



**HAL**  
open science

# Mmm whatcha say? Uncovering distal and proximal context effects in first and second-language word perception using psychophysical reverse correlation

Paige Tuttösi, H. Henny Yeung, Yue Wang, Fenqi Wang, Jean-Julien Aucouturier, Angelica Lim

## ► To cite this version:

Paige Tuttösi, H. Henny Yeung, Yue Wang, Fenqi Wang, Jean-Julien Aucouturier, et al.. Mmm whatcha say? Uncovering distal and proximal context effects in first and second-language word perception using psychophysical reverse correlation. Interspeech 2024, Sep 2024, Kos Island, Greece. hal-04789026

**HAL Id: hal-04789026**

**<https://hal.science/hal-04789026v1>**

Submitted on 18 Nov 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Mmm whatcha say? Uncovering distal and proximal context effects in first and second-language word perception using psychophysical reverse correlation

Paige Tuttösi<sup>1,2</sup>, H. Henny Yeung<sup>3</sup>, Yue Wang<sup>3</sup>, Fenqi Wang<sup>3</sup>, Guillaume Denis, Jean-Julien Aucouturier<sup>2</sup>, Angelica Lim<sup>1</sup>

School of Computing Science<sup>1</sup>, Dept. of Linguistics<sup>3</sup>, Simon Fraser University, Canada  
Université de Franche-Comté, SUPMICROTECH, CNRS, Institut FEMTO-ST<sup>2</sup>, France

ptuttosi@sfu.ca

## Abstract

Acoustic context effects, where surrounding changes in pitch, rate or timbre influence the perception of a sound, are well documented in speech perception, but how they interact with language background remains unclear. Using a reverse-correlation approach, we systematically varied the pitch and speech rate in phrases around different pairs of vowels for second language (L2) speakers of English (/i/-/ɪ/) and French (/u/-/y/), thus reconstructing, in a data-driven manner, the prosodic profiles that bias their perception. Testing English and French speakers ( $n=25$ ), we showed that vowel perception is in fact influenced by conflicting effects from the surrounding pitch and speech rate: a congruent proximal effect 0.2s pre-target and a distal contrastive effect up to 1s before; and found that L1 and L2 speakers exhibited strikingly similar prosodic profiles in perception. We provide a novel method to investigate acoustic context effects across stimuli, timescales, and acoustic domain.

**Index Terms:** speech perception, context effects, reverse correlation

## 1. Introduction

In human-to-human interaction, the ability of a speaker to adapt to an interlocutor is invaluable. Humans will modify their speech to interlocutors with reduced comprehension abilities, e.g., babies [1] or second language (L2) speakers [2]. Moreover, taking cues from multiple modalities, a speaker is able to perceive when they are not being understood and make adjustments to their speech production to increase clarity, both in terms of linguistic [3, 4] and paralinguistic (e.g. emotional [5, 6]) contexts. While synthesized speech is quickly approaching human speech clarity and naturalness, it often still lacks the ability to adapt to interlocutors, especially in a controllable manner. A progression towards adaptive synthesised speech is to first understand, in a fine grained manner, how speech can be adjusted to facilitate perception.

One classical view on vowel perception is that it is a local process categorizing speech sounds by comparison with a pre-learned auditory representation, presumably a spectral one in formant space [7]. Speech sounds, however, are highly variable within and across speakers and it is widely documented that word perception also operates relative to its acoustic context [8, 9]. For instance, following a phrase that is spoken quickly, a sound can be perceived to be longer than it actually is [10].

There is debate, however, about the exact temporal characteristics of such context effects in ecological sentences, and whether the locus of contextual influences depends on the speech cue that is involved [11]. For instance, [10] suggested that the influence of preceding speech rate on phoneme distinction may be limited to a temporal window of one or two adjacent

phonemes, while [12] suggest that spectral context effects result from a form of speaker’s vocal tract length normalization, which benefits from exposure to long-term spectral cues accumulated over possibly several sentences. Whether and how such proximal and distal effects can interact or compete, and the exact timescale at which they occur, remain poorly understood [8].

Determining the acoustic and temporal characteristics of information intake in sentence perception is methodologically challenging for speech-perception research. While hypothesis-driven experimental paradigms can establish the causal influence of specific cues on word contrasts by systematically varying their intensity (e.g. 13 distal speech rates on the perception of a target word in [13]), assessing the relative weight of several temporal regions of interest becomes quickly impractical [10]. To document such temporal dynamics, several studies have relied on eye-tracking in visual search tasks (e.g. printed words, or a ‘visual world’ paradigm, where objects corresponding to each word are shown), and compared the time course of eye fixations to the occurrence of contextual cues [14, 11]. However, this type of study can be underpowered, and there may be other methodologies that permit the automatic extraction, in a data-driven manner, of a listener’s mental representation of what acoustic profile (i.e. what cues, and where) drives a specific speech sound contrast in one direction or another.

