Audio-visual factors in stop debuccalization in consonant sequences

Keith Johnson

Department of Linguistics

UC Berkeley

Abstract

This study tested the hypothesis that some language sound patterns originate in factors relating to the perception of speech as an audio-visual signal. Specifically, this was a study of stop debuccalization (a process in which the place of articulation of a stop changes from oral to glottal). Phonetically trained listeners were asked to identify stop consonants in medial clusters as “p”, “t”, “k” or “ ʔ” in audio-visual recordings of nonwords that had the test segments [p], [t], [k], and [ʔ] in syllable coda followed by [m], [n], [l] or [ŋ] in the onset of the next syllable. The relative timing of the audio and video streams were manipulated so that in 2/5 of the stimuli the audio was advanced relative to the video and in 2/5 of the stimuli the audio was delayed. This manipulation mainly affected the perception of [p] in homorganic [p.m] clusters such that when the audio was advanced listeners reported more “k” and “ʔ” than when the audio offset was unchanged or when the audio was delayed relative to the video. The implications of this type of effect in audio-visual speech perception on linguistic sound patterns is discussed.

1. Introduction

In debuccalization, the place of articulation of a consonant changes from oral to glottal. Stops debuccalize to the glottal stop: [p],[t],[k] > [ʔ], and fricatives debuccalize to a glottal fricative ([f], [θ],[s],[z],[x] > [h]). The two main environments where debuccalization has been found are in consonant sequences and in word final position (O'Brien, 2012).

This report concerns only one kind of debuccalization - of stops in consonant sequences. The main idea tested in this report is that debuccalization may be influenced by variability in the temporal alignment of audio and visual phonetic cues. This was tested by shifting the sound track of movies so that the audio and video signals are out of sync with each other. Speech perception is known to be affected by both auditory and visual information (e.g., McGurk and MacDonald, 1976; Summerfield, 1979). Early research clearly demonstrated that seeing a speaker’s face in conversation can significantly improve the accuracy of speech perception (Sumby and Pollack, 1954; Erber, 1975). And research on the integration of audio and visual signals in speech perception indicates both that listeners/viewers tolerate a fairly wide discrepancy of AV synchronization (Nygaard, 1993; Munhall, et al., 1996) and that attention to one channel or the other can cognitively alter the relative timing of the two (Spence & Parise, 2010).

This last point is important because if we are to posit that AV temporal misalignment could have any impact on sound change in the real world (outside the lab), it is necessary to show that AV temporal alignment variation may occur in nature. Audio-visual temporal misalignment certainly occurs in the physical world due to the different speeds of sound and light (table 1). As a result of this, listeners have experience with some degree of AV temporal misalignment for at least some kinds of events in the world in which the audio signal lags slightly behind the visual signal because the speed of light is many orders of magnitude faster than the speed of sound.

Table 1. Temporal asynchrony in audio and visual arrival times.

1 meter: (1/speed of sound) - (1/speed of light) = 0.0029 sec

10 meters: (10/speed of sound) - (10/speed of light) = 0.0294 sec

100 meters: (100/speed of sound) - (100/speed of light) = 0.2938 sec

Speed of sound = 340.29 m/s; speed of light = 299,792,458 m/s.

It is also important to note that research has found that the cognitive integration of audio and visual information is malleable (Fujisaki et al., 2004) and in particular that attention to a channel can speed the processing in that channel (Sternberg & Knoll, 1973; Spence et al., 2001; Weiß et al. 2013).

Debuccalization of stops in consonant sequences has been observed in a number of languages.
For example, in many varieties of English the phrase 'get ready' may be pronounced with a glottal stop instead of the "t" [gɛʔ ɹɛdi]. O'Brien (2012) presented a comprehensive literature review detailing nearly 20 languages that have synchronic or diachronic stop debuccalization processes. He found that stop debuccalization is more common with dorsal consonants than with coronal or labial consonants. Of nineteen stop debuccalization cases that he found, 16 of them involved velar consonants [k] or ejective [k'], while only 11/19 cases were with [t] or [t'], and only 8/18 involved labials. There was also an implicational relationship. No language had only debuccalization of labial, while eight debuccalized dorsals only, labials and coronals unaffected, and in three only coronal consonants debuccalized. The tendency, then, is for back consonants to debuccalize before front ones do. There is no good acoustic reason for this - dorsal stops have distinctive formant transitions (the "velar pinch" of F2 and F3) not found the glottal stop, and dorsal stops tend to be released with a double stop release burst that is almost never seen in glottal stops. The phonetic affinity between velar stops and glottal stops is that in both of these the place of articulation of the stop closure is not as visually apparent as it is in labials or coronals (ref).

