The predominant pitch of semivowels[1]

Greg Finley
University of California, Berkeley

A remarkable characteristic of speech perception is that it can be active when listening to nonspeech. Numerous studies have shown speech perception operating to some degree upon nonspeech using a wide variety of stimuli and conditions (Remez & Rubin 1993, Shannon *et al.* 1995, Liebenthal *et al.* 2003, Berent *et al.* 2010, Iverson *et al.* 2011, Finley 2012), demonstrating either that nonspeech is intelligible as speech or that speech and language abilities affect the processing of nonspeech sounds. Taking a perspective that considers speech perception a cognitive process that operates on auditory input, there is evidence that the computation of linguistic outputs can be performed on a range of inputs that is not restricted to speech (Berent *et al.* 2010). This suggests that the effects observed in these studies are all a reflection of the same cognitive module responsible for perceiving natural speech, and that this module can be further explored, and perhaps even reverse-engineered, by observing its behavior given various types of auditory input.

One exemplary case of linguistic output from obviously nonspeech input is a phenomenon known as the predominant pitch (PP) of vowels. It has long been observed that pure tones of different pitches evoke different vowel qualities for listeners. An early thorough and systematic psychophysical evaluation of this phenomenon was by Farnsworth (1937), who polled subjects hearing tones ranging in frequency from 375 to 2400 Hz as to which of 13 English vowels best matched each tone. The most common three choices were /u/, /o/, and /i/, for which the respective median pitches were 500, 550, and 1900 Hz; the back /ɔ/, /ɑ/ had medians of 700 and 825 Hz and, taken together as for speakers without that contrast, constitute the fourth most common choice. Other studies (Fant 1973, Kuhl *et al.* 1991) vary in exact pitch ranges but confirm the trend of the back rounded vowels on the low end and the high front at the high end.

The identified pitch actually corresponds fairly well to the second vowel formant, although this is difficult to prove outright given the variance of vowel categories between speakers and the rather vague, noncategorical nature of the PP effect. Given the importance of F2 in phonetic categorization (especially for English vowels, as the distribution of rounding serves only to exaggerate the range of possible F2 values), it might not be surprising for a very sparse auditory input to be automatically attributed to this cue. If this is the case, then PP may be the very transparent application of speech perception to nonspeech auditory input, reflecting the same processes that are employed in normal speech listening.

A key difference between these stimuli and speech, however—beyond the obvious spectral differences—is in the temporal dynamics of the signal. Speech perception depends on processing both steady and time-varying spectra, but natural speech overwhelmingly comprises the latter. If PP does reflect the natural recruitment of speech perception, and not some other auditory effect, then it should operate also on stimuli with temporal modulation at rates similar to speech. Auditory pure tone stimuli used previously to gauge PP could fit the bill if modified to modulate in frequency over time. Of course this would also change the prediction of what speech sound such a tone would evoke: if a steady tone evokes a vowel, then the analogous speech

1 This work was presented as a talk at the 166th meeting of the Acoustical Society of America in San Francisco, CA on December 3, 2013. The title that appears in the conference program is 'Simple auditory elements induce perception of a phonetic feature'. This work was supported by NIH grant 1R01DC011295.

correspondence for a frequency-modulated (FM) tone would be a glide—a consonant that is articulated as a transition between two states of open articulation and acoustically resembles a continuum between two vowels. An experiment was designed to see if listeners would match FM tones to glides to and examine the PP effect on segments that were adjacent to one another rather than totally isolated.

Experiment: Audiovisual mixed speech and nonspeech

This experiment was designed to utilize both the visual and auditory modalities as a way of simultaneously representing speech and nonspeech. Nonspeech FM tones constituted the audio, and speech CV syllables the video. Two possible consonants were compared: the bilabial glide /w/ and the palatal glide /j/. Visually, the difference between these is hugely salient, particularly the rounding on /w/; acoustically, a major difference between them is the direction of F2 modulation, which sweeps downward from near the top of the natural F2 range for /j/ and up from the bottom of the range for /w/. (This assumes both glides are in onset position, and would naturally be reversed when following a vowel.)

Two visually distinct vowels, back low /ɑ/ and back mid /o/, were chosen to round out the CV syllables. All sounds are present in the inventory of American English. (The high back /u/ was not chosen given the absence of obvious transition between /w/ and /u/, and high front /i/ was not chosen given the lack of transition between /j/ and /i/.) The visual syllables to be tested, then, were /wɑ wo jɑ jo/. Audio stimuli were designed two correspond temporally to the speech syllables, and likewise consisted of two connected stages: a frequency sweep during the consonant and a steady pitch during the vowel. These sounds were made to vary in the direction of the sweep and also in the pitch range traversed by the tone. The stimuli are discussed in further detail below.

The visual modality of speech has been used to test PP before: in one experiment, Kuhl *et al.* (1991) displayed visual /a/ and /i/ video and asked subjects to match them to one of six steady tones (750, 1000, 1500, 2000, 3000, or 4000 Hz). This study uses similar methods, although the method of presentation is slightly different, and the visual and auditory stimuli are matched to CV syllables rather than bare vowels.