In this work, we investigate the use of a classic experimental method, psychophysical **reverse-correlation** [15], which has seen a recent surge of interest in speech perception research for its ability to uncover an individual’s mental representation of what prosodic pattern drives, e.g., judgements of a speaker’s dominance [16] or confidence [17]. Specifically, we use a phase-vocoder technique to systematically and concurrently vary the pitch and speech rate in phrases surrounding pairs of vowels/words that are known to be difficult for L2 speakers: English (/i/-/ɪ/) and French (/u/-/y/); we then use reverse correlation to reconstruct the prosodic profiles that bias the perception of these word pairs in one direction or the other.

To illustrate the ability of this procedure to identify fine differences between groups of participants, we collected data from bilingual French and English speakers, both on the same French and English stimuli. The research literature is equivocal on whether non-native language processing is able to recruit similar acoustic context mechanisms as in native language. For instance, Kang, Johnson, and Finley [18] found that French-L2 speakers failed to use vowel context (i.e. to compensate for coarticulation) when judging fricative sounds when such vowels were unfamiliar; but others have found differing results [19]. Here, we specifically ask what surrounding pitch and speech rate cues L1 speakers use to differentiate pairs of sounds, and whether L2 speakers are sensitive to the same prosodic profiles.

## 2. Procedure & Stimulus generation

### 2.1. Procedure

Reverse correlation is an experimental paradigm aimed at discovering the signal features that govern a participant’s judgement by analysing their responses to large sets of stimuli whose acoustic characteristics have been systematically manipulated [20]. Specifically, we presented participants with a series of 250 trials, each consisting of a single base recording containing an ambiguous target word. For each trial, the base recording was manipulated with a different random profile of pitch and speech rate, and participants were asked which of two target words they heard (1-interval, 2-alternative forced choice). Responses are then analysed to reconstruct the prosodic profile that maximizes the likelihood of responding to one option or the other (for a review, see [21]).

This study included 4 different experiments: (A) two involving random manipulations of the isolated target word (English or French), aimed at establishing a baseline for the intrinsic pitch and rate of the two alternatives, and (B) two involving manipulations of phrases containing the target word, aimed at uncovering extrinsic acoustic context effects. When participants took part in more than one experiment, the order was randomized across language and type of stimuli (word and phrase). For each experiment, the order of response options (e.g. “pill/peel” or “peel/pill”) was randomized across participants. Each experiment lasted, on average, 12 minutes.

### 2.2. Participants

N=114 participants took part in the study: N=54 French-L1 speakers (female: 24, M=32.4yo  $\pm$  9.7) and N=60 English-L1 speakers (female: 33, M=30.0yo  $\pm$  10.7). Each experiment (2 word tasks, 2 phrase tasks) was carried out with  $n=25$  French-L1 and  $n=25$  English-L1 speakers, some of whom participated in more than one experiment. Participants had a self-rated proficiency in the L2 language (English or French) ranging from 2-5 (1: no proficiency, 5: fluent) with a mode of 4. English participants were recruited primarily in anglophone Canada from Simon Fraser University and French participants in France from SUPMICROTECH-ENSMM, as well as on Prolific. The study received internal ethics approval from both universities.

### 2.3. Target words

In each language, we selected pairs of vowels that are known to be difficult for English or French L2 speakers. For English, we selected /i/ and /ɪ/. The vowel /i/, as in “beat” /bit/, is a high, tense vowel that exists in both French and English. In contrast, vowel /ɪ/, as in “bit” /bit/, is lower and lax, and does not exist within the French phonetic inventory. Often, French-L1 speakers learning English will replace /ɪ/ by /i/, for example, saying “sheep” [ʃip] when they mean to say “ship” /ʃɪp/ [22].

For French, we selected the vowels /u/ and /y/. Once again, this pair of vowels is known to be difficult for English-L1-French-L2 speakers in both perception and production [23, 24]. The vowel /u/, such as in “fou” /fu/ (mad), is a high, back vowel, and exists in both French and English. In contrast, the vowel /y/, such as in “fût” /fy/ (cask), is a high, front vowel, and does not exist within the English lexicon. English speakers tend to overcompensate for the lack of /y/ vowel in their language production and often replace /u/ with /y/, such as mispronouncing “beaucoup” /boku/ (a lot) with the rather immodest phrase [boky] [25].