It has also been observed that stops in homorganic clusters are more likely to debuccalize than are stops in heterorganic clusters. Consider for example the process of glottal stop formation in Hayu (Michailovsky & Mazaodon, 1974). The final /p/ in /dip/ becomes a glottal stop when the suffix /-me/ is added: [diʔme]. All of the voiceless stops evidently do this: /put-no/ -> [puʔno], /pʊk-ŋo/ -> [pʊʔŋo]. Once again a visual phonetic mechanism may be involved. In homorganic clusters, the timing of the movement of articulators, particularly the lips but also movement of the jaw for coronal stops, may be visually indistinct. The listener/viewer may see motion but not be able to attribute the phonetic place of articulation information to a particular temporal portion of the speech signal. Whether this leads to perceptual "loss" of place of articulation in a homorganic stop more often than in a heterorganic stop may depend on the particular phonotactic probabilities of the language, but at least in homorganic clusters the adjacent homorganic segment provides for the listener a docking site for the audio or visual perceptual cues for place of articulation so that if acoustic or visual cues for the coda stop are indistinct the stop can be perceptually debuccalized without leaving inexplicible place cues floating unaccounted for.

The hypothesis tested in this report is that this type of phonological phenomenon is impacted by temporal misalignment in the perceptual integration of audio and visual cues. The acoustic signal provides the main cues for the voicelessness and obstruction of [p], for example, while the visual signal provides key information for the [labial] place of articulation in [p]. If these two streams (manner of articulation and place of articulation) are misaligned perceptually, debuccalization may by hypothesis result. The results will provide some evidence for this and point toward further tests that may be done to explore this idea.

2 Methods

2.1 Materials. A phonetically trained male speaker of American English (a native of Santa Barbara, CA) read a list of two syllable nonce words modeled phonotactically on words in Hayu (table 2). The talker was not aware of the purpose of the experiment. He reviewed the word list prior to the recording session in order to become familiar with the phonetic make up of the words. The test words were presented in IPA phonetic transcription on a teleprompter so that the talker looked directly into the camera while reading. The order was randomized and the list was presented twice in different random orders. As the talker read the items a video camera (Canon XF100) recorded an MXF file containing uncompressed audio, 16 bit samples at 48kHz sampling rate, and MPEG-1/2 HDV video with 1920x1080 pixels at 59.94006 frames per second (16.68 ms per frame).

The word list (Table 2) was composed to 32 pronouncable nonwords formed by inserting [p], [t], [k] and [ʔ] in eight different frames. The perception of these stop consonants is the focus of the experiment reported here. The vowel before the test coda consonants was [ɑ] in two of the frames and [i, u, o, ɪ, æ, e] in the remaining six frames. The consonant following the test consonants were [m, n, ŋ, l]. There were eight homorganic intervocalic sequences, as indicated in table 2, and twenty-four heterorganic sequences.

Table 2. Phonetic sequences tested in the experiment. Homorganic sequences are underlined.

dip.me	dup.mo	pep.ni	pɑp.nu	tɪp.ŋæ	tæp.ŋi	sɑp.lo	sop.lɑ
dit.me	dut.mo	pet.ni	pɑt.nu	tɪt.ŋæ	tæt.ŋi	sɑt.lo	sot.lɑ
dik.me	duk.mo	pek.ni	pɑk.nu	tɪk.ŋæ	tæk.ŋi	sɑk.lo	sok.lɑ
diʔ.me	duʔ.mo	peʔ.ni	pɑʔ.nu	tɪʔ.ŋæ	tæʔ.ŋi	sɑʔ.lo	soʔ.lɑ

2.2 Pretest. The audio track of the recording session was extracted (all AV processing was done on a Linux platform using the software package FFMPEG; Niedermayer & Biurrun, 2015) and a Praat (Boersma & Weenik, 2015) TextGrid file was produced using the Penn Forced Aligner (Yuan & Liberman, 2008). The alignments were inspected and hand-corrected as necessary. Audio waveforms corresponding to each test word were then extracted into separate sound files and the amplitude in the files was normalized so that each clip had the same peak amplitude level (3 dB below the maximum representable in 16 bit samples). This was done using a custom script that read the TextGrid file (Sprouse, 2015) and used the software package SOX (Bagwell & Norskog, 2015) to extract and modify the audio. These audio files were presented to a panel of ten trained phoneticians in a glottal stop detection task. The instruction was to listen for glottal stop in the coda of the first syllable of each two syllable stimulus.

Table 3. Responses in the audio-only glottal stop detection pre-test.

