Tracking the time course of perception of coarticulation: A replication

Robert Xu

1 Introduction

1.1 Background

This study is a replication of Beddor et al. (2013), where the authors investigated the time course of perception of coarticulation, specifically nasalization, using an eye tracking experiment. Coarticulation is the result of overlapping gestures of the speech organs in speech production. Due to the neighboring or nearby sounds, sounds deviate from their "canonical" forms and have context-specific articulatory and acoustic realizations. Although coarticulation could obscure the target sound, more recent theories in speech perception stipulates that coarticulation could provide useful information for the listeners. In this line of thought, listeners are regarded active participants in speech communication, and process the rich, time-varying speech input signal using all sorts of available information including contextual information like coarticulation. If this were true, it would be expected that listeners would determine the following sound once anticipatory coarticulatory information occurs in the signal rather than waiting to hear the actual sound.

This study investigated one kind of coarticulation commonly occurring in American English: nasalized vowels before nasals. There are no phonemic nasalized vowels in American English. However, the velum typically lowers in anticipation of a following nasal consonant during the production of a vowel, coloring the vowel with a nasalized flavor. Listeners could use the nasalization in their perception, supported by various studies cited in Beddor et al. (2013) using a wide range of experimental paradigms. Malécot (1960)'s early tape-editing experiment showed that listeners could identify a VN sequence by the nasalized vowel alone especially if it is followed by a voiceless coda.

Beddor (2009) found that listeners would need a shorter N if the nasalized vowel was longer to identify the postvocalic nasal consonant. Other studies (Flagge et al. 2006, Warren et al. 1987, Lahiri et al. 1991, Ohala et al. 1995) have shown that listeners would react faster to VN sequences with a nasalized V than those without on both the behavioral (identification, gating paradigm) and the neural level (MEG).

However, although these past studies have shown that listeners indeed benefit from the nasalization in their perception, these experimental techniques could not show how the listeners are using the coarticulatory information as the perception takes place in real time. Therefore, Beddor et al. (2013), replicated here, takes a further step by using the visual world paradigm, in order to track the moment-by-moment perception of the nasalization. In this paradigm, listeners are presented with both audio and visual stimuli. Their visual fixation in response to the audio is tracked and continuously updated throughout each trial. Therefore it generates rich data sensitive to the time course of the perception in comparison to reaction time or the gating paradigm which only have a single judgment thus one data point per perception. As Beddor et al. (2013) points out, the eye tracking study is also different from the MEG study, which also offers a good temporal resolution, in that MEG shows neural responses to passive presentations of coarticulation while the eye-tracking paradigm "provides direct evidence of active use of coarticulatory information".

In addition, two types of findings motivate the design of this study. First, the duration and quality of the nasalization vary due to both linguistic and sociolinguistic factors. The different intrinsic velo-settings of different vowels may influence the quality of the nasalization, while the extent of the nasalization can be highly variable across dialects and speakers. Second, the duration profile of the nasalization is systematically sensitive to the voicing of the coda in a CVNC syllable, which could also be useful information for listeners in perception.

1.2 The current study

In this study, listeners hear either a CVC or a CVNC word, and are instructed to look at that word on the screen against a competition word. The primary dependent variable is the latency of the first correct fixation after they hear the target word. Two hypotheses are

tested (the third hypothesis in Beddor et al. (2013) was not tested in order to deduce the duration of the experiment).

Hypothesis 1 states that participants use the nasalization of a vowel to recognize the following nasal shortly after that information is available in the speech signal. Therefore, when listeners listen to a CVNC sequence with a nasalized vowel, their visual fixation should start before they hear the N. Furthermore, if the nasalization starts earlier, the fixation should be earlier as well.

Hypothesis 2 claims that participants that the listeners will use the coarticulatorily nasalized vowel as a better indicator of a following N than using the oral vowel an indicator of a following C. This is based on the asymmetry that a nasalized vowel appears to be more informative than an oral one due to the listeners' compensation of coarticulation. This predicts that the latency of the fixation should not be different between a word with or without N when listeners hear a word with an oral vowel.

2 Methods

2.1 Participants

In comparison to the original study, which has twenty-three college participants, this replication only has two listeners. They are both native English speakers from North America, without any known hearing deficits and with normal/corrected-to-normal vision. An additional participant was recruited but was excluded because of failure in tracking her eye movements with the eye-tracker.