We selected words that had the same beginning and ending

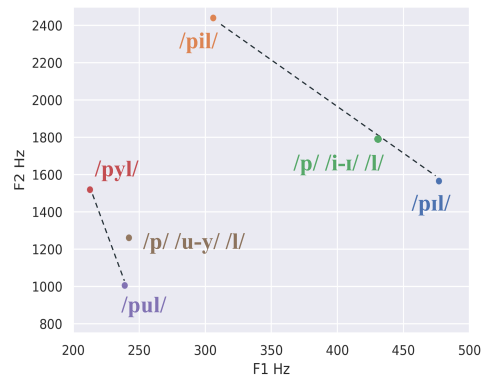


Figure 1:  $F1$ - $F2$  plot of initial and final formants for ambiguous vowels /u-y/ and /i-ɪ/.

consonants across languages to maintain consistency in confounding factors such as co-articulation. The selected words were “pill” /pil/ and “peel” /pi:l/ in English, and “pull” /pyl/ and “poule” /pul/ in French.

### 2.4. Phrase stimuli

For our phrase experiments, we used the phrase “I heard them say” (FR: “je l’ai entendu dire”) preceding the word in an attempt to control contextual bias. The phrases were generated using the Hugging Face interface for CoquiXTTS<sup>1</sup>. The language was set to English for the English phrase and French for the French phrase. An L1 male reference voice was provided and no other modifications were made to the TTS settings.

### 2.5. Stimulus manipulation

To control for response bias in the 1-interval, 2-alternative task (i.e. one alternative being judged more common than the other), we generated morphed sounds intermediate between each of the two vowel pairs /i/-/ɪ/ and /u/-/y/. To do so, we used the Praat software [26] to modify the F1 and F2 formants of each original target word, increasing their resemblance by steps of 10Hz (original formants: /i/ F1: 305.89Hz, F2: 2440.77Hz; /ɪ/ F1: 476.85Hz, F2: 1565.45Hz; /u/ F1: 238.90Hz, F2: 1005.67Hz; /y/ F1: 212.61Hz, F2: 1519.49Hz). We then modified the Whisper ASR algorithm [27] (v20231117, medium multilingual) to access the log-probability of the two possible word interpretation at every step in the formant grid; using the difference in these log-probabilities to select the transformation step that resulted in the smallest difference in prediction probability between the two target words. The resulting formants can be seen in Fig. 1 with the final /i/-/ɪ/ vowel having F1: 436.77Hz, F2: 1722.68Hz (synthesized from “peel” /pi:l/), and the final /u/-/y/ vowel having F1: 238.13Hz, F2: 1258.68Hz (synthesized from “poule” /pul/). These ambiguous words served as base stimuli in the word-task, and were inserted manually in their original phrase (at a zero-crossing 120ms after the end of the last word “say/dire”) to serve as base stimuli for the phrase-task.

Second, we generated the reverse-correlation stimuli from these ambiguous base sounds using the open-source CLEESE toolbox [28]. A voice transformation toolbox that creates random fluctuations around an audio file’s original contour of pitch and speech rate. The pitch contour of the recordings was artificially flattened to a constant 120Hz. We then transformed the

<sup>1</sup><https://huggingface.co/spaces/coqui/xtts>

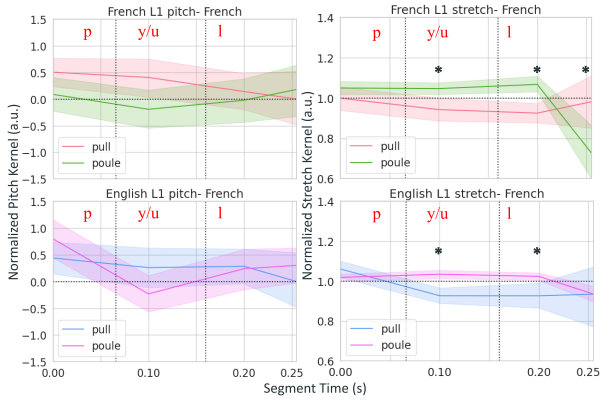


Figure 2: Pitch (left) and speech rate (right) kernels for French-L1 (top) and L2 (bottom) speakers for the French words “pull” and “poule”. In all figures, colored areas mark 95% confidence intervals on the mean, and \* marks time segments that differ statistically at  $\alpha = 0.05$ . Smaller values for the rate kernel mean shorter duration, i.e. faster speech rate.