	/p/	/t/	/k/	/ʔ/
'ʔ'	25	33	13	96
'~ ʔ'	135	127	147	64

These data are from the most accurately perceived instance of each nonce word listed in table 2. The target consonants are shown in the columns and the response options are shown in the rows. The percent correct glottal stop detections (hits) was 96/(96+64) = 0.6 and the percent of false alarms to [t] tokens (the most confusable consonant) was 33/(33+127) = 0.21. Thus, d' = Z(0.6) - Z(0.21) = 1.073.

The overall d' value for glottal stop detection (using the proportion of 'ʔ' responses to [t] tokens as the false alarm rate) was 0.54. Taking the most accurately perceived example of each test word (recall that each word was recorded twice) the response matrix shown in Table 3 results. This gives a d' value for glottal stop detection of 1.073. The most accurately perceived token of each nonce word in table 2 was used in the main audio-visual experiment. The pretest shows that glottal stops were present and detectable above chance in the stimuli that were used in the main experiment. The results also show that there may have been some "glottal reinforcement" of /t/ - the tendency in American English to pronounce coda /t/ with a nearly simultaneous glottal stop [ʔt] -

because glottal stops were heard more often in the /t/ tokens than in the /p/ or /k/ tokens. This was not a large effect, the intended /t/ tokens were much less likely to be identified with 'ʔ' than was the intended glottal stop, and the main manipulation in this experiment is to alter the time alignment of the audio and video streams, so even this small bit of glottal reinforcement in /t/ won't affect the interpretation of the AV time-alignment manipulation, particularly in changes in the perception of /p/ and /k/.

2.3 Audio/Visual Stimuli. Video clips of the tokens selected in the pretest were extracted from the AV recording. Each clip began 300 ms prior to the tagged word onset in the TextGrid file and ended 300ms after the word offset. The audio clips used in the pretest were added to the extracted video clips at five different audio offset values (-0.2, -0.1, 0, 0.1, 0.2 seconds relative to the video signal). The discrepancy between the audio and video signals ranged from 6 to 12 frames in the video (100 ms/16.7 ms/frame = 6 frames). Thus, with an audio offset value of -0.2 sec, the audio started 12 frames earlier in the movie than it did in the original recording, and with an audio offset value of 0.2 sec the audio signal started 12 frames later than it did in the original recording. This resulted in 160 audio/visual stimuli -- 32 test items with five degrees of audio/video asynchrony.

2.4 Listeners. Eighteen participants were tested. They were students at UC Berkeley who had taken a course on phonetics and learned to transcribe glottal stop in phonetics exercises. The listeners were all native speakers of American English, and reported no speech or hearing problems. The research study reported here was approved by the Committee for the Protection of Human Subjects at UC Berkeley (IRB Protocol #2013-07-5474) . Informed consent was obtained orally as approved by the IRB.

2.5 Procedure. Listeners were seated in a sound attenuating booth, and given headphones to wear (AKG K240). They were seated in front of a 15” LCD computer monitor which was about 1/2 meter in front of them and the room light in the booth was moderately dim. Listeners were given a keyboard and asked to press one of four keys in response to movies of a person saying nonce words. The response options were the 'p', 't', 'k', and '/' keys on a standard computer keyboard.

Specifically the instructions were:

> In this experiment you will hear 320 two-syllable words.
> Each word has the structure CVCCV.
> Your task is to identify the second C as /p/, /t/, /k/, or /?/ (glottal stop).
>
> That is, was it CVpCV, CVtCV, CVkCV or CV?CV?
>
> Press the letter 'p', 't','k', or '?' to enter your answer
> Don't use the <shift> key for '?' - the key '/' is the expected answer.

Ready to start? press any key

The experiment session started with 32 practice trials drawn randomly from the 160 test stimuli. Listeners were given feedback during the practice session. They were simply told if their response was correct or incorrect, and not what the correct answer should have been. Then after a short break the main session began. Responses were collected without feedback, and without pressure to answer quickly. Each of the 160 AV stimuli was presented twice in a randomized list of 320 trials.

3 Results

Listeners' ability to detect glottal stop in these stimuli was much worse than that of the more expert pre-test listeners. Calculating d' as we did for the pre-test we derive a value of only 0.42. As Table 4 also shows, listeners were biased against responding with 'ʔ' especially for [p] tokens. Glottal stop tokens were twice as likely to be identified as glottal stop than were any of the other tokens, but despite this the glottal stop target consonants were more likely to be identified as 't' than they were to be identified as 'ʔ'. This reflects the tendency for coda /t/ to be glottalized in American English.

Table 4. Confusion matrix for data pooled and converted to percentages across all conditions.

