Perceptual relevance of long-domain phonetic dependencies

Noël Nguyen, Zsuzsanna Fagyal, Jennifer Cole

Perceptual relevance of long-domain phonetic dependencies

Noël Nguyen

Laboratoire Parole et Langage, UMR 6057
CNRS & Université de Provence, Aix-en-Provence, France
noel.nguyen@lpl.univ-aix.fr

Zsuzsanna Fagyal

French Department, University of Illinois at Urbana-Champaign, USA
zsfagyal@uiuc.edu

Jennifer Cole

Department of Linguistics, University of Illinois at Urbana-Champaign, USA
jscole@uiuc.edu

Abstract

This study is concerned with the perceptual relevance of regressive vowel harmony in French. Two experiments were conducted. The first experiment showed that acoustic variations in a non-final vowel depending on the final vowel can be detected by listeners. The second experiment revealed that vowel harmony can facilitate the identification of the final vowel. Implications for current models of speech perception are discussed.

1 Introduction

Previous studies have shown that listeners are sensitive to acoustic cues to phonological contrasts that are distributed over long temporal intervals in the perception of speech (e.g. Martin and Bunnell, 1981, 1982). In many current models of speech perception and word recognition, however, listeners focus upon acoustic properties that are confined to short stretches of time. For example, in Marslen-Wilson and Warren's (1994) direct access featural model, the emphasis is put on coarticulatory effects between segments that are adjacent to each other. In Lahiri and Reetz' (2002) Featurally Underspecified Lexicon model, the speech signal is mapped onto phonological features through a short (20-ms) sliding time window. In Stevens' (2002) model, acoustic cues to distinctive features are assumed to be located in the vicinity of local acoustic landmarks in the speech signal.

Short-domain acoustic cues are usually given a greater perceptual weight for several reasons. First, the auditory trace of continuous speech is known to fade rapidly. Acoustic properties that extend over about a half-second may therefore need to be converted into an intermediate, more abstract representation prior to being passed onto higher levels of processing, contrary to local acoustic cues, which listeners may be able to process more directly (Remez, 2003). Second, local acoustic cues are thought to be generally more prominent than long-domain ones. For example, coarticulatory effects gradually decrease in magnitude as the distance from the assimilating segment increases along the time axis. In most models of speech perception, listeners are assumed to concentrate on the most prominent acoustic properties associated with each phonological contrast, at the expense of more fine-grained properties, and this results in long-domain cues being often regarded as deprived of any real perceptual significance. Third, the segmental approach that often prevails in speech perception studies ("There is ample evidence that words are stored in memory in terms of sequences of segmental units, and that these segmental units are further represented in terms of the values of a set of binary features", Stevens, 2002) has also led researchers to minimize the potential relevance of long-domain acoustic properties for the listener.

This well-established approach has been questioned in several recent studies which suggest that long-domain acoustic properties do matter in speech perception. Hawkins and Slater (1994) found

that when synthesized sentences are presented in noise, including V-to-V coarticulatory effects in these sentences can improve their intelligibility by 15%. West (1999) showed that listeners are sensitive to long-domain resonance effects associated with /l/ and /r/ in the perception of the /l/-/r/ contrast. Hawkins and Nguyen (2003) found that syllable-onset /l/s can provide listeners with cues to voicing in syllable-coda obstruents. These findings are inconsistent with a model of speech perception in which acoustic information is integrated over a time window of short or very short duration.

The present study aims to provide further empirical evidence for the role of long-domain acoustic properties in speech perception. It focusses on the perception of regressive vowel harmony in French. We sought to determine to what extent these effects, which extend across a syllable boundary, between two non-adjacent segments, are used by the listener to speed up lexical access.