Subjects

Volunteers were recruited from the undergraduate student population of UC Berkeley. The first 14 subjects recruited comprised Group 1, and the latter 8 Group 2; each group performed an identical task with minor differences between them in the audio stimuli (see below). The entire task took no longer than 30 mintues. No subjects reported any history of language or hearing disorders. Participants were compensated with either cash or course credit.

Stimuli

All stimuli were short video clips (1.25 sec), with video and audio generated separately. Videos were unaltered clips of a 27-year-old male native American English speaker pronouncing CV syllables in isolation. The articulation was slightly exaggerated for clarity. Videos were cropped to a 360x360 resolution and compressed using lossy (MPEG) compression; however, compression artifacts were minor and did not in any way obscure the phonetics of the image.

When presented as stimuli, videos were uniformly stretched to fill the vertical dimension of the screen.

The audio from the original video recording was discarded and replaced with synchronized FM tones. These tones were of the same approximate length as the spoken syllables and were aligned manually by the experiment designer for best impressionistic coherence with the video, with the ultimate result being a video with a speaker who appeared to be uttering FM tones. (Audio was originally aligned automatically with video using an amplitude threshold of the original audio track as recorded by the video camera; however, the variability in the audio track, due in part to intrinsic vowel amplitude, led to some AV combinations looking much more plausible than others due only to the timing of tone onset, which was not what this experiment was designed to measure.)

The FM tones were generated algorithmically and written to WAV files using GNU Octave. The instantaneous angular frequency (ω) of the tones was defined using a logistic function, with a minimum approaching the starting frequency of the tone (ω_s) and a maximum that approached the final frequency (ω_f). Equation (1) below shows frequency as a function of time (t). Also provided are the starting and ending frequencies, the desired start time of the sweep (t_0), the length of the sweep (τ, set to 80 ms for these stimuli), and a parameter α ($0 < \alpha < 1$), which roughly designates the ratio of the sweep range left untraversed by the end of the sweep length to the total difference between the min and max frequency limits. (For these stimuli, α was set at 0.1, meaning that roughly 90% of the frequency change should happen by τ seconds after t_0.

$$\omega(t) = \omega_s + \frac{\omega_f - \omega_s}{1 + e^{\frac{2}{\tau}\ln(\alpha)(t-(t_0+\frac{\tau}{2}))}} \quad (1)$$

The 'start' and 'end' frequencies always differed by 3.5 Bark. The rate of FM, however, was not warped to conform to critical hearing bands, which may have interfered with the percept of a steadily sweeping tone. The function of the tone itself was defined as the cosine of the value of the cumulative integral of the logistic frequency function, as shown in (2); this function *x* was calculated for every sample at a sampling rate of 22050 Hz and written to a WAV file.

$$x(t) = \cos\left(2\pi \int_0^t \omega(T)\, dT\right) \quad (2)$$

The first portion of the tone, corresponding to the video consonant, was set to begin approximately when growth of the logistic function becomes noticeable; the interval over which growth slows significantly represents the transition from consonant to vowel portion (at length τ). An amplitude envelope similar to that of speech was applied to all tones, with a quick attack (15 ms) at the beginning of the consonant and a gradual decline (to 70% of max amplitude) followed by a rapid offset (15 ms) at the end of vowel. The total length of nonzero amplitude was 210 ms. Note that this effectively muted any steady tone at starting frequency, leaving behind only a rapid transition beginning at t_0 followed by a hold at ending frequency.

Finally, a separate dynamic amplitude envelope was applied to each tone. This envelope matched the value of an equal loudness contour (for 40 phon, about the level of normal

conversation) at the value of the instantaneous frequency of that tone. This resulted in tones all being matched for loudness: transitions that passed through the range from ~600–900 Hz, for example, had amplitude over these intervals slightly reduced to account for the comparative high sensitivity of the auditory system to tones at those frequencies. For well-formed WAV output, tones were normalized as a group; that is, they were not individually normalized, and differences between different tones in maximum amplitude were preserved. Digital audio was not compressed at any point.

Outlined above is the process applied to all stimuli. Variance between stimuli was in four key dimensions: two visual, spoken consonant and spoken vowel; and two acoustic, direction of the frequency sweep and the final frequency of the sweep. Consonants varied between /w/ and /j/ and vowels between /ɑ/ and /o/. Direction of frequency sweep varied between up and down. Final frequency varied between 700 Hz and 960 Hz for Group 1, between 1081 Hz and 1479 Hz for Group 2. Recall also that starting frequency is 3.5 Bark from final frequency, so two tones with opposite direction and same final frequency will have distinct starting frequencies 7 Bark apart from each other. A summary of all 12 unique audio stimuli is given in the table below.

	Low	High
Block 1: up	Group 1: **300** to **700** Hz Group 2: **569** to **1081** Hz	Group 1: **484** to **960** Hz Group 2: **838** to **1479** Hz
Block 1: down	Group 1: **1274** to **700** Hz Group 2: **1853** to **1081** Hz	Group 1: **1664** to **960** Hz Group 2: **2500** to **1479** Hz
Block 2	Group 1: **700** Hz Group 2: **1081** Hz	Group 1: **960** Hz Group 2: **1479** Hz

Table 1: The starting and ending frequencies for all audio stimuli.

All possible combinations of audio and video were generated, resulting in 16 different AV stimuli for Block 1 for each of the two groups. Block 2 was identical except that the starting frequencies were the same as the end frequencies, resulting in steady tones. Sweep direction was meaningless for this block, so only 8 distinct AV stimuli were generated for each group.