stimuli by randomly manipulating their pitch and duration independently in  $n$  successive windows of 100ms ( $n=4$  for words;  $n=13$  for phrases), each of which with a factor sampled from a normal distribution (pitch:  $\mu=0$ ,  $\sigma=100$  cents (i.e. 1 semitone); duration:  $\mu=0\%$ ,  $\sigma=1\%$  (i.e. doubling or halving the window’s duration); both distributions clipped at  $\pm 2\sigma$ ). These values were chosen so as to cover the range observed in naturally produced utterances and were linearly interpolated between successive time points to ensure a natural sounding transformation.

### 3. Results

#### 3.1. Validation of ambiguity

We explored our success at creating ambiguous sounds using our combined PRAAT/Whisper approach by analysing participant response bias. Because random manipulations were centered on zero, we expected a 50% response rate for all alternative options. For the word experiments, English-L1 speakers answered “peel” 52% of the time and “poule” 64% of the time; French-L1 speakers answered “peel” 54% of the time and “poule” 70% of the time. For phrases, English speakers responded with 59% “peel” and 59% “poule”, while French-L1 speakers’ responses were 54% “peel” and 66.6% “poule”. It appears that it was more difficult to manipulate the original sound for native French speakers in French, especially for a lone word; discussions with participants concurred with these results.

#### 3.2. Word reverse-correlation

For each participant and each word task (EN/FR), we computed first-order kernels from reverse-correlation data using the *classification image* method [21]. We computed the average random pitch and speech rate transformation profile of the recordings classified as one response option (e.g. “pill”), and compared it with the average pitch and speech rate profile of the recordings classified as the other response (e.g. “peel”). Kernels were normalized by dividing them by the root-mean-square sum of their values. For each participant and each task, this procedure resulted in two dim=4 vectors of pitch and rate of speech values for words, representing what pitch and speech rate transformations should be applied to a given base word in order to increase the likelihood of recognizing a particular target. In the follow-

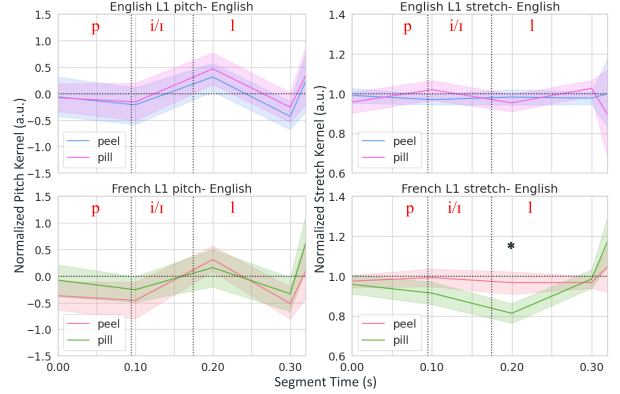


Figure 3: Pitch (left) and speech rate (right) kernels for English-L1 (top) and L2 (bottom) speakers for the English words “peel” and “pill”. Smaller values for the rate kernel mean shorter duration, i.e. faster speech rate.

ing, we test for differences between pitch and speech rate kernel values in each task using paired t-tests at every time point.

*French words.* Our prediction was that French-L1 (FL1) speakers’ perception of /y/, a high/front vowel, would be driven by higher pitch and faster speech rate compared to /u/. Our results confirmed this (Fig.2) for speech rate (FL1 - 0.1s:  $t(24)=2.80$ ,  $p=.010$ ; 0.2s:  $t(24)=3.52$ ,  $p=.002$ ), and pitch (at 0s and 0.1s, albeit non-statistically). French-L2 (EL1) shared the same pattern of data, with a non-statistical pitch increase at  $t=0.1$ s and faster speech rate at 0.1s:  $t(24)=4.57$ ,  $p<.001$ ; and 0.2s:  $t(24)=2.73$ ,  $p=.012$ . Although statistically weak, this pattern of results (higher pitch/faster rate for /y/) was confirmed by segments located on the target word in the phrase kernels.