Vowel harmony (henceforth, VH) in French is described as a word-level anticipatory process affecting non-final mid vowels that assimilate in height to the final tonic vowel. The non-final vowel will tend to be mid-high before a high or mid-high vowel (e.g. *aimer* [eme] "to love"), and mid-low before a low or mid-low vowel (*aimable* [ɛmabl] "kind"). Most authors consider that VH is optional, possibly speaker-dependent, and is more likely to occur in informal speech. In a recent companion study (Fagyal et al., 2003; Nguyen and Fagyal, 2003), speech data were collected for seven native speakers of French with a view to explore the acoustic correlates of VH. Both spectral and durational differences were found in mid vowels depending on the following, word-final vowel. Mid vowels tended to be slightly longer, and have a lower F1 frequency and/or a more extreme F2 frequency before a mid-high vowel than before a mid-low vowel. These findings are in keeping with the assumption that VH is an assimilation in vowel height, although an alternative interpretation involving the tense/lax distinction can be offered.

2 Experiment 1: Perceptual relevance of vowel harmony

The goal of this experiment was to determine to what extent V2-to-V1 assimilatory effects can be perceived by listeners. To address this issue, we employed a modified version of the ABX discrimination task, designed by Majors (1998), and which allowed the natural recordings analyzed in the production study to be used as stimuli. Listeners were presented with series of three (C)V sequences and were asked to determine whether the third sequence was more similar to the first one or to the second one. Unlike the standard ABX design, these sequences were all acoustically different from each other. Two of them, however, had been extracted from the same phonetic context. As an example, one of the series used was [no no no], where the first and third sequences had been excerpted from two tokens of *notice* [notis] and the second one from *nota* [nota]. Listeners were in that case expected to judge the third sequence to sound more like the first one than the second one, assuming that V2-dependent variations would perceptually prevail upon all other acoustic differences among the stimuli.

2.1 Method

The material was extracted from 29 pairs of disyllabic words divided into two subsets, referred to as the *épice-épate* and *notice-nota* subsets (using one of the word pairs included in each subset as a generic label). The first syllable contained a mid vowel that was phonemically identical in both words (either /e/ or /o/). The second syllable contained the vowel /i/ for one member of the pair and the vowel /a/ for the other. The syllable boundary fell between the first vowel and the onset of the second syllable (either a single consonant or a consonant cluster). These words were recorded twice, in isolation, by a native speaker of the Norther variety of French spoken in Paris (Speaker B).

The first syllable in each word was segmented and saved to a separate file. The signal amplitude was linearly decreased to 0 over the last 30 ms at the end of the vowel to avoid any audible click, and

the maximum amplitude value was normalized over the entire set of stimuli. The two repetitions for each word pair were then used to construct triads of the form $A_1B_1A_2$, $B_1A_1A_2$, $A_1B_1B_2$, and $B_1A_1B_2$, where A and B refer to the first and second member of a word pair, respectively, and the index (1 or 2) refers to the repetition. The duration of the ISI was 500 ms.

There were 116 triads in total, which were presented in four blocks. The four triads associated with each word pair appeared in separate blocks. The triads within each block were played in a random order to the listener.

Six trained phoneticians (all native speakers of French) took part in the experiment. They were asked to identify which of the two first stimuli the third one sounded more like, by pressing a button on a response box.

2.2 Results

Mean proportions of correct responses for each V1-V2 combination (referred to using one of the carrier words) over the six listeners are shown in Figure 1 (left panel). For example, Figure 1 indicates that V1 in an /e/–/i/ word such as *épice*, repetition 2, was perceived as more like V1 in the same word, repetition 1, than V1 in the corresponding /e/–/a/ word *épate*), in 60% of the cases (leftmost vertical bar). The mean proportion of correct responses was above the 50% chance threshold for all V1-V2 combinations. A three-way chi-square test showed that the difference between the proportion of correct responses and that of incorrect responses was statistically significant ($\chi^2 = 40.73, \mathrm{df} = 1, p < 0.001$).