Procedure

Subjects viewed stimuli on a computer screen and listened over headphones. Audio was adjusted to a comfortable listening volume, similar to normal speech. Each trial consisted of two stimuli presented with an intervening 0.5 sec of silence and black screen. Within a trial, stimuli always shared the same audio but had different video. All possible pairings according to this constraint was generated, resulting in 48 different trial types (4 audio, 4 * 3 = 12 video pairings) for Block 1 and 24 (2 audio, 12 video pairings) for Block 2. For every trial, subjects were asked to indicate via keypress which of the two clips they had just seen, the first or second, had a better match between video and audio.

Subjects saw three repetitions of all Block 1 trials in random order during Block 1 (about 12 minutes) and two repetitions of all Block 2 trials in random order during Block 2 (4 minutes). (Between these sessions, subjects also saw a block [about 8 minutes] of similar stimuli that were analyzed for a different experiment.)

Results

All blocks of both groups show large numbers of all four syllables identified and generally balanced numbers for each type of response. Participants chose the first video of the pair in 2,168 of 4,224 trials (51.3%), so there is no reason to believe that choices were biased by presentation order. Identification of vowels was also largely split within each block: Group 1 identified the round vowel /o/ 53.9% and 56.5% of the time for Blocks 1 and 2, respectively; Group 2, 52.1% and 58.1%. Even for consonant identification, no single block showed an absolutely overwhelming preference for either /w/ or /j/: Group 1 chose /w/ in 72.1% and 49.3% of trials; Group 2, 52.1% and 42.4%.

The data were modeled to examine how video selection was predicted by the varying aspects of the auditory stimulus. (Recall that all trials consisted of a pair of clips with the same audio and different video.) Video parameters (consonant and vowel) were modeled as dependent variables with audio parameters (direction and end frequency for Block 1, direction only for Block 2) as independent variables. Because there were two consonant choices and two vowel choices, each was modeled as a binomial response in a mixed effects logistic regression model.[2]

Two models were fit for each block for each group: one modeled vowel choice and the other consonant choice, both using the two varying audio parameters as categorical predictors. The interaction between these predictors was also tested. Each model considered only those trials in which a choice was actually given between the two possible responses; that is, trials with two videos having the same consonant, such as /wɑ/ and /wo/, were excluded from the consonant model, and trials with two videos having the same vowel, such as /wo/ and /yo/, were excluded from the vowel model. This resulted in one third of all responses being excluded from each model, and one third of all responses (the video pairings /wɑ/, /jo/ and /wo/, /jɑ/) being shared by both models in each block.

Group 1

The fixed effects as handled by the models were ‘end’ (the final frequency of the FM sweep), either 700 Hz or 960 Hz, and ‘direction’, either up or down. Subject ID was the only random effect. Tables showing the estimated regression coefficients for these effects and their interaction upon consonant and vowel are given below for Block 1.

Group 1, Block 1: Consonant (/w/ or /j/)

predictor	coeff.	std. error	*z*-score	*p*
end (960 Hz)	0.232	0.206	1.13	0.26
direction (down)	1.58	0.188	8.42	<< 0.01 *
end : direction	0.0299	0.259	0.115	0.91

2 A multinomial regression model with all four videos as possible outcomes was also tested; however, treating the four as separate choices ignores the principled commonalities between them and ultimately makes it more difficult to interpret the effects of acoustic parameters on consonant and vowel. The multinomial model also had the disadvantage of implying the availability of a four-way choice for each trial, which was not the case.

Group 1, Block 1: Vowel (/ɑ/ or /o/)

predictor	coeff.	std. error	*z*-score	*p*
end (960 Hz)	–0.384	0.157	–2.44	0.015 *
direction (down)	0.659	0.164	4.03	<< 0.01 *
end : direction	–0.288	0.227	–1.27	0.20

When auditory stimuli were FM tones that varied in direction of modulation as well as frequency range, these variatons affected choices of vowel and consonant. Vowel choice was affected strongly by both direction and range, as shown in the table above: a 700 Hz ending tone and downward sweep both predict /o/ identification. The only reliable effect on consonant choice is direction, which shows that the upsweep predicts /w/ identification and the downsweep /j/.

Recall that these models are fitted to subsets of trials (those trials relevant to the consonant or vowel choice being modeled), and neither gives a complete picture of what visual syllable was matched to each sound over the entire block. Figure 1 below shows the proportions of all trials in Block 1 by chosen syllable and tone type (n = 2016).