	'p'	't'	'k'	'ʔ'
[p]	80%	5.1%	7.7%	7.2%
[t]	5.9%	58.1%	18.1%	17.9%
[k]	3.4%	6.3%	77.9%	12.3%
[ʔ]	8.3%	41.6%	15.4%	34.7%

The confusion matrices in Table 5 show the main results of this experiment. There are twenty confusion matrices in table 5, one for each combination of audio/video asynchrony and following consonant context. The confusion data are also illustrated graphically in Figure 1. Statistical analysis of these data (Table 6) indicates that each of these factors (target consonant, context consonant, and audio/video asynchrony) helps explain the perceptual data. The data were fit with a set of GLM logistic regression models, one for each of the four response options. Thus, one model coded responses as either "p" or not-"p" and predicted this binary outcome with the three predictor variables and their interactions. The predictor variables are target consonant ([p], [t], [k]), context consonant ([m], [n], [l], or [ŋ]) and amount of audio-offset (-.2, -.1, 0, .1, .2 seconds of audio/video temporal asynchrony, entered as a continuous predictor variable). Responses to glottal stop stimuli were not included in the models because the aim of the experiment was to determine when oral consonants would be debuccalized, so we didn't have any research aims involving the perception of glottal stops. The glottal stop tokens were presented in the experiment as fillers to encourage the use of the glottal stop response alternative.

The results are largely unchanged but statistically weaker if the audio-offset effect is entered as an ordered categorical factor rather than as a continuous variable. As shown in Table 6, target consonant, context consonant, and their interaction had a significant impact on the use of all of the response alternatives. This can be seen in the confusion matrix as well. For example, target identification was most accurate in the context of [_l] at a little over 60% correct, while [t] identification in the context of [ŋ] was not very accurate at all at only a little over 10% correct.

Table 5. Raw confusion matrices for each combination of audio offset (-0.2, -0.1, 0, 0.1, 0.2 sec audio-visual asynchrony), and following consonant. Data are the number of responses in each combination of conditions.

		[_l]				[_m]				[_n]				[_ŋ]			
		“p”	“t”	“k”	“?”	“p”	“t”	“k”	“?”	“p”	“t”	“k”	“?”	“p”	“t”	“k”	“?”
-0.2	[p]	**60**	2	0	9	**17**	13	25	16	**69**	2	1	0	**62**	4	5	1
	[t]	2	**59**	2	9	4	**45**	8	15	7	**48**	5	12	3	**9**	42	16
	[k]	2	1	**66**	2	0	6	**53**	12	1	5	**59**	5	1	5	**53**	11
-0.1	[p]	**64**	0	0	8	**20**	10	26	13	**68**	0	1	3	**64**	3	1	2
	[t]	0	**65**	1	6	3	**44**	7	15	5	**51**	3	12	0	**10**	43	16
	[k]	0	1	**64**	7	3	2	**51**	13	3	5	**60**	4	1	5	**53**	11
0	[p]	**65**	1	0	4	**39**	9	11	12	**66**	2	2	0	**70**	0	2	0
	[t]	1	**60**	2	9	7	**47**	2	16	5	**52**	4	11	0	**8**	45	15
	[k]	1	0	**63**	6	6	4	**53**	8	1	10	**51**	7	0	6	**60**	5
0.1	[p]	**66**	1	0	4	**40**	12	11	7	**68**	2	2	0	**66**	2	2	1
	[t]	0	**63**	0	6	17	**34**	4	15	4	**50**	2	16	2	**13**	42	12
	[k]	2	0	**64**	3	7	5	**41**	14	6	8	**50**	7	2	6	**52**	11
0.2	[p]	**64**	0	1	5	**39**	8	14	11	**67**	0	3	2	**64**	2	2	4
	[t]	0	**62**	0	10	16	**39**	2	13	4	**50**	2	13	3	**11**	39	16
	[k]	0	1	**62**	8	7	6	**37**	19	5	8	**49**	8	0	5	**52**	12