*English words.* Our prediction was that EL1 speakers’ perception of /i/, a high/tense vowel, would be driven by higher pitch and slower speech rate compared to /ɪ/. Our results did not confirm this prediction (Fig.3): while the speech rate kernels for /i/ showed a significantly slower rate for L2 speakers (0.2s:  $t(24)=3.13$ ,  $p=.005$ ); for pitch, the kernels were not significant and, if anything, indicated the opposite, a preference for increased pitch in /i/ for L1 speakers. Again, this pattern of results (higher pitch and a slower rate for /i/) was confirmed by segments located on the target word in the phrase kernels.

#### 3.3. Phrase reverse-correlation

For each participant and phrase task (EN/FR), we computed reverse-correlation kernels using the same procedure as for the words, resulting in dim=13 kernels for pitch and speech rate.

*French phrases.* Phrase reverse-correlation data confirmed, this time significantly, that within the target word hearing /y/ is driven by higher pitch in FL1 (1s:  $t(24)=-2.63$ ,  $p=.015$ ) and faster speech rate in both FL1 and EL1 speakers (FL1 - 1.0s:  $t(24)=2.99$ ,  $p=.006$ ; EL1 - 1.0s:  $t(24)=4.04$ ,  $p<.001$ , 1.1s:  $t(24)=3.71$ ,  $p=.001$ ). Regarding the surrounding context of the word, the literature predicts the existence of contextual contrast effects, i.e. that /y/ is driven by *lower* pitch and speech rate before the target word. In actuality, our data revealed a mix of contrastive and congruent contextual influences: pitch (Fig. 4-left) was not associated with any contrastive effect, but rather a proximal congruent effect 200-300ms pre-target (FL1- 0.7s:  $t(24)=-2.18$ ,  $p=.040$ ; EL1- 0.8s:  $t(24)=-4.05$ ,  $p<.001$ ), i.e. an increase of pitch immediately before the target word biases the response towards the higher-pitch alternative /y/. Speech rate

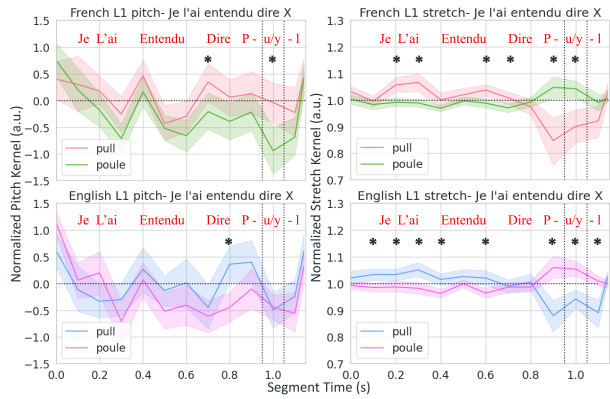


Figure 4: Pitch (left) and speech rate (right) kernels for French-L1 (top) and L2 (bottom) speakers for the French phrases containing “pull” and “poule”. Smaller values for the rate kernel mean shorter duration, i.e. faster speech rate.

(Fig. 4-right) exhibited the expected long-term (distal) contrastive effect (FL1 - 0.2s:  $t(24)=-3.72$ ,  $p=.001$ , 0.3s:  $t(24)=-3.86$ ,  $p=.001$ , 0.6s:  $t(24)=-2.89$ ,  $p=.008$ , 0.7s:  $t(24)=-2.24$ ,  $p=.018$ ; EL1 - 0.1s:  $t(24)=-3.35$ ,  $p=.003$ , 0.2s:  $t(24)=-2.53$ ,  $p=.018$ , 0.3s:  $t(24)=-4.12$ ,  $p<.001$ , 0.4s:  $t(24)=-2.99$ ,  $p=.006$ , 0.6s:  $t(24)=-3.46$ ,  $p=.002$ ), where a slower speech rate at the beginning of a phrase and a proximal congruent effect 100ms pre-target (FL1 - 0.9s:  $t(24)=3.25$ ,  $p=.003$ ; EL1 - 0.9s:  $t(24)=3.76$ ,  $p=.001$ ) biases the response towards the faster alternative /y/, resulting in an overall scissor-shape profile. This pattern of result was remarkably conserved in L2 speakers (Fig. 4-bottom).