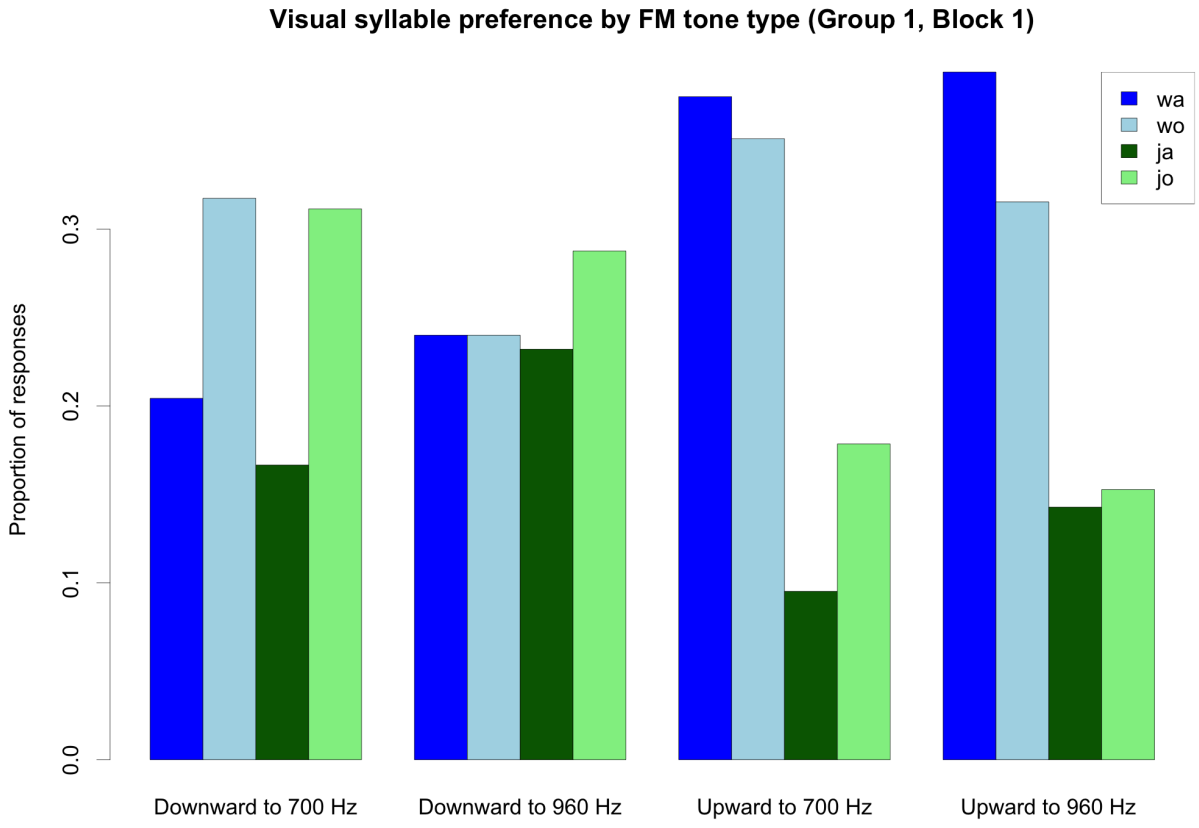

Fig. 1: Visual syllable proportions by auditory stimulus type for Group 1, Block 1

This plot reveals at least one aspect of consonant choice that is not clear in the regression model. Judging from the proportions of responses, /w/ is vastly preferred for upward tones, although there seems to be no preference at all between /w/ or /j/ for the downward tones. For vowel choice, when considering the effect of end frequency on the upward and downward groups separately, there are reliably more /o/ identifications for the lower range than the higher.

Vowel choice becomes easier to interpret in Block 2. Recall that the auditory stimuli for Block 2 were tones of a constant frequency. This eliminates direction as a predictor, leaving only frequency (700 Hz or 960 Hz).

Group 1, Block 2: Consonant (/w/ or /j/)

predictor	coeff.	std. error	*z*-score	*p*
frequency (960 Hz)	0.348	0.201	1.74	0.082

Group 1, Block 2: Vowel (/ɑ/ or /o/)

predictor	coeff.	std. error	*z*-score	*p*
frequency (960 Hz)	–1.40	0.208	–6.74	<< 0.01 *

Tone frequency has a clear effect on vowel but no reliable effect on consonant, with a higher tone predicting a visually unrounded vowel, exactly as in Block 1. Figure 2 shows the proportions of responses for Block 2.

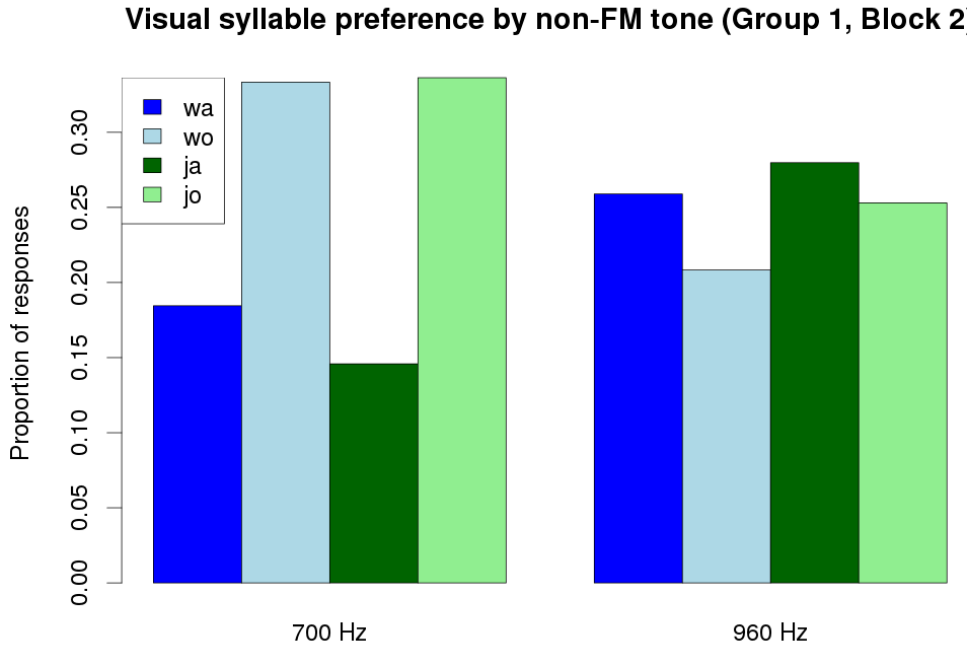

Fig. 2: Visual syllable proportions by auditory stimulus type for Group 1, Block 2