Table 6. Analysis of Deviance Tables. Results of GLM analysis of responses fitting separate binomial logistic regression models to predict the four different response alternatives that listeners used to identify the test segments. The factors are “targetC” - the targe consonant [p], [t], or [k]; "adjC" the context consonant [m], [n], [l], or [ŋ]; and "av_offset" – the AV temporal asynchrony (in seconds).

```
-------------------------------------------------------------------------------
Response: labial “p”
                       Df Deviance Resid. Df Resid. Dev  Pr(>Chi)    
NULL                                    4235     5172.2              
targetC                 2  2700.47      4233     2471.7 < 2.2e-16 ***
adjC                    3   141.47      4230     2330.3 < 2.2e-16 ***
av_offset               1    26.31      4229     2303.9 2.911e-07 ***
targetC:adjC            6   293.10      4223     2010.8 < 2.2e-16 ***
targetC:av_offset       2     0.50      4221     2010.3 0.7778645    
adjC:av_offset          3    19.37      4218     1991.0 0.0002294 ***
targetC:adjC:av_offset  6    11.27      4212     1979.7 0.0803253 .  
---------------------------------------------------------------------------

Response: dorsal “k”
                       Df Deviance Resid. Df Resid. Dev  Pr(>Chi)    
NULL                                    4235     5452.8              
targetC                 2  1868.19      4233     3584.6 < 2.2e-16 ***
adjC                    3   175.17      4230     3409.4 < 2.2e-16 ***
av_offset               1    19.08      4229     3390.3 1.251e-05 ***
targetC:adjC            6   578.19      4223     2812.2 < 2.2e-16 ***
targetC:av_offset       2     0.58      4221     2811.6   0.74803    
adjC:av_offset          3     9.15      4218     2802.4   0.02734 *  
targetC:adjC:av_offset  6    10.36      4212     2792.1   0.11033    
---------------------------------------------------------------------------

Response: coronal “t”
                       Df Deviance Resid. Df Resid. Dev Pr(>Chi)    
NULL                                    4235     4587.3             
targetC                 2  1429.78      4233     3157.6   <2e-16 ***
adjC                    3   281.40      4230     2876.2   <2e-16 ***
av_offset               1     0.03      4229     2876.1   0.8638    
targetC:adjC            6   271.38      4223     2604.7   <2e-16 ***
targetC:av_offset       2     3.55      4221     2601.2   0.1696    
adjC:av_offset          3     1.61      4218     2599.6   0.6572    
targetC:adjC:av_offset  6     3.85      4212     2595.7   0.6973    
---------------------------------------------------------------------------

Response: debuccalized “/”
                       Df Deviance Resid. Df Resid. Dev  Pr(>Chi)    
NULL                                    4235     3186.2              
targetC                 2   76.911      4233     3109.2 < 2.2e-16 ***
adjC                    3   59.123      4230     3050.1 9.048e-13 ***
av_offset               1    0.272      4229     3049.8   0.60200    
targetC:adjC            6   66.642      4223     2983.2 1.992e-12 ***
targetC:av_offset       2    6.782      4221     2976.4   0.03367 *  
adjC:av_offset          3    0.930      4218     2975.5   0.81815    
targetC:adjC:av_offset  6    5.135      4212     2970.4   0.52657    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Both the “p” and “k” logistic models showed reliable main effects for audio-offset, and reliable two-way interactions between adjacent consonant, and audio-offset, though this interaction was only marginally significant in the “k” analysis ($\chi^2 = 9.15$, df = 3, p=0.027). The interactions are quite apparent in Figure 1, which shows the confusion matrix data (Table 5) in graphic form. Notice that in Figure 1A the proportion of "p" responses given to [pm] sequences is dramatically lower when the audio leads the video by 100 or 200 milliseconds. Concomitantly, the number of "k" responses to these stimuli (Figure 1C) is greater for the audio-lead stimuli. The presentation in Figure 1 is designed to reflect the statistical analysis, where we had one model for each response alternative. The plot symbols in each panel stand for the stimulus target consonant so Figure 1A shows that most "p" responses were given to [p] stimuli. Figure 1C shows that [t] in [tŋ] sequences and [p] in [pm] sequences tended to be identified as "k" more than [t] or [p] in other consonant environments.