2.3 Discussion

Our results show that initial syllables were judged to sound more like each other when originating from two repetitions of the same disyllabic word, compared to a pair of words with different final syllables. Since the most systematic difference between the two members of each word pair lied in the final vowel (/i/ for one member and /a/ for the other), it may be assumed that the observed response patterns reflect the effect that final vowels have on how initial vowels (and possibly word-onset consonant(s), if any), are produced. This effect appears to be more perceptually salient than the acoustic variations that occur between repetitions for each word.

The subjects in this pilot experiment were trained phoneticians who were expected to experience less difficulty than naive listeners with the modified ABX discrimination task, which involves making two-way comparisons between syllables that show small, subphonemic differences. It therefore remains to be ascertained whether our results can be generalized to naive listeners. In addition, although our data suggest that listeners are able to detect acoustic variations in V1 depending on V2, we need to determine to what extent listeners can anticipate the identity of V2 on perceiving V1. These two issues are addressed in the next experiment.

3 Experiment 2: Role of vowel harmony in V2 detection

This experiment aimed to examine whether vowel harmony facilitates the identification of the final vowel in a disyllabic word. Drawing on a previous series of experiments carried out by Martin and Bunnell (1981, 1982) on the perception of V-to-V anticipatory coarticulation in English, we had listeners perform a V2 detection task in words whose initial syllable either was the original one, or was cross-spliced from another word with a different final syllable. In the cross-spliced version of *notice* [notis], for example, the first syllable was excerpted from *nota* [nota]. We assumed that the initial vowel would in that case contain inappropriate cues to the identity of the final vowel and that listeners would therefore detect the final vowel less rapidly in the cross-spliced version compared to the original version.

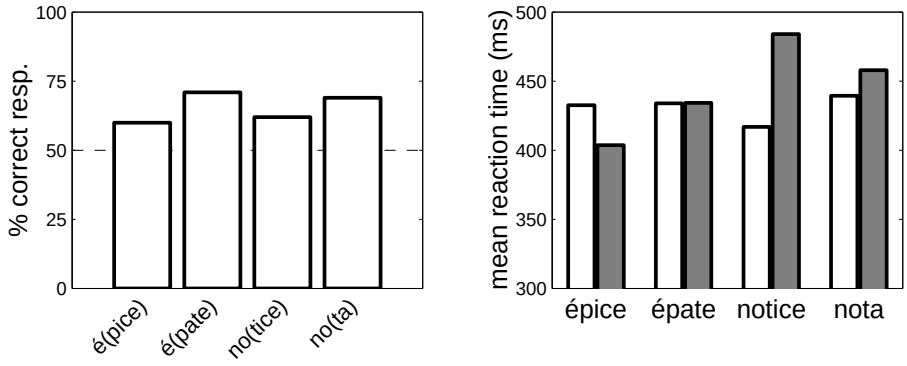

Figure 1: Left panel: Mean proportion of correct responses for each V1-V2 combination in the ABX discrimination task. Right panel: Mean reaction times for hit responses in the V2 detection task, for each V1-V2 combination and for both the original (white bars) and cross-spliced (grey bars) stimuli.

3.1 Method

The material was made up of 24 pairs of words from the *épice-épate* and *notice-nota* sets as recorded within a carrier sentence (first repetition) by Speaker B in the production study. The maximum amplitude value was normalized across sentences. Cross-spliced versions of the sentences were constructed by exchanging target words' initial syllables (as well as the preceding words) in each pair. A careful listening of the material revealed no audible discontinuity at the cross-splicing point. This yielded 96 sentences (48 original, 48 cross-spliced) that were presented twice to each listener in 4 blocks of 48. Each sentence appeared once (either in the original or cross-spliced version) in each block. The two members of a given sentence pair appeared one in the first half of the block and the other in the second half. The order of the sentences within each half-block was randomized.

Sixteen naive undergraduate students in language sciences took part in the experiment. Sentences were presented to them over headphones, with a 3-s ISI. The subjects were asked to press a button on a response box as fast as possible whenever a vowel visually specified prior to the presentation of the sentence was contained in the final syllable of the target word. The vowel to detect alternated between *i* and *a* from one block to the following one. The subject's reaction time was measured from the offset of the first vowel in the target word.