Note that /o/ is selected for the low tone roughly twice as often as /ɑ/, while /ɑ/ may have a slight advantage for the higher tone. (For glides, there also appears to be a very slight preference for /j/ on the higher tone, which is in the same direction as the model would predict, although pitch was not determined to be a statistically significant predictor for glide.) Given the clarity of vowel preference without FM, it appears that the effects on vowels in Block 1 were confounded somewhat by the strong consonant preference effects.

Group 2

Below are the tables showing regression coefficients for Group 2. The only difference between these and the Group 1 models is in the two available ending frequencies (or steady frequencies, for Block 2): 1081 Hz is the low, and 1479 Hz the high.

Group 2, Block 1: Consonant (/w/ or /j/)

predictor	coeff.	std. error	*z*-score	*p*
end (1479 Hz)	0.588	0.290	1.99	0.047 *
direction (down)	2.96	0.280	10.6	<< 0.01 *
end : direction	0.547	0.402	1.36	0.17

Group 2, Block 1: Vowel (/ɑ/ or /o/)

predictor	coeff.	std. error	*z*-score	*p*
end (1479 Hz)	–0.643	0.208	–3.09	0.0020 *
direction (down)	0.539	0.218	2.48	0.013 *
end : direction	–0.692	0.302	–2.30	0.021 *

As before vowel selection is influenced by both audio parameters. The directions of the effects are the same as for Group 1, although for Group 2 the interaction between range and direction appears to be a significant predictor on vowel, operating in the same direction as range but the opposite direction as sweep direction. This may indicate that for the tone for which the predictors 'clash'—that is, for the high-range, downward sweep—the static PP effect wins out and predicts /ɑ/.

Another divergence from Group 1 is the effect of range on consonant, with a higher tone predicting /j/ over /w/. Recall that these tones are higher overall than those seen by Group 1, and thus closer to the pitch range that PP associates with /i/ and, hypothetically, /j/. Given the special suitability of the high downsweeping tone for /j/, an interaction between the two effects might have been expected, but it was not observed to be significant. The effect of direction is still very strong, as it was for Group 1.