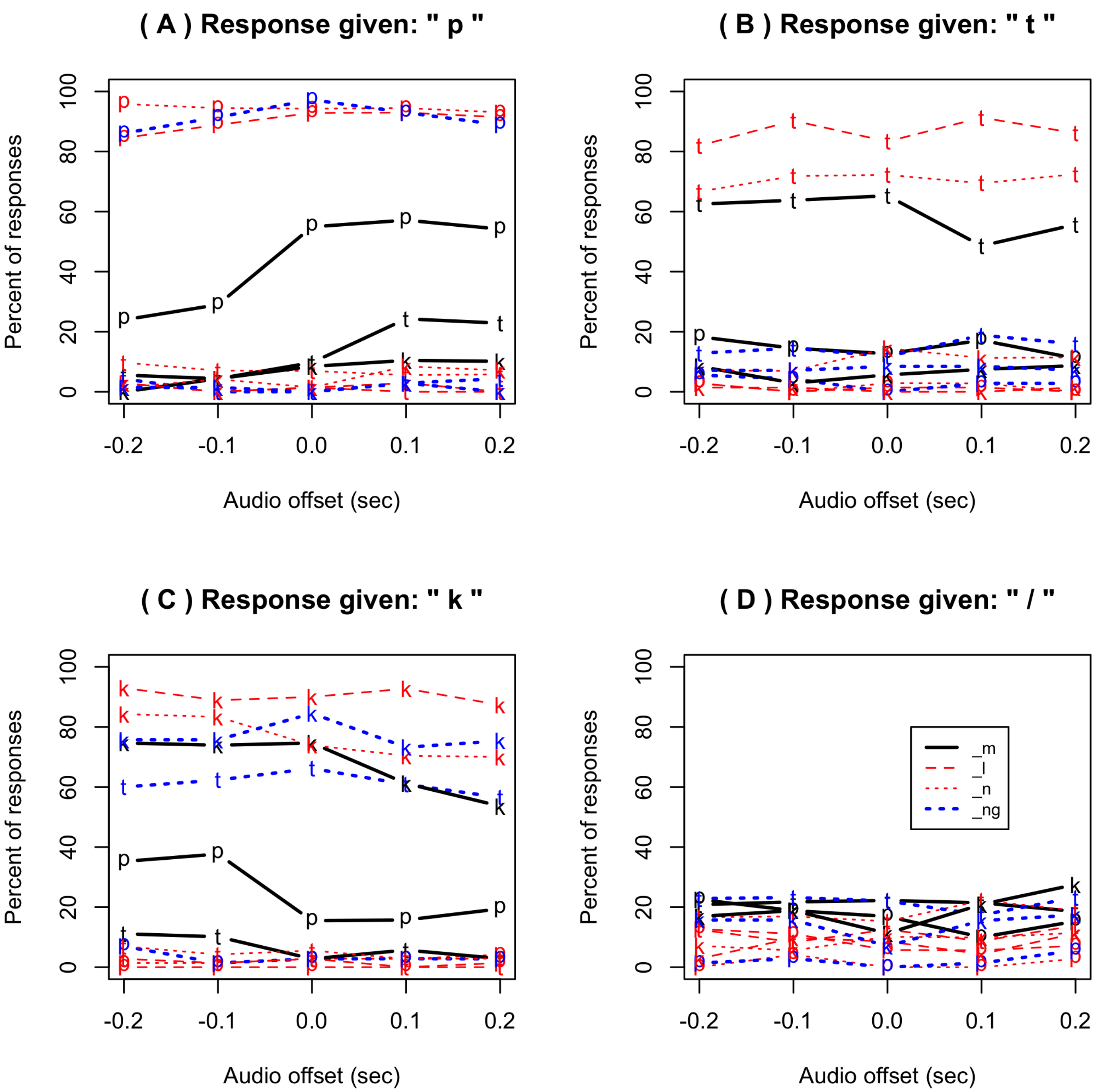

Figure 1. The confusion matrix data (Table 5) are presented here in graphic form. Percent responses are plotted in a separate graph for each response option. The target consonant identity (the consonant produced by the speaker) are indicated with the plot symbols used to plot each function. Data are plotted in separate functions according to identity of the following consonant ([m], [l], [n] or [ŋ]), and as a function of the audio-offset value (negative = audio leads the video stream, and positive = audio lags the video stream). For example, panel (C) shows that subjects responded "k" to the [kŋ] sequence about 80% of the time and responded "k" to the [tŋ] sequence about 60% of the time, regardless of audio offset.

Figure 2 shows the interaction of target consonant and audio offset in the likelihood of identifying the target consonant as a glottal stop - i.e. in the perceptual debuccalization rate. This interaction was marginally significant (χ^2 = 6.78, df = 2, p=0.034). Figure 2 indicates that the debuccalization rate (averaged over consonant cluster type) was highest for [t] and lowest for [p], and that debuccalization for [k] and possibly [p] were influenced by AV temporal asynchrony while debuccalization for [t] was not. In a logistic model predicting the proportion of debuccalization responses from target consonant, adjacent consonant, and audio offset, we have a slope for audio offset for [t] targets of 0.056 (not reliably different from zero, p = 0.91), a slope for audio offset for [k] targets of 1.203 (p = 0.039) and a slope of -1.248 for [p] targets (p = 0.097). This analysis is at the edge of what can be reliably determined from this data set, but points to an interpretation of these data which suggests that the rate of debuccalization was higher for [k] tokens when the audio lagged the video, while the debuccalization rate was somewhat higher for [p] when the audio led the video. The higher debuccalization rate for [t] is likely tied to "glottal reinforcement" of /t/ in American English and was not affected by audio-visual temporal asynchrony.

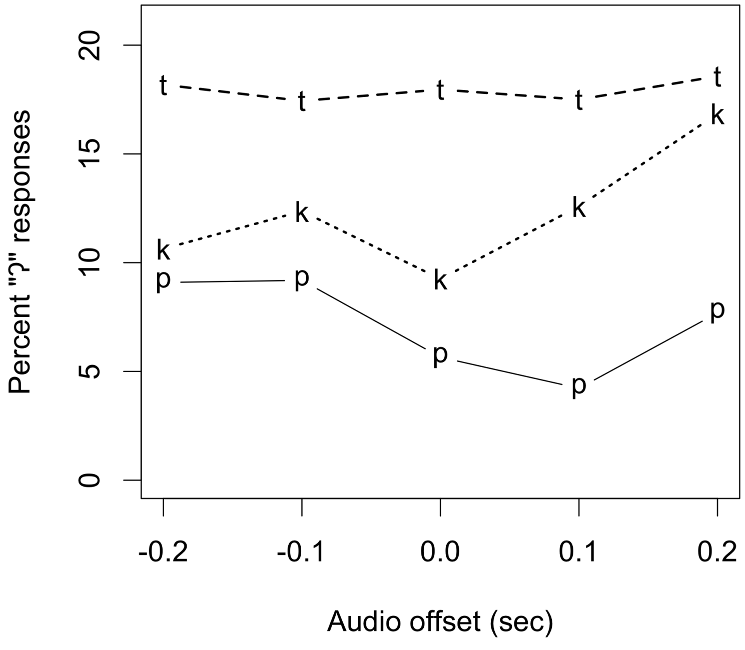

Figure 2. Percent debuccalization responses. Responses for [p], [t] and [k] target consonants as a function of the degree of AV temporal misalignment (audio offset).

4 Discussion

This study found that the perception of coda consonants in word medial clusters depends on the place of articulation of the neighboring consonant in the cluster. Accuracy of labial stop perception in A/V synchronous stimuli (audio-offset = 0) drops from 93% in [pl] sequences to 55% in [pm] sequences, while accuracy of coronal stop perception shows an inverse effect with accuracy less than 12% in [tŋ] sequences and 83% accuracy in [tl] sequences. Cross-linguistic research on this topic needs to be conducted because it isn't at this point clear what proportion of these perception results is due to the availability of acoustic cues in the particular stimuli and what portion is due to listener expectations driven by linguistic/structural characteristics that are unique to English. As the type counts in Table 7 suggest, there is reason to wonder if listeners might have been guided by English phonotactics. Yet, this is clearly not the only explanation for the interactions of target consonant and cluster type in the perceptual data. The table shows that that [pm] is a relatively rare sequence while [tl] is relatively common. This is compatible with the idea that listeners' responses were based at least in part on their language-guided expectations that [pl] and [tl] sequences are relatively common while [pm] and [tŋ] sequences are unusual. Complicating this story is that responses to [k] showed no great sensitivity to cluster type despite the fact that [kŋ] sequences are not found at all in the phonemic structure of English.

