

Tone-evoked vowels and semivowels

Gregory Finley

University of California, Berkeley

Abstract

Prior research has shown that listeners associate pure tones of certain frequencies with certain vowels, with strong associations between high tones and the vowel [i] and between low tones and back rounded vowels [u] and [o]. Studies of this phenomenon with speakers of several languages are reviewed in detail and compared with similar results from experiments involving filtered vowels and sine-wave speech. New experiments were conducted in which listeners matched frequency-modulated tones with video of glide–vowel syllables, as glides are a natural dynamic analogue to tone-evoked vowels. Results show that the direction, frequency range, and rate of modulation can affect judgments on both parts of the glide–vowel syllable. The labiovelar glide [w] was associated with modulation through the range associated with [u], and the palatal glide [j] with the range associated with [i]. Relative effects of tone frequency are also observed, indicating that in the absence of full spectral cues, the categorical boundaries of tone-evoked vowels are flexible and context dependent. Implications for theories of speech perception and for models of spectral recognition are discussed.

Tone-evoked vowels and semivowels

A common tactic in researching speech perception is to ask targeted questions using nonspeech stimuli. Many of the more interesting discoveries have come out of this strategy; even without robust phonetic cues or a recognizable human voice, listeners hear speech. Famous examples of these types of stimuli include sine-wave speech (SWS; Remez *et al.*, 1981), noise-vocoded speech (Shannon, 1995; Davis *et al.*, 2005), and one- or two-formant speech (Delattre *et al.*, 1952; Finley, 2012). Researchers usually conclude that speech perception is robust enough to occur even with certain natural cues missing; in some cases, they insist that a speech percept is possible because the cues preserved by a particular type of nonspeech are the indispensable ones. Other studies are concerned more with identifying the bare minimum cue fidelity for intelligibility—for example, reducing the number of noise bands used for speech vocoding by Shannon and colleagues.

Intelligibility of words and sentences is certainly a sufficient criterion for nonspeech being recognized as speech, but the parity between speech and nonspeech sounds can be much more subtle. Probably the most extreme case of a speech percept drawn from a decidedly nonspeech acoustic stimulus is the phenomenon by which single pure tones of different frequencies resemble different vowels to a listener, with lower tones generally being identified with back vowels and higher tones (in the range of 1.5 to 4 kHz) with front vowels. The correspondence between vowels and tones is not as robust as that between speech and more spectrally complex nonspeech, but past studies have shown the association to be predictable and repeatable. This phenomenon, which I call *Vokalcharakter* following Köhler (1910), has received sporadic attention for some time but is lacking both a comprehensive review and an adequate explanation; this paper seeks to remedy the first point and offer direction towards addressing the second. I

will also present results of original experiments and discuss their relevance to what is currently known. I begin, however, with the review of *Vokalcharakter*, followed by a discussion of experiments with acoustically similar stimuli—most notably, filtered vowels and SWS.

Tones and vowels

Although the earliest systematic studies appear in the early 20th century, the relationship between vowel quality and a single frequency peak was remarked upon much earlier. Referring to photostats of Isaac Newton's original notebooks (*c.* 1665), Ladefoged (1967, p. 65) transcribes: 'The filling of a very deepe flaggon wth a constant streame of beere or water sounds y^e vowells in this order w, u, ω, o, a, e, i, y' (the symbols are Ladefoged's best printed equivalents to Newton's handwriting). Helmholtz (1954; final German version, 1877) notes that the major resonance of the back vowels from high to low constitute an ascending series of tones, which is continued by the higher resonance of the front vowels from low to high. Köhler (1910; as summarized by Weiss, 1920) ascribed the property of *Vokalcharakter* to pure tones, with categories occurring roughly every octave: 256 Hz corresponds to [u], 526 Hz to [o], and 1066 Hz to [a].

Weiss (1920) carried out what is probably the earliest systematic experimental mapping between pitches and vowels, asking listeners to match the sounds of tuning forks (ranging from 128 to 1152 Hz) with one of eight vowels. Unfortunately, Weiss's results are difficult to interpret due to high test/retest variability, as well as variability between the populations studied. (Note also that the sounds of tuning forks do not have the same spectral or temporal characteristics of constant-amplitude pure tones.) The most thorough study for English is by Farnsworth (1937), who played tones ranging from 375 to 2400 Hz generated by a beat frequency oscillator and

asked listeners to identify the vowel. The most common vowel choices were [u], [o], and [i], for which the respective median frequencies were 500, 550, and 1900 Hz; [ɔ] and [ɑ] had medians of 700 and 825 Hz and, if lumped together, constitute the fourth most common choice. Overall, the results suggest a continuum similar to Newton's.

Systematic research on *Vokalcharakter* is, happily, not restricted to English. Engelhardt and Gehrcke (1930) address German vowels, Fant (1973) Swedish vowels, and Chiba and Kajiyama (1958) Japanese vowels. (Note that the latter study does not directly test the mapping between pitch and vowel but does identify a ‘principal formant’ that characterizes each of the five Japanese vowels and speculates that this alone is sufficient to identify the vowel.) The availability of these languages is actually quite fortuitous because all feature rounded front or unrounded back vowels, and a natural question to ask from the English data alone would be how rounding changes a vowel’s associated tone. Results show that the effect of rounding is dwarfed by that of place: the German and Swedish studies indicate that [y] tends to favor a slightly lower tone than [i] but not as low as [e] (Fant’s results show that central [ɯ] is rarely associated with any tone), and the supposed unique principal formant of Japanese [ɯ] is still hypothesized to be lower than that of [o] (350 vs. 500 Hz). The Fant and Engelhardt studies are also valuable because they include responses to tones of up to 4 kHz, and both show that listeners overwhelmingly choose [i] above 3 kHz, while [y] dominates in the 2 to 3 kHz range. How the boundaries between front vowels differ for speakers of languages without rounded front vowels—or, put another way, what English-speaking listeners would do with the space that German and Swedish listeners seem to allocate to [y]—is a question that would probably require direct study of both speaker groups with very high tones.

The studies mentioned above have relied on imagined vowels, usually by presenting listeners with a word bank, one with each possible vowel response. Kuhl *et al.* (1991) showed that this phenomenon can also operate across modalities: given video-only presentations of spoken vowels, listeners tended to match an [ɑ] face with lower tones (750, 1000, or 1500 Hz) and an [i] face with higher tones (2, 3, or 4 kHz). The results from the audiovisual condition qualitatively matched those with imagined speech and also recorded vowels. Though their study tested only [i] and [ɑ] productions, the audiovisual presentation method is extensible to the entire continuum of tone-evoked vowels for English speakers, as there are salient and generally unambiguous visual articulation for three broad categories: mid