York: Cambridge University Press.

Ngon, C., Martin, A., Dupoux, E., Cabrol, D., Dutat, M., & Peperkamp, S. (2013). (Non)words, (non)words, (non)words: Evidence for a protolexicon during the first year of life. *Developmental Science*, *16*(1), 24–34.

Ota, M., & Skarabela, B. (2016). Reduplicated words are easier to learn. *Language Learning and Development*, *12*, 380–397.

Ota, M., & Skarabela, B. (2017). Reduplication facilitates early word segmentation. *Journal of Child Language*, *45*, 204–218.

Papoušek, M., & Hwang, S.-F. C. (1991). Tone and intonation in Mandarin babytalk to presyllabic infants: Comparison with registers of adult conversation and foreign language instruction. *Applied Psycholinguistics*, *12*(04), 481–504.

Phillips, J. R. (1973). Syntax and vocabulary of mothers' speech to young children: Age and sex comparisons. *Child Development*, *44*, 182–185.

Pierrehumbert, J. B. (2003). Phonetic diversity, statistical learning, and acquisition of phonology. *Language and Speech*, *46*(2–3), 115–154.

Reilly, J. S., & Bellugi, U. (1996). Competition on the face: Affect and language in ASL motherese. *Journal of Child Language*, *23*(1), 219–239.

Rost, G. C., & McMurray, B. (2009). Speaker variability augments phonological processing in early word learning. *Developmental Science*, *12*(2), 339–349.

Rost, G. C., & McMurray, B. (2010). Finding the signal by adding noise: The role of noncontrastive phonetic variability in early word learning. *Infancy*, *15*(6), 608–635.

Ryalls, B. O., & Pisoni, D. B. (1997). The effect of talker variability on word recognition in preschool children. *Developmental Psychology*, *33*(3), 441.

Sagisaka, Y., Takeda, K., Abel, M., Katagiri, S., Umeda, T., & Kuwabara, H. (1990). A large-scale Japanese speech database. In *Proceedings of International Conference on Spoken Language Processing* (ICSLP'90, Kobe), Vol. *2*, pp. 1089–1092.

Saji, N., & Imai, M. (2013). Onomatope kenkyu no shatei—chikadzuku oto to imi [The Role of Iconicity in Lexical Development]. In K. Shinohara & R. Uno (Eds.), Onomatope kenkyu no shatei - chikadzuku oto to imi (Sound Symbolism and Mimetics) (pp. 151–166). Tokyo, Japan: Hituji Syobo.

Sakoe, H., & Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *26*(1), 43–49.

Schatz, T. (2016). *ABX-Discriminability Measures and Applications* (Unpublished doctoral dissertation). Paris: Ecole Normale Supérieure.

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 F. Bimbot et al. (Ed.), *Proceedings of Interspeech* (pp. 1781–1785).

Soderstrom, M. (2007). Beyond babytalk: Re-evaluating the nature and content of speech input to preverbal infants. *Developmental Review*, *27*(4), 501–532.

Soderstrom, M., & Wittebolle, K. (2013). When do caregivers talk? The influences of activity and time of day on caregiver speech and child vocalizations in two childcare environments. *PLoS ONE*, *8*(11), e80646.

Trueswell, J. C., Lin, Y., Armstrong, B., Cartmill, E. A., Goldin-Meadow, S., & Gleitman, L. R. (2016). Perceiving referential intent: Dynamics of reference in natural parent–child interactions. *Cognition*, *148*, 117–135.

Tsuji, S., Nishikawa, K., & Mazuka, R. (2014). Segmental distributions and consonant-vowel association patterns in Japanese infant-and adult-directed speech. *Journal of Child Language*, *41*(06), 1276–1304.

Uther, M., Knoll, M. A., & Burnham, D. (2007). Do you speak E-NG-LI-SH? A comparison of foreigner-and infant-directed speech. *Speech Communication*, *49*(1), 2–7.

Vallabha, G. K., McClelland, J. L., Pons, F., Werker, J. F., & Amano, S. (2007). Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences*, *104*(33), 13273–13278.

Varadarajan, B., Khudanpur, S., & Dupoux, E. (2008). Unsupervised learning of acoustic sub-word units. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers* (pp. 165–168). Association for Computational Linguistics.

Versteegh, M., Anguera, X., Jansen, A., & Dupoux, E. (2016). The Zero Resource Speech Challenge 2015: Proposed approaches and results. *Procedia Computer Science*, *81*, 67–72.

Werker, J. F., & Curtin, S. (2005). PRIMIR: A developmental framework of infant speech processing. *Language Learning and Development*, *1*(2), 197–234.

2365

Appendix

Table A1
Acoustic discriminability comparisons on common words in ADS versus infant-directed speech (IDS), and in adult-directed speech (ADS) versus read speech (RS) (Exp. 1). Individual scores for separation (in radians), variability (in radians) and overall acoustic discrimination