*English phrases.* As above, the effect of pitch and rate of speech manipulation within the target word was consistent with the word task: contrary to theoretical predictions, /i/ is driven by higher pitch, significantly in L1 speakers (0.9s:  $t(25)=-2.69$ ,  $p=.013$ ) and, in L2 speakers, by the expected faster speech rate (0.9s:  $t(24)=2.15$ ,  $p=.042$ , 1.0s:  $t(24)=3.61$ ,  $p=.001$ ). Contrary to the French results, pitch showed contrastive effects for L1 speakers within the phrase, both distally (0.1s:  $t(25)=3.23$ ,  $p=.003$ , 0.3s:  $t(25)=3.68$ ,  $p=.001$ ) and in the immediate proximity of the target word (0.8s:  $t(25)=2.91$ ,  $p=.007$ ). Speech rate for both L1 and L2 speakers showed the same scissor-shape pattern as the FR phrases, with distal contrastive effects (EL1 - 0.1s:  $t(25)=-2.52$ ,  $p=.018$ , 0.2s:  $t(25)=-2.31$ ,  $p=.030$ , 0.4s:  $t(25)=-3.47$ ,  $p=.002$ , 0.5s:  $t(25)=-2.34$ ,  $p=.028$ ; FL1 - 0.1s:  $t(24)=-3.02$ ,  $p=.006$ , 0.3s:  $t(24)=-3.53$ ,  $p=.002$ , 0.4s:  $t(24)=-3.07$ ,  $p=.005$ , 0.5s:  $t(24)=-2.75$ ,  $p=.011$ ), and a strong proximal congruent effect (EL1: 0.7s:  $t(25)=3.77$ ,  $p=.001$ , 0.8s:  $t(25)=5.98$ ,  $p<.001$ ), however, the rate difference on the target word was not significant for EL1; FL1: 0.8s:  $t(24)=4.25$ ,  $p<.001$ ). As before, this pattern of result was remarkably conserved across L1 (Fig. 4-top) and L2 speakers (Fig. 4-bottom).

#### 4. Discussion & Limitations

This study provides a proof of concept for using the reverse-correlation paradigm to uncover basic phonetic findings, such as the effects of intrinsic pitch or surrounding speech rate on vowel recognition. In four experiments with  $N=114$  French and English participants, we found systematic acoustic context effects across 2 word pairs and 2 sentences that were, in general, consistent with predictions from the speech perception literature, notably for distal contrast effects [8].

Compared to traditional hypothesis-driven experimental

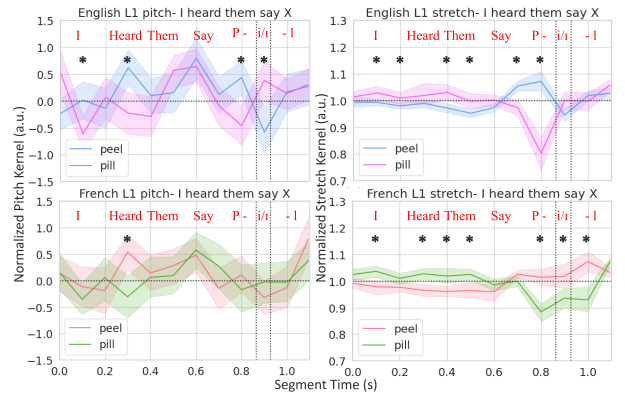


Figure 5: Pitch (left) and speech rate (right) kernels for English-L1 (top) and L2 (bottom) speakers for the English phrases containing “peel” and “pill”. Smaller for the rate kernel mean shorter duration, i.e. faster speech rate.

paradigms, reverse correlation offers several key advantages. First, it allows for uncovering the precise chronometry of context effects and aligning cues with specific phrase elements. In particular, we found that contrastive context effects became congruent proximal to the target, with a tipping point around 200-300ms pre-target. Second, while speech sound perception typically integrates multiple cues and is context dependent [9], phonetic studies are often concentrated on examining a single cue at a time. Reverse correlation offers a methodology to explore combinations of cues at once in a single experiment, producing results more akin to every day human perception. Future work could include additional audiovisual cues, such as formant manipulation [29] or orofacial gestures [30], as well as quantifying the relative perceptual weight of each cue, taking inspiration from feature selection approaches [31].