Table 7. Word type-counts for consonant combinations.

	m	n	l	ŋ
p	6	13	214	0
t	27	60	165	0
k	23	34	165	0

Number of lexical items in the CMU pronouncing dictionary (Rudnicky, 2014) that have medial clusters of a stop (rows in the table) followed by resonant (columns in the table) as tested in this study. The homorganic clusters are underlined.

4.1 Perceptual factors that favor debuccalization. Labial stops in homorganic sequences are unstable -- likely to be misperceived. And when the audio signal appears to precede the video signal [p] may be perceived as "ʔ" in as many as 15% of the words that it occurs in. It should also be noted that an even greater number of "k" percepts occur when [p] is in an asynchronous homorganic sequence.

In general, the impact of audio-visual asynchrony was limited to sequences that involve labials. When the audio signal preceded the video signal, so that the audio stop closure occurred before the video lip closure, then listeners tended to 'loose' the [p] – reporting instead that they heard a 'k', 't', or '?'. But this only happened when there was an [m] following the [p]. The following [m] seems to have given the listener a way to parse the visual labial gesture, giving it an perceptual anchor in the signal which freed the stop percept from the labiality percept. On the other hand, if the audio signal lagged behind the visual signal (as it sometimes does in nature), there were small increases in the number of [t] → “p” misperceptions and a few more [k] → “ʔ” misperceptions both only in the [_m] environment. In the case of [t] → “p”, the audio lag puts the [t] acoustic closure closer in time to the visual lip closure of [m]. It isn't clear why audio lag would cause [k] → “ʔ” misperceptions.

One thing that *is* clear, though, is that these results point to the granularity problem discussed by Pierrehumbert (2001). We see here some small seeds of debuccalization in a perceptual effect involving audio-visual integration. But the effect is limited to a very specific phonetic context which is much more circumscribed than the phonological consequences of sound change. The results of this experiment suggest that disruption of audio-visual perceptual integration plays a role in initiating a debuccalization sound change. But the constrained set of environments and experimental conditions that result in debuccalized percepts also suggests that this phonetic “cause” of the sound change provides only a small push in a direction that must be ultimately completed by other processes.

4.2 Naive listeners. The listeners in this experiment were phonetics students who had been taught to transcribe glottal stop as an ordinary consonant. One question that naturally arises is whether this type of data could have any realistic implications for actual sound change scenarios. How should we imagine that the factors governing the perceptual emergence of debuccalized stops could come to play in the phonologies of speakers who don't have the concept of [ʔ] as a possible segment in phonology? One answer is that we have identified a condition for the loss of labiality but not a positive condition for the emergence of [ʔ]. When the labiality of [p] is displaced through audio-visual asynchrony, [k] was more likely than [ʔ] to emerge in perception. So, in a way, the present results don't have any clear implications