	ADS Versus IDS							ADS Versus RS (control)		
		Separation		Variability		ABX_{score}			ABX_{score}	
Speaker	No. of Types	ADS	IDS	ADS	IDS	ADS	IDS	# Types	ADS	RS
F039	43	0.47	0.45	0.35	0.38	0.83	0.77	19	0.82	0.94
F047	67	0.47	0.49	0.34	0.37	0.87	0.83	25	0.88	0.92
F118	66	0.49	0.50	0.39	0.39	0.79	0.81	23	0.79	0.95
F233	64	0.41	0.39	0.34	0.36	0.79	0.74	25	0.78	0.93
F302	54	0.49	0.50	0.40	0.43	0.76	0.76	24	0.74	0.89
F367	70	0.46	0.45	0.32	0.35	0.88	0.84	26	0.85	0.94
F478	71	0.44	0.48	0.35	0.40	0.82	0.78	28	0.80	0.95
F483	69	0.46	0.48	0.35	0.37	0.83	0.84	23	0.80	0.91
F484	81	0.45	0.46	0.37	0.39	0.81	0.77	32	0.82	0.95
M002	73	0.46	0.46	0.31	0.37	0.91	0.80	27	0.90	0.92
M013	67	0.45	0.46	0.30	0.37	0.89	0.81	31	0.88	0.93
M024	77	0.48	0.49	0.35	0.38	0.89	0.83	31	0.85	0.93
M025	43	0.49	0.46	0.31	0.37	0.89	0.83	21	0.88	0.91
M044	86	0.47	0.46	0.35	0.35	0.86	0.83	32	0.85	0.93
M120	46	0.45	0.49	0.34	0.41	0.84	0.79	19	0.82	0.89
M125	59	0.48	0.51	0.33	0.38	0.90	0.85	31	0.91	0.92
M312	94	0.47	0.48	0.37	0.38	0.83	0.77	29	0.81	0.91
M374	90	0.47	0.46	0.35	0.37	0.87	0.81	-	-	-
M417	57	0.43	0.43	0.40	0.38	0.73	0.75	-	-	-
M425	78	0.44	0.46	0.34	0.34	0.86	0.86	27	0.82	0.9
M432	49	0.51	0.49	0.41	0.37	0.79	0.81	23	0.81	0.91
M480	68	0.47	0.47	0.36	0.38	0.83	0.82	32	0.84	0.94
M	66.91	0.46	0.47	0.35	0.38	0.84	0.80	26.4	0.83	0.92
SD	14.41	0.02	0.03	0.03	0.02	0.05	0.03	4.3	0.04	0.02

Table A2
Phonological and acoustic discriminability comparisons in adult-directed speech (ADS) versus infant-directed speech (IDS) (Exp. 2 & 3). Individual mean normalized edit distance (NED) and overall acoustic discriminability before and after removal of onomatopoeias. Values are computed as the mean of the corresponding values from 100 word samplings per speaker

	With Onomatopoeias							No Onomatopoeias				
	No. of	% Onom.		NED		ABX_{score}		No. of	NED		ABX_{score}	
Speaker	Types	ADS	IDS	ADS	IDS	ADS	IDS	Types	ADS	IDS	ADS	IDS
F039	82	6.1	35.4	0.883	0.870	0.85	0.82	77	0.882	0.867	0.85	0.82
F047	178	2.2	20.2	0.873	0.879	0.90	0.84	174	0.873	0.873	0.9	0.84
F118	139	1.4	26.6	0.876	0.876	0.84	0.85	137	0.876	0.871	0.84	0.85
F233	167	1.8	22.2	0.869	0.876	0.82	0.77	164	0.868	0.872	0.82	0.77
F302	117	0.9	50.4	0.877	0.880	0.80	0.79	116	0.877	0.873	0.8	0.79
F367	187	1.1	39.0	0.873	0.879	0.91	0.88	185	0.872	0.876	0.91	0.89
F478	221	3.2	8.1	0.870	0.876	0.82	0.81	208	0.869	0.873	0.82	0.81
F483	168	0	37.5	0.870	0.880	0.87	0.89	168	0.870	0.872	0.87	0.90
F484	250	0.8	18.0	0.872	0.874	0.85	0.86	232	0.872	0.870	0.85	0.85
M002	194	2.1	27.3	0.868	0.876	0.92	0.85	190	0.867	0.873	0.92	0.85
M013	196	2.0	19.9	0.867	0.874	0.91	0.89	176	0.868	0.872	0.91	0.88
M024	229	1.3	31.0	0.863	0.877	0.92	0.90	226	0.862	0.870	0.91	0.90
M025	120	0.8	30.0	0.873	0.881	0.94	0.89	119	0.873	0.874	0.94	0.89
M044	212	1.9	29.7	0.875	0.877	0.87	0.89	208	0.874	0.872	0.87	0.89
M120	102	1.0	63.7	0.868	0.879	0.87	0.82	101	0.867	0.876	0.87	0.81
M125	194	1.0	24.2	0.862	0.872	0.95	0.89	192	0.862	0.867	0.95	0.88
M312	248	2.8	24.6	0.868	0.870	0.86	0.87	241	0.868	0.865	0.86	0.87
M374	260	1.9	19.6	0.866	0.875	0.89	0.87	255	0.864	0.871	0.89	0.87
M417	156	0.6	34.6	0.870	0.880	0.79	0.84	155	0.870	0.872	0.79	0.84
M425	191	3.7	20.4	0.872	0.877	0.87	0.90	184	0.870	0.874	0.87	0.90
M432	139	0.7	43.9	0.869	0.877	0.82	0.84	138	0.869	0.870	0.82	0.84
M480	202	2.0	25.7	0.875	0.885	0.89	0.85	185	0.875	0.875	0.88	0.85
M	179.64	1.79	29.64	0.871	0.877	0.87	0.86	174.14	0.870	0.872	0.87	0.85
SD	48.72	1.31	12.18	0.005	0.004	0.04	0.04	46.11	0.005	0.003	0.04	0.04