2.2 Stimuli

Three out of five sets of minimal CV(N)C quadruplets from Beddor et al. (2013) were used as stimuli in this experiment. The two sets were not used due to consideration of experiment duration. They are also the sets that have the lowest lexical frequency. Within each set of the quadruplets, as shown in Table 1, the words differ in the presence of the nasal consonant and in the voicing of the final consonant. The auditory stimuli were produced by a female native speaker, as oppose to a male speaker in beddor et al. (2013).

The speaker was recorded in a sound-treated room saying randomized repetitions of these words embedded in the carrier phrase "Please look at ____". This is the same carrier phrase used in the experiment for pitch and intonation consideration, as opposed to the original experiment using a simpler carrier phrase "Say ____".

Table 1 The quadruplets of stimuli

CVT	CVD	CVNT	CVND
bet	bed	bent	bend
let	led	lent	lend
set	said	sent	send

In order to control the time course of the coarticulatory information over the words, the auditory stimuli provided to the experiment participants were manipulated versions of the original recording. All stimuli were cross-spliced in Praat by wave-editing. For each CVC-CVNC pair of matching voicing, the initial CV sequence was taken from an original CVC token in the recording (e.g. for a *bet-bent* pair, [b] and the onset of [e] are taken from an original *bet* token). For the CVC tokens for the experiment, Beddor et al. (2013) took the remainder of the word from a second CVC token. However, because this did not work out for all the words in my recordings due to unnaturalness caused by creakiness and pitch difference, I took the remaining part from the same CVC word. Because the duration of my recorded words were longer than those used in Beddor et al. (2013), this means that I cut out some waves in the middle of the words for each CVC words. For the CVNC tokens for the experiment, I followed the Beddor et al. (2013) treatment where the remaining part came from an original CVNC word (the nasalized vowel, the nasal and the coda). In all the original CVNC words, the nasalization of the vowels was clearly audible and could be identified acoustically with lower amplitude of the formants.

Following this manipulation procedure, one version of CVC and two versions of CVNC tokens were generated. The two CVNC tokens differ in the proportions of the oral versus the nasalized vowels. For the early nasalization onset version, the oral part of the vowel (the part from the original CVC token) takes 20% of the duration of the vowel to the nasalized part (the part from the original CVNC token) with 80%. For the late nasalization onset version, the proportion is 60% to 40%. The duration of the vowel and

nasal parts of each word strictly followed the average duration from Beddor et al. (2013) with no more than 3ms difference, as shown in Table 2.

Table 2 Average duration (in ms) of VN portions of target stimuli in Beddor et al. (2013)

	Oral vowel	Nasalized vowel	Nasal
CVT	134.51		
CVD	185.96		
CVNT early onset	26.85	100.85	51.33
CVNT late onset	79.01	50.76	51.33
CVND early onset	36.00	136.75	92.37
CVND late onset	99.47	75.18	92.37

While Beddor et al. (2013) used drawings to represent each word for the visual stimuli, this replication instead used plain words for the following considerations. First, the drawings may not be the best representation of the words, especially for word pairs like *bend-bent*. Moreover, to train the participants to match the words with the drawings, the participants went through an extensive training session in Beddor et al. (2013), which would take too long to perform in this replication. However, there might be two shortcomings replacing the drawings with orthographical representations. First, the critical visual difference in the orthography (e.g. "n" in *lent-let*, the final letter in *let-led*) would have different distances from the fixation cross. This was solved by always counterbalancing the sides each word occurred on the screen. Second, the word lengths of some word-pairs, especially those with or without nasals, were different. Hopefully this effect could be minimized by counterbalancing the sides of the visual stimuli and the frequency of all the audio stimuli.

For each trial, there consisted of a single auditory stimulus and two visual stimuli. Table 3 shows all the visual word pairs, as well as all the auditory stimuli that occur to each word pair. Each word pair occurred 6 times (2 counterbalanced sides × 3 repetitions). Thus there were 36 CVT-CVD stimuli (3 word pairs × 2 auditory stimuli × 6 times), 54 CVT-CVNT and CVD-CVND each (3 × 3 × 6), and 72 CVNT-CVND (3 × 4 × 6), making a total of 216 trials. This number is different from Beddor et al. (2013)'s 360 trials because: (a) 2 sets of quadruplets were cut; (b) an additional condition was cut

resulting one less auditory stimuli for the CVT-CVNT/CVD-CVND pairs; (c) different repetitions (8 for CVT-CVD and 4 for others in Beddor et al. 2013).