for listeners who have no prior conception of [ʔ]. Factors such as adjacent context and temporal asynchrony that were seen to modulate the rate of [ʔ] perception in this experiment may only be operative for speakers who know [ʔ] as a possible consonant. This implies that language contact or some other source of [ʔ] in language may be necessary before the debuccalization factors that have been identified here come into play. Of course, we could reasonably expect that children learning language would be less constrained by habit or expectation than are adults. At any rate, despite the fact that the experiment was done with phonetics students, it isn't hard to imagine realistic language change scenarios where their performance in this study is relevant.

Acknowledgements. I am very grateful for the support I have been given in conducting this research study.

Funding for this research was provided by the National Science Foundation (grant number BCS 1147583). Students Annelisse Rehn, Jackson Harris, and Sarah Bakst ran subjects. Jevon Heath served as the speaker of the stimuli, and several lab members in the UC Berkeley Phonology lab provided expert judgements about the stimuli.

References

Bagwell C, Norskog L. Sox: The Swiss army knife of sound processing programs Version 14.4.2. [Computer program]. 2015. Available: http://sox.sourceforge.net/

Boersma P, Weenik D. Praat: Doing phonetics by computer Version 5.4.11. [Computer program]. 2015. Available: http://www.praat.org/

Erber NP. Auditory-visual perception of speech. *Journal of Speech & Hearing Disorders. 1975;* **40**(4): 481-492.

Fujisaki W, Shimojo S, Kashino M, Nishida S. Recalibration of audiovisual simultaneity. *Nat Neurosci. 2004; 7*: 773–8 doi:10.1038/nn1268

McGurk H, MacDonald H. Hearing lips and seeing voices. *Nature. 1976;* **264:** 746-748.

Michailovsky B. Mazaudon M. Notes on the Hayu language. *Kailash: A Journal of Himalayan Studies.* 1973; **1(2):** 135-152.

Munhall, KG, Gribble P, Sacco L, Ward M. Temporal constraints on the McGurk effect. *Perception & Psychophysics. 1996;* **58 (3):** 351-362.

Niedermayer M, Biurrun D. FFMPEG [Computer program]. Version 2.7.1 2015. Available: http://www.ffmpeg.org/

Nygaard LC. Phonetic coherence in duplex perception: Effects of acoustic differences and lexical status. J. Experimental Psychology: Human Perception and Performance. 1993; 19(2): 268-286.

O'Brien JP. *An Experimental Approach to Debuccalization and Supplementary Gestures.* PhD Dissertation, UC Santa Cruz: 2012. Available: https://escholarship.org/uc/item/1cm694ff

Pierrehumbert, J. (2001) Why phonological constraints are so coarse-grained. *Language and Cognitive Processes 16* (5/6), 691-698.

Rudnicky A. CMU Pronouncing Dictionary [Computer resource]. Version 0.7b. 2015. Available: http://svn.code.sf.net/p/cmusphinx/code/trunk/cmudict/

Spence C. Shore DI, Klein RM. Miltisensory prior entry. *J. Exp. Psychol. General.* 2001; **130**(4): 799-832.

Spence C., Parise C. Prior-entry: A review. *Consciousness and Cognition. 2010;* 19: 364–379.

Sprouse R. Audiolabel [Computer library]. 2015. Available: https://github.com/rsprouse/audiolabel/

Sternberg S, Knoll RL. The perception of temporal order: Fundamental issues and a general model. In S. Kornblum editor. Attention and Performance IV. New York: Academic Press; 1973. pp. 629-685.

Sumby WH, Pollack I. Visual contribution to speech intelligibility in noise. *J. Acoust. Soc. Am. 1954;* **26:** 212-215.

Summerfield Q. Use of visual information in phonetic perception. *Phonetica.* 1979; **36**(4--5):314--331.

Weiß K, Hilkenmeier F, Scharlau I. Attention and the speed of information processing: Posterior entry for unattended stimuli instead of prior entry for attended stimuli. *PloS ONE. 2013;* 8,e54257.

Yuan J, Liberman M. Speaker identification on the SCOTUS corpus, *Proceedings of Acoustics; 2008:*. 5687-5690.