3.2 Results

Our analyses were restricted to hit responses for which the reaction times were comprised between 200 and 1200 ms. On average, the reaction time was 14 ms longer for the cross-spliced stimuli (445 ms) compared to the original stimuli (431 ms), and this difference proved statistically significant in a by-subject repeated-measure ANOVA with V1, V2 and cross-splicing as independent variables ($F(1,14) = 6.914, p < 0.05$). The mean reaction times associated with hit responses for each V1-V2 combination (referred to using one of the carrier words) and for both the original and cross-spliced stimuli are shown in Figure 1 (right panel).

There were significant V1 × cross-splicing as well as V1 × V2 × cross-splicing interactions ($F(1,14) = 7.215, p < 0.05$; $F(1,14) = 9.08, p < 0.01$). Figure 1 reveals that the mean RT difference between the original and cross-spliced stimuli was greater for the /o/-/i/ set of words (+67 ms) than for the other words. Oneway ANOVAs performed for each word set separately indeed showed that this difference reached statistical significance for the /o/-/i/ words only ($F(1,14) =$

$16.826, p = 0.001$).

3.3 Discussion

When the first vowel contained conflicting cues to the second vowel, the subjects' responses were delayed relative to a control condition, and this delay proved significant for the /o/-/i/ words. Thus, our results partly confirm the assumption that listeners are sensitive to regressive vowel harmony effects and use them to predict the identity of an upcoming vowel.

Two factors may contribute to explain the fact that cross-splicing had no measurable impact on the response patterns for the *épice-épate* set of word pairs. First, the interval between V1 offset and V2 onset was on average longer for the *épice-épate* set (100 ms) than for the *notice-nota* set (80 ms). Since the cross-splicing point was located at the end of V1, a longer intervocalic interval may have made the mismatch between V1 and V2 less perceptually salient. Second, vowel harmony effects were found to be acoustically larger and more systematic in the production study for the *notice-nota* set than for the *épice-épate* set, particularly as regards F2 frequency (F2 in V1 was higher in frequency before /i/ than before /a/ by 51 Hz on average for the *épice-épate* words, and by 61 Hz for the *notice-nota* words). It may be the case that larger F2 variations in a lower frequency range (i.e. in a back rounded vowel, as opposed to a front unrounded one) had a greater influence on the identification of the following vowel.

Our results also reveal a perceptual asymmetry between the *notice* and *nota* words. Replacing the first syllable of *notice* by that of *nota* had a disruptive effect on V2 detection, whereas replacing the first syllable of *nota* by that of *notice* only led to a nonsignificant increase in V2 detection time. A potential explanation for this asymmetry lies in the phonetic quality of /o/ in the material used for this experiment. Regardless of the phonetic context, /o/ was produced with an advanced tongue position, as is common in the Northern variety of French spoken by our speaker. It may therefore be possible that the fronter quality shown by /o/ in the cross-spliced version of *nota* was not judged by listeners as being inconsistent with the following vowel being /a/. Conversely, the backer quality of /o/ in the cross-spliced version of *notice* may have been perceived as inappropriate in the context of a following front vowel.

4 General discussion

Taken together, Experiments 1 and 2 suggest that regressive vowel harmony in French is perceptually relevant and may facilitate the identification of word-final vowels. Our findings lend support to the view that long-domain acoustic properties associated with phonological contrasts may be brought into play along with short-domain properties in the perception of speech. They more particularly show that acoustic information is integrated over a time window that may extend across one syllable boundary at least. In that respect, these findings are at variance with models of speech perception that exclusively focus upon local acoustic properties. In addition, our results do not seem to be compatible with theories that view syllables as the primary units of perception, and therefore as the locus of acoustic information integration.