Table 3. All the visual pairs and auditory stimuli for each pair

	CVT-CVD	CVT-CVNT	CVD-CVND	CVT-DCVD
Visual pairs	bet-bed	bet-bent	bed-bend	bent-bend
	let-led	let-lent	led-lend	lent-lend
	set-said	set-sent	said-send	sent-send
Auditory stimuli	CVT	CVT	CVD	CVNT-early
				CVNT-late
	CVD	CVNT-early	CVND-early	CVND-early
		CVNT-late	CVND-late	CVND-late

2.3 Procedure

All the participants were tested with a screen-based eye-tracker (Tobii Pro TX300) with a sampling frequency at 300Hz (Beddor et al. (2013) used an Eyelink headband-mounted camera with a sampling frequency at 500Hz). They were seated in a chair at an appropriate height for the camera. Room lighting was not direct. Before the experiment started, they went through a calibration procedure until the criterion was reached for both eyes. Data from the left eye were used in this study. After the calibration, they were given instructions on the experiment and performed a practice session of ten random trials to get familiar with the procedure. The experiment took place in a single session, but the participant could take a rest whenever they wanted to between trials.

Each trial consisted the following sequence of events. First, two words appeared on the screen symmetrically on two sides for 2 seconds for the participants to get familiar with the words and their sides in this trial. Then the words disappeared and fixation cross appeared on the screen while the listeners heard “Please look at the fixate cross.” The machine let the participants pass to the next event when it captured their eye gaze at the central area around the fixation cross. Next the words reappeared on the two sides of the cross, while the listeners heard “Now look at …” At this point, they heard the target word, and they should then look at that word accordingly as the fixation cross disappeared. A sample of the visual stimuli is shown in Figure 1. All the trials were randomized throughout the session. This procedure was comparable to the one used in Beddor et al.

(2013), except that their experiment took much longer and therefore had to be split into sessions and separate visits.

Figure 1 Sample for a trial

2.4 Data analysis

The eye movements of the participants were monitored starting from the onset of the target word. The measurements were latency of initial correct fixation: how long did it take for the participants' eye gaze entered the area of interest. The latency would only be counted if it took no longer than 1000ms (the participant did not react in the trial or simply took too long rendering an outlier) and no shorter than 200ms (the participant's eye gaze might be wandering and happened to enter the area not as a response of the auditory stimuli). As a result, data from 5 trials were excluded from the analysis for each participant respectively. The data tracking the right eye were used in the data analysis.