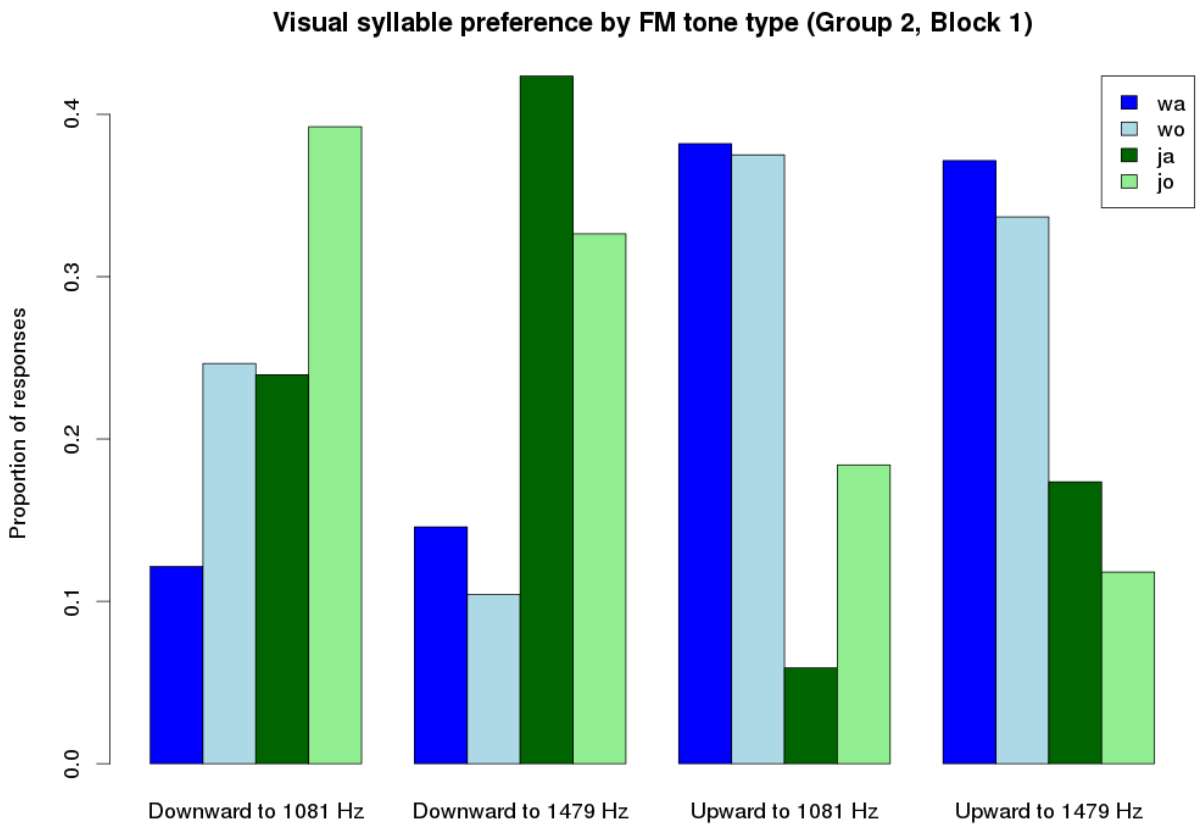

Fig. 3: Visual syllable proportions by auditory stimulus type for Group 2, Block 1

Taking all responses together, the distribution is somewhat like Group 1. A correlation test between the 16 different counts (4 syllables for each of 4 tones) from Group 1 and Group 2 shows $r^2 = 0.58$ ($p < 0.01$). A visible difference in this group, however, is that /j/ is chosen by a large margin over /w/ for the downsweep, in contrast to Group 1's lack of clear preference for either glide given the downsweep. Given tones of higher frequency than those given Group 1, Group 2 enthusiastically choose /j/ for the downsweep. The vowel effect for Group 2 is similar to that of Group 1, although a bit clearer when examining all responses. The lower end tone (1081 Hz), which persists over the length of the vowel, shows more /o/ responses than the higher end tone. This trend persists and is clearer in Block 2, the tables for which are given below.

Group 2, Block 2: Consonant (/w/ or /j/)

predictor	estimate	std. error	*z*-score	*p*
frequency (1479 Hz)	0.175	0.262	0.668	0.50

Group 2, Block 2: Vowel (/ɑ/ or /o/)

predictor	estimate	std. error	*z*-score	*p*
frequency (1479 Hz)	–1.40	0.208	–4.89	<< 0.01 *

These results are also qualitatively similar to Group 1: steady pitch has an unambiguous effect on vowels and no reliable effect on consonants. These counts correlate with those from Group 1 fairly well, $r^2 = 0.72$ ($p < 0.01$).

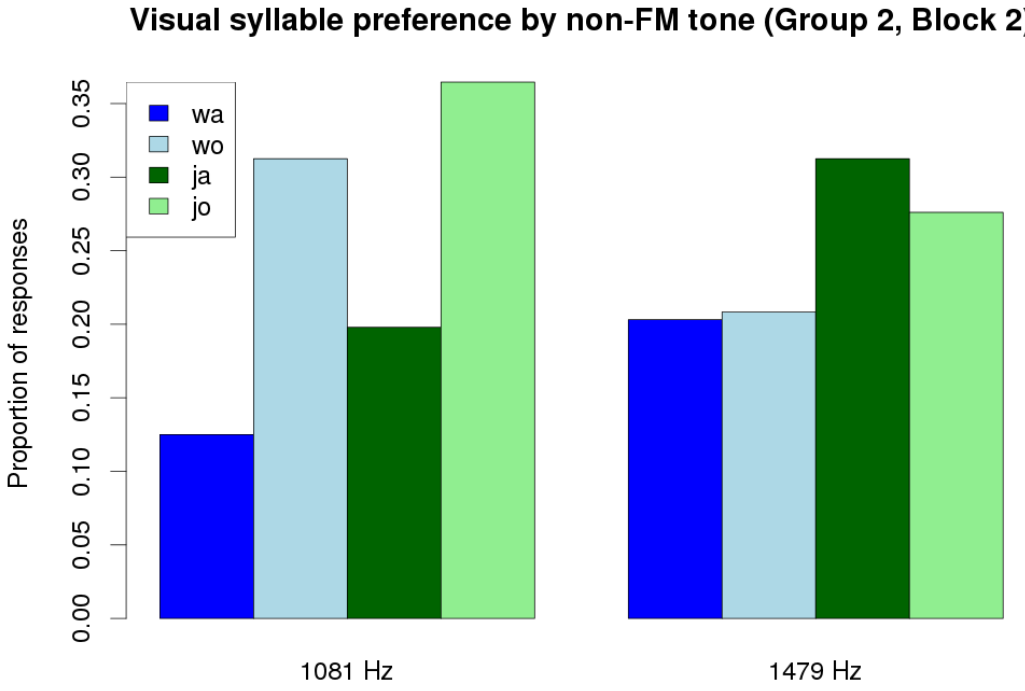

Fig. 4: Visual syllable proportions by auditory stimulus type for Group 2, Block 2

Actually, this distribution's similarity to Group 1's is surprising simply because the tones *are* considerably higher in frequency. The high tone for Group 1, 960 Hz, is just below the *low* tone for Group 2, 1081 Hz. The pitch difference between these is only about two semitones (versus about 5.5 semitones for the high-low contrast within each block), yet the syllables chosen for the 960 Hz and 1081 Hz tones are very different, as confirmed by a chi-squared test ($X^2 = 24.3$, $p < 0.01$); note for comparison that a chi-squared test taken over the entirety of Block 2 on syllable counts by Groups 1 and 2 fails to show significant independence between them ($X^2 = 7.41$, $p = 0.39$). Broadly speaking, the tone's status as low or high within its block is more predictive than its absolute frequency, which suggests that there is some context effect happening within the limited range of stimuli of the experimental block.