Despite this potential, the work remains preliminary and suffers from several limitations. First, we saw weaker and less-consistent effects for pitch than speech rate, especially within the FR phrase and in the FL1 group. Since French is not a language with lexically specified word stress, it is possible that pitch cues are less important for FL1 than EL1 speakers. Another possible explanation may be that the selected phrase “je l’ai entendu dire”, with e.g., more consonants than the English phrase, lends itself comparatively poorly to pitch transformations. Future work should therefore investigate the generalizability of the results with more word pairs and more phrase contexts. Second, results in the EN word task (preference for increased pitch in /i/ for L1 speakers) were inconsistent with theoretical predictions of higher intrinsic pitch for /pil/. While this effect was replicated in the phrase task, and was consistent with the contrast effects obtained in the rest of the phrase, it would be interesting to clarify why this occurred and examine possible individual differences in how participants combine pitch and spectral cues in this sound. For example, it is possible that EL1 speakers weigh spectral cues higher than speech rate or pitch cues, whereas the L2 speakers rely more heavily on prosody to disambiguate vowel tensivity [32, 33]. Finally, our experiments used a 1-interval, 2-alternative design, in an attempt to keep the experiment short and feasible in an online setting. However, such as design potentially introduces response bias (section 3.1) and other decision variables that can obfuscate purely phonetic mechanisms (see e.g. pitch kernels that significantly depart from zero in the same direction for *both* response options, Figure 5-left), and it would be interesting to reproduce these results with a longer, 2-interval 1-alternative task.

## 5. Acknowledgements

This work was supported by NSERC Discovery Grant 06908-2019, the France Canada Research Fund, the Mitacs Globalink Research Award, and the Fondation Pour l’Audition (FPA RD-2021-12). The authors thank P. Maublanc, R. Guha, and A. Adl Zarrabi for their valuable discussions; V. Yang, B. Burkanova, C. Zhang, and M. Durana for their help running our study; and the Rajan Family for their support. This work has been conducted in the framework of the EIPHI Graduate school (ANR-17-EURE-0002 contract).