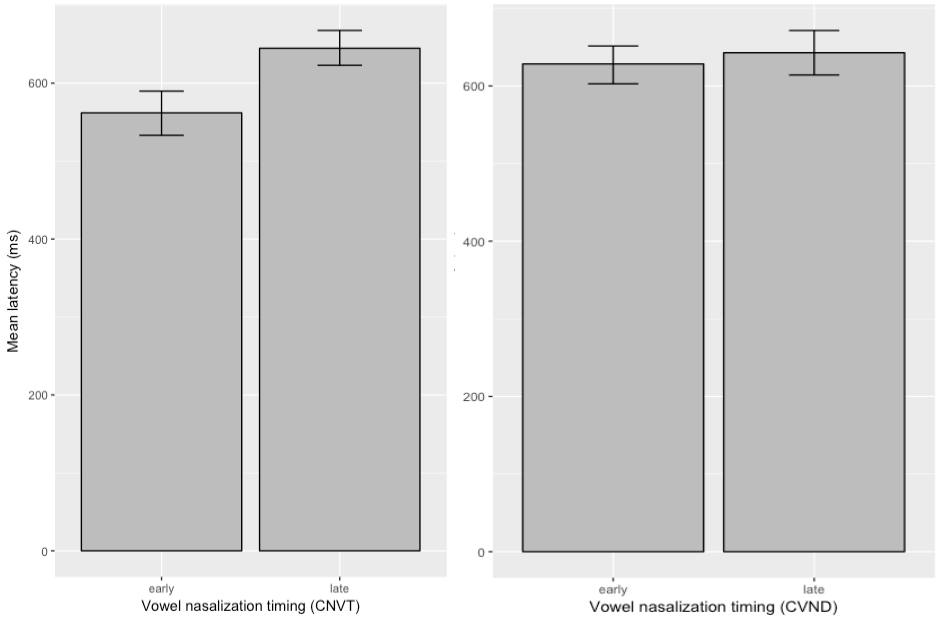
3 Results

Hypothesis 1 states that the listeners will use the coarticulatory information soon after it becomes available in the speech signal. Therefore, when facing a CVNC-CVC visual pair and listen to a CVNC target, the listeners are predicted to look at CVNC earlier when the auditory stimulus has an early nasalization onset. In other words, the fist fixation latency

for CVNC-early auditory stimulus should be smaller than that for CVNC-late in trials with CVNC-CVC visual pairs.

Figure 2 shows the mean latencies of the first correct fixation on trials where participants heard a CVNC and saw a CVC-CVNC word pair. The mean latency for the early nasalization onset tokens appears to be smaller than the late tokens, especially for the trials with voiceless codas. A linear mixed-effects model was computed on the latencies in these trials. The fixed effects were Timing of Nasalization (early, late), the Voicing of the Coda (voiced, voiceless), as well as their interaction; the random effect was the Target Word. No significant effects were found, indicating that there is no significant difference in the latencies between the trials with early and late nasalization onsets, disconfirming the first hypothesis.

Figure 2 Mean latencies of first correct fixations on trials with auditory CVNC and visual CVC-CVNC pair, grouped by voicing condition (left: voiceless coda; right: voiced coda)

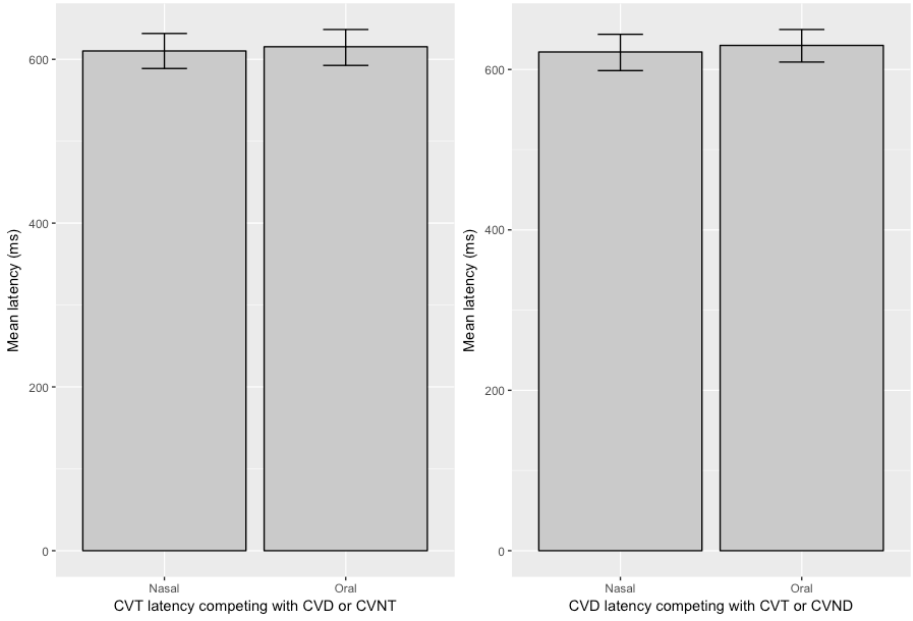


Hypothesis 2 claims a nasalized vowel is more informative than an oral vowel; therefore an oral vowel, unlike a nasalized vowel, will not help the listeners to predict whether the following consonant is nasal or not. This reasoning predicts that when listeners heard target word $CVC_{\alpha voice}$, they will use the oral vowel to predict that the word is a CVC sequence and not a CVNC sequence. Therefore, listeners will not look at the

target word faster in a trial where the visual competitor is $CVNC_{\alpha voice}$ than in a trial with the competitor $CVC_{-\alpha voice}$. For example, when listeners hear *bet*, the latency to look at bet should be not be smaller if the competitor is *bent* than if it is *bed*. In fact, we would expect there should not be a significant difference between those latencies.