Discussion

In this experiment, the visual CV syllables chosen by subjects was predictable in part by features of the concurrent acoustic stimulus. It should be reiterated that at no point were subjects explicitly choosing speech sounds; rather, they were choosing from four visibly different consonant/vowel articulations, as seen in a head-on view. (From a rough survey of participants viewing videos without sound, /o/, /w/, and /ɑ/ were fairly consistently identified correctly, whereas there was considerable variation in how subjects interpreted /j/, although all indicated that it was some type of palatal or coronal sound.) Thus, to say that a subject chose /wo/ for a given sound is not as accurate as saying that the subject chose a face that began a very rounded articulation and gradually opened, with the lips remaining somewhat rounded for the duration (and similarly for the other syllables). For simplicity, the remainder of this discussion will refer to subjects' choices by the sounds articulated by the speaker in the video; the implications of these choices actually being visual articulations rather than phones will be briefly discussed later.

Vowels

The effects observed here that tone frequency has on segment identification are consistent with past work on vowel PP for both vowel and consonant identification. A lower tone was a better match for /o/ over /ɑ/, but this persisted as a relative effect as much as an absolute one. Although the lower tone heard by Group 2 is a better match for /ɑ/ than /o/, according to the PP literature, listeners still preferred the rounded vowel for this pitch, presumably because there was a higher tone and a higher-PP vowel available in the same block. This relative effect, as well as the fact that identification boundaries for the PP effect are so difficult to determine, suggests that these boundaries are especially malleable. Just as phoneme boundaries can shift due to context, as in normalizing a speaker's vowel space in the context of a conversation, the boundaries for vowel identification of tones also shifts given the context, in this case the other trials of an experimental block.

The PP of vowels was also shown to be relative over a much shorter duration: the syllable itself. Though the most dramatic effect of FM direction was on consonant identification, a positive effect was still observed on vowels. For both groups, /o/ was a more likely choice if the second (vowel) portion of the tone was at the lowest frequency of the FM sweep. Similarly to the block-level effect, the absolute-pitch boundaries for /o/ and /ɑ/ were modulated by the effects of context, in this case the immediately neighboring part of the tone. This may be an effect of the same process as the block-level relative effect described above; at any rate, the PP vowel identification boundaries are clearly malleable and shift in the same directions as would be predicted by contrast effects and normalization of speech.

A final remark on vowels concerns the steady pitch in Block 2 and the lack of a consonant effect. When steady tones were presented alongside CV visual syllables, the tone's pitch acted as a predictor on vowel but not on consonant, suggesting that when trying to match a CV syllable to this simple auditory stimulus—which could be considered a highly degraded speech signal, in a way—the listener's first priority is to determine the vowel. The predominance of the vowel in this case should not be surprising, given that the vowel is the part of the syllable most sonorous and therefore most robust to degradation. (The long duration of the vowel relative to the glide would also predict the vowel's prioritized status; note as well that length is an acoustic parameter positively correlated with sonority.) It is also possible, although certainly not

proven from these results, that the lack of acoustic modulation during the consonant phase of the tone causes listeners to assume that the tone is associated only with the vowel. Recall that when the FM was present in Block 1, the magnitude of the consonant effect made the simple PP effect on vowels more difficult to interpret. Listeners appear to be quite adept at associating the modulation with a consonant, but only if this modulation is present; otherwise, the tone seems not to evoke any consonant.

Glides

When FM was part of the stimulus, however, a clear association was shown between the upsweep and /w/ and between the downsweep and /j/, although the latter was much clearer for the higher-range sweeps seen by Group 2. The vowel qualities associated with these glides sit at the endpoints of the PP continuum, with /u/ corresponding to the bottom of the scale and /i/ to the top. Thus, any /wV/ syllable (other than /wu/), if converted into a trace of its vowel continuum's PP, should show movement from the bottom of the range to some other point, and similarly from the top of the range for any /jV/ syllable (other than /ji/).

As with vowels, it appears that the effect is relative as well as absolute, since for both groups an upsweep strongly predicted /w/, even when the frequency range was far from what PP ranges for the vowel /u/ would predict—the highest upward sweep, heard by Group 2, moves from about 838 Hz to 1479 Hz, which is well above where /u/ falls on the PP scale. There was no clear preference for /j/ over /w/ when hearing a downsweep, however, until moving up into higher frequencies.

This asymmetry suggests that the PP effect for /w/ (i.e., that an upsweep sounds like /w/) is more robust to changes in absolute frequency range than the analogous effect for /j/. An explanation may lie in the formant transitions of the two glides. With (initial) /w/, F1 and F2 sweep upward parallel to each other at close enough frequencies that they should not be resolvable separately (Bladon 1986), whereas F1 and F2 for /j/ move in opposite directions. (F2 and F3 of /j/ do also move in parallel, but they should be far enough apart in frequency to remain perceptually discrete cues.) Thus, the importance of a single dynamic spectral feature to /w/ identification is greater than to /j/, and a rising FM pure tone would be more similar to the rising F1-and-F2 of /w/ than a falling tone to the falling F2 /j/.