## 6. References

- [1] E. K. McClay, S. Cebioglu, T. Broesch, and H. H. Yeung, “Rethinking the phonetics of baby-talk: Differences across canada and vanuatu in the articulation of mothers’ speech to infants,” *Developmental science*, vol. 25, no. 2, p. e13180, 2022.
- [2] C. Redmon, K. Leung, Y. Wang, B. McMurray, A. Jongman, and J. A. Sereno, “Cross-linguistic perception of clearly spoken english tense and lax vowels based on auditory, visual, and auditory-visual information,” *Journal of phonetics*, vol. 81, p. 100980, 2020.
- [3] T. Biro, A. J. Olmstead, and N. Viswanathan, “Talker adjustment to perceived communication errors,” *Speech communication*, vol. 138, pp. 13–25, 2022.
- [4] K. Maniwa, A. Jongman, and T. Wade, “Acoustic characteristics of clearly spoken english fricatives,” *The Journal of the Acoustical Society of America*, vol. 125, no. 6, pp. 3962–3973, 2009.
- [5] R. Banse and K. R. Scherer, “Acoustic profiles in vocal emotion expression,” *Journal of personality and social psychology*, vol. 70, no. 3, pp. 614–636, 1996.
- [6] J.-A. Bachorowski, “Vocal expression and perception of emotion,” *Current directions in psychological science : a journal of the American Psychological Society*, vol. 8, no. 2, pp. 53–57, 1999.
- [7] A. M. Liberman and D. H. Whalen, “On the relation of speech to language,” *Trends in cognitive sciences*, vol. 4, no. 5, pp. 187–196, 2000.
- [8] C. Stilp, “Acoustic context effects in speech perception,” *Wiley interdisciplinary reviews. Cognitive science*, vol. 11, no. 1, pp. e1517–n/a, 2020.
- [9] B. McMurray and A. Jongman, “What information is necessary for speech categorization? harnessing variability in the speech signal by integrating cues computed relative to expectations,” *Psychological review*, vol. 118, no. 2, p. 219, 2011.
- [10] R. S. Newman and J. R. Sawusch, “Perceptual normalization for speaking rate: Effects of temporal distance,” *Perception & psychophysics*, vol. 58, no. 4, pp. 540–560, 1996.
- [11] E. Reinisch and M. J. Sjerps, “The uptake of spectral and temporal cues in vowel perception is rapidly influenced by context,” *Journal of phonetics*, vol. 41, no. 2, pp. 101–116, 2013.
- [12] K. Johnson, E. A. Strand, and M. D’Imperio, “Auditory–visual integration of talker gender in vowel perception,” *Journal of phonetics*, vol. 27, no. 4, pp. 359–384, 1999.
- [13] C. C. Heffner, L. C. Dille, D. J. McAuley, and M. A. Pitt, “When cues combine: How distal and proximal acoustic cues are integrated in word segmentation,” *Language and Cognitive Processes*, vol. 28, no. 9, pp. 1275–1302, 2013.
- [14] K. B. Shatzman and J. M. McQueen, “Prosodic knowledge affects the recognition of newly acquired words,” *Psychological Science*, vol. 17, no. 5, pp. 372–377, 2006.
- [15] A. A. Jr and J. Lovell, “Stimulus features in signal detection,” *The Journal of the Acoustical Society of America*, vol. 49, no. 6B, pp. 1751–1756, 1971.
- [16] E. Ponsot, J. J. Burred, P. Belin, and J.-J. Aucouturier, “Cracking the social code of speech prosody using reverse correlation,” *Proceedings of the National Academy of Sciences - PNAS*, vol. 115, no. 15, pp. 3972–3977, 2018.
- [17] L. Goupil, E. Ponsot, D. Richardson, G. Reyes, and J.-J. Aucouturier, “Listeners’ perceptions of the certainty and honesty of a speaker are associated with a common prosodic signature,” *Nature communications*, vol. 12, no. 1, pp. 861–861, 2021.
- [18] S. Kang, K. Johnson, and G. Finley, “Effects of native language on compensation for coarticulation,” *Speech Communication*, vol. 77, pp. 84–100, 2016.
- [19] N. Viswanathan, J. S. Magnuson, and C. A. Fowler, “Similar response patterns do not imply identical origins: an energetic masking account of nonspeech effects in compensation for coarticulation,” *Journal of Experimental Psychology: Human Perception and Performance*, vol. 39, no. 4, p. 1181, 2013.
- [20] R. Adolphs, L. Nummenmaa, A. Todorov, and J. V. Haxby, “Data-driven approaches in the investigation of social perception,” *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 371, no. 1693, p. 20150367, 2016.
- [21] R. F. Murray, “Classification images: A review,” *Journal of vision*, vol. 11, no. 5, pp. 2–2, 2011.
- [22] P. Inverson, M. Pinet, and B. G. Evans, “Auditory training for experienced and inexperienced second-language learners: Native french speakers learning english vowels,” *Applied psycholinguistics*, vol. 33, no. 1, pp. 145–160, 2012.
- [23] J. E. Flege, “The production of “new” and “similar” phones in a foreign language: evidence for the effect of equivalence classification,” *Journal of phonetics*, vol. 15, no. 1, pp. 47–65, 1987.
- [24] E. S. Levy, “On the assimilation-discrimination relationship in american english adults’ french vowel learning,” *The Journal of the Acoustical Society of America*, vol. 126, no. 5, pp. 2670–2682, 2009.
- [25] J. L. Sturm, “Explicit phonetics instruction in l2 french: A global analysis of improvement,” *System*, vol. 41, no. 3, pp. 654–662, 2013.
- [26] P. Boersma and D. Weenink, “Praat: doing phonetics by computer,” version 6.4.06, retrieved 25 February 2024 from <http://www.praat.org/>.
- [27] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” 2022.
- [28] J. J. Burred, E. Ponsot, L. Goupil, M. Liuni, and J.-J. Aucouturier, “Cleese: An open-source audio-transformation toolbox for data-driven experiments in speech and music cognition,” *PLoS one*, vol. 14, no. 4, p. e0205943, 2019.
- [29] E. Ponsot, P. Arias, and J.-J. Aucouturier, “Uncovering mental representations of smiled speech using reverse correlation,” *The Journal of the Acoustical Society of America*, vol. 143, no. 1, pp. EL19–EL24, 2018.
- [30] M. Liu, Y. Duan, R. A. Ince, C. Chen, O. G. Garrod, P. G. Schyns, and R. E. Jack, “Facial expressions elicit multiplexed perceptions of emotion categories and dimensions,” *Current Biology*, vol. 32, no. 1, pp. 200–209, 2022.
- [31] S. Garg, G. Hamarneh, A. Jongman, J. A. Sereno, and Y. Wang, “Computer-vision analysis reveals facial movements made during mandarin tone production align with pitch trajectories,” *Speech Communication*, vol. 113, pp. 47–62, 2019.
- [32] S. Ylinen, M. Uther, A. Latvala, S. Vepsäläinen, P. Iverson, R. Akahane-Yamada, and R. Näätänen, “Training the brain to weight speech cues differently: A study of finnish second-language users of english,” *Journal of Cognitive Neuroscience*, vol. 22, no. 6, pp. 1319–1332, 2010.
- [33] C. Redmon, K. Leung, Y. Wang, B. McMurray, A. Jongman, and J. A. Sereno, “Cross-linguistic perception of clearly spoken english tense and lax vowels based on auditory, visual, and auditory-visual information,” *Journal of phonetics*, vol. 81, p. 100980, 2020.