Our production study (Fagyal et al., 2003; Nguyen and Fagyal, 2003) indicated that vowel harmony in French is a fine-grained and accent-specific phenomenon. Acoustic variations shown by the first vowel depending on the second vowel were subtle and mostly confined to the Parisian variety of French, as opposed to the South-East variety. Thus, it is unsurprising that the perceptual effects reported here were small, all the more since listeners in Experiment 2 (tested in Aix-en-Provence) are for most of them speakers of SE French, and may therefore not be used to paying attention to vowel harmony effects as a reliable source of information in lexical access. In such conditions, it is in fact interesting to observe that vowel harmony did have an effect on the listeners' responses. V2-dependent variations in V1 can at best be considered as enhancing the primary cues to V2 identity,

which are located in the core portion of V2 itself and at the vowel's edges (in the vicinity of the boundary with the preceding consonant, in particular). The fact that listeners proved sensitive to vowel harmony may be taken as evidence for exemplar-based models of speech perception, in which each word is associated with a set of detailed phonetic representations in long-term memory. In such a framework, vowel harmony patterns would be directly stored in the exemplars alongside primary segmental cues. If we assume that word recognition involves a direct mapping between the speech input and the exemplars, vowel harmony would contribute to facilitating the lexical search, by enhancing the level of activation of the target word as the non-final (assimilated) vowel is being processed. This exemplar-based account leads to some interesting predictions. For example, it may be assumed that a listener should be more sensitive to vowel harmony when the speaker produces it in a more systematic manner, and when the listener is more familiar with the speaker's voice. In such a case, a detailed representation of vowel harmony patterns should form in the listener's memory, whereas smaller and/or less consistent vowel harmony effects should give rise to more generic and abstract memory traces. These issues will be explored in future work.

Acknowledgements

This work is partly supported by the "France and the World" cooperative research program between the University of Illinois at Urbana-Champaign and the CNRS.

References

Fagyal, Z., Nguyen, N., and Boula de Mareüil, P. (2003). From dilation to coarticulation: is there vowel harmony in French? *Studies in Linguistic Sciences*, 32:1–21.

Hawkins, S. and Nguyen, N. (2003). Effects on word recognition of syllable-onset cues to syllable-coda voicing. In Local, J., Ogden, R., and Temple, R., editors, *Papers in Laboratory Phonology VI*, pages 38–57. Cambridge University Press, Cambridge, UK.

Hawkins, S. and Slater, A. (1994). Spread of CV and V-to-V coarticulation in British English: Implications for the intelligibility of synthetic speech. In *Proceedings of ICSLP 94*, volume 1, pages 57–60, Yokohama.

Lahiri, A. and Reetz, H. (2002). Underspecified recognition. In Gussenhoven, C. and Warner, N., editors, *Papers in Laboratory Phonology VII*, pages 637–675. Mouton de Gruyter, Berlin, Germany.

Majors, T. (1998). *Stress dependent harmony: Phonetics origin and phonological analysis*. PhD thesis, University of Texas, Austin, TX.

Marslen-Wilson, W. and Warren, P. (1994). Levels of perceptual representation and process in lexical access - words, phonemes, and features. *Psychological Review*, 101:653–675.

Martin, J. and Bunnell, H. (1981). Perception of anticipatory coarticulation effects. *Journal of the Acoustical Society of America*, 69:559–567.

Martin, J. and Bunnell, H. (1982). Perception of anticipatory coarticulation effects in vowel-stop consonant-vowel sequences. *Journal of Experimental Psychology: Human Perception and Performance*, 8:473–488.

Nguyen, N. and Fagyal, Z. (2003). Acoustic aspects of vowel harmony in French. In *Proceedings of the XVth International Congress of Phonetic Sciences*, pages 3029–3032, Barcelona, Spain.

Remez, R. (2003). Establishing and maintaining perceptual coherence: Unimodal and multimodal evidence. *Journal of Phonetics*, 31:293–304.

Stevens, K. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *Journal of the Acoustical Society of America*, 111:1872–1891.

West, P. (1999). Perception of distributed coarticulatory properties in English /l/ and /ɹ/. *Journal of Phonetics*, 27:405–426.