Figure 3 shows the mean latencies in trials where the target is a CVC word (e.g. *bent*), while the competitor is the other CVC word (*bed*) or the corresponding CVNC word (*bent*). The mean latencies between the oral competitor and the nasal competitor do not seem to be different. This was confirmed by a mixed-effects model, where the fixed effect were Competitor Nasal (nasal, oral) and Target Voicing (voiced, voiceless) and the random effect is the Target Word The model detected no significant effect. The lack of effect confirms hypothesis 2 that the oral vowel would not be a helpful indicator for the nasality of the following sound. However, since it is difficult to detect effects with a small pool of data from only two participants, the lack of effect here should be taken with caution.

Figure 3 Mean latencies of $CVC_{\alpha voice}$ target against $CVNC_{\alpha voice}$ or $CVC_{-\alpha voice}$ competitor (left: target with voiceless coda; right: target with voiced coda)



4 Discussion

This study investigated the dynamic correspondence between perception and the gestural information encoded in the acoustic signal using eye tracking data that monitor the time course of perception. Because eye tracking captures perceptual evens with high temporal resolution, this paradigm could show when listeners attend to certain sub-phonemic information in the speech stream.

This replication focused on the first two hypotheses in Beddor et al. (2013), using first fixation latency as a measurement to test whether listeners benefit from nasalized (hypothesis 1, does benefit) and oral vowel (hypothesis 2, does not benefit) in predicting the next sound. With very limited number of participants (only two), no effect was shown after including the target word as a random effect for both hypotheses. For hypothesis 1, the lack of effect contradicted Beddor et al. (2013)'s results. However, while the effect was not shown, the data did show the predicted direction, especially for the voiceless targets, where early nasalized onset had a shorter latency for the perception of the nasal. Given more participants, it is hopeful that there will be an effect to confirm the claim in the paper. For hypothesis 2, the lack of effect was expected. However it could be just a result of the limited amount of data. Moreover, although the authors expected no effect in hypothesis 2 in their study, they in fact found that an oral vowel would shorten the latency of a following oral consonant when the coda was voiceless. However, this effect was found to be specific to the word *watt*, which was not included in this replication. The authors stipulated that the word-specific effect might have something to do with the lip-rounding of the onset [w].

An additional potential problem I found in replicating this experiment was that the cross-splicing might not give full control over the degree of nasalization in the experiment tokens. It is possible that the speaker might have different degree or timing of nasalization in different productions, or across voicing contexts. Furthermore, The degree of nasalization is notoriously unreliable to measure acoustically. Therefore, cross-splicing a chunk of nasalized vowels from a second token might not guarantee the consistence in the kind of information that the nasalized vowel could provide, especially across voicing distinctions. This might be why both the authors and my replication have found some inconsistent patterns across the voicing conditions.

This study shows the potential of the eye-tracking paradigm in understanding moment-to-moment perception. It could be used to explore long-distance coarticulation and other phonetic variations, or to capture social perception of meaningful sub-phonemic features in real-time interactions, especially with more complicated visual stimuli, such as interactants in action.

5 Reference

Beddor, P. S. (2009). A coarticulatory path to sound change. *Language*, 85, 785-821.

Beddor, P. S., Mcgowan, K. B., Boland, J. E., Coetzee, A. W., & Brasher, A. (2013). The time course of perception of coarticulation. *The Journal of the Acoustical Society of America*, *133*(4), 2350–2366.

Flagg, E. J., Oram Cardy, J. E., and Roberts, T. P. L. (2006). MEG detects neural consequences of anomalous nasalization in vowel-consonant pairs. *Neurosci. Lett*., 397, 263–268.

Lahiri, A., & Marslen-Wilson, W. (1991). The mental representation of lexical form: A phonological approach to the recognition lexicon. *Cognition,* 38, 245–294.

Malécot, A. (1960). Vowel nasalization as a distinctive feature in American English. *Language*, 36, 222-229.

Ohala, J. J., and Ohala, M. (1995). Speech perception and lexical represen- tation: The role of vowel nasalization in Hindi and English. in *Phonology and Phonetic Evidence: Papers in Laboratory Phonology IV*, edited by B. Connell and A. Arvaniti (Cambridge University Press, Cambridge, UK), pp. 41–60.

Warren, P., and Marslen-Wilson, W. (1987). Continuous uptake of acoustic cues in spoken word recognition. *Percept. Psychophys*., 41, 262–275.