Theoretical discussion

The major contribution of this study was to show the extension of a phenomenon relating speech and nonspeech sounds to a condition that is more reminiscent of natural speech. Whereas previous studies of predominant pitch, the phenomenon by which a pure tone with no harmonic structure evokes a slight percept of a vowel, were limited to vowels in isolation, the experiment here generalizes this finding to those consonants whose articulation can be related to vowel quality. Though the difference between CV syllables and natural speech is still tremendous, these findings do indicate that PP is not limited to static spectra and can indeed operate on segmental transitions, which are important units in the processing of longer streams of speech.

The fact that PP applies to these transitions suggests that it holds some utility in describing the relationship between auditory perception and speech perception. As it stands, the PP phenomenon could be explained in one of two ways: either as the action of a general auditory spectrotemporal analysis strategy, upon which speech perception also relies; or as the

consequence of an auditory stimulus serving as input to a speech perception module, which attempts to pick out phonetically important cues and decode the nonspeech input as if it were speech. Because this experiment involved comparing this nonspeech stimulus to seen speech, not heard speech, the latter explanation may be more relevant to the results. In fact, the admission of nonspeech as input to the speech module may have been prompted or enabled by pairing the sounds with visual speech in the AV stimulus.

The other important outcome of this work is a demonstration of the flexibility of the vowel categories evoked by PP. Temporally proximal stimuli, either on the scale of seconds (the 1.25-second stimulus itself) or minutes (a 'conversational' unit like an experimental block), affected how listeners associated the tones with vowels, and the directions of these identification boundary shifts were consistent with effects of contextual contrast that underlie phonetic compensation, such as normalization for different speakers' vowel spaces. If anything, the capacity for contextual normalization of category boundaries seems higher in the case of the pure tone stimuli than it does for natural vowels. Further study is needed to determine if a mechanism is indeed shared between phonetic normalization and the relative PP effect here; if so, that finding would be another important piece of evidence that speech processing is occurring on these nonspeech stimuli.

Finally, these results should also be considered in relation to theories of speech perception that rely on the listener's recovery of a speaker's articulations (Liberman *et al.* 1967, Liberman & Mattingly 1985, Fowler 1986). For an experiment such as this one that pairs sound with visible gestures, a gesture-based account in fact provides a parsimonious explanation of the association between speech and nonspeech: the pairings of tones and faces are accomplished by matching the acoustic correlates of the gestures seen with the sounds heard, and thus no invocation of phonemes or other linguistic structures is necessary. If the interpretation of speech relies not at all on hearing sounds as articulatory gestures, then the PP effect in an AV experiment is not a direct association between the two modalities of the stimulus, but rather requires the participation of some higher-level representation (such as a phoneme or other high-level phonetic/linguistic knowledge). While this study certainly does not constitute evidence in favor of a gestural account, its results are consistent with an model of speech perception in which phonetic analysis arises from or is directly assisted by knowledge of articulation.

References

Berent, I., Balaban, E., Lennertz, T., Vaknin-Nusbaum, V. (2010). Phonological Universals Constrain the Processing of Nonspeech Stimuli. *Journal of Experimental Psychology: General* 139 (3), 418–435.

Bladon, A. (1986). Phonetics for hearers. In G. McGregor (ed.) *Language for Hearers*. Oxford: Pergamon.

Fant, G. (1973). *Speech Sounds and Features*. Cambridge, MA: MIT Press.

Farnsworth, P. (1937). An Approach to the Study of Vocal Resonance. *Journal of the Acoustical Society of America* 9, 152–155.

Finley, G. (2012). Partial effects of perceptual compensation need not be auditorily driven. *UC Berkeley Phonology Lab Annual Report*, 169–188.

Fowler, C. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics* 14, 3–28.

Iverson, P., Wagner, A., Pinet, M., Rosen, S. (2011). Cross-language specification in phonetic processing: English and Hindi perception of /w/-/v/ speech and nonspeech. *Journal of the Acoustical Society of America* 130 (5), EL297–EL303.

Kuhl, Patricia K., Williams, Karen A., and Meltzoff, Andrew N. 1991. Cross-Modal Speech Perception in Adults and Infants Using Nonspeech Auditory Stimuli. *Journal of Experimental Psychology: Human perception and Performance 17* (3), 829–840.

Liberman, A., Cooper, F., Shankweiler, D. (1967). Perception of the speech code. *Psychological Review* 74 (6), 431–461.

Liberman, A. & Mattingly, I. (1985). The motor theory of speech perception revised. *Cognition* 21 (1), 1–36.

Liebenthal, E., Binder, J., Piorkowski, R., Remez, R. (2003). Short-term reorganization of auditory analysis induced by phonetic experience. *Journal of Cognitive Neuroscience* 15 (4), 549–558.

Remez, R. & Rubin, P. (1993). On the Intonation of Sinusoidal Sentences: Contour and Pitch Height. *Haskins Laboratories Status Report on Speech Research, SR-113*, 33–40.

Shannon, R., Zeng, F., Kamath, V., Wygonski, J., Ekelid, M. (1995) Speech recognition with primarily temporal cues. *Science* 270 (5234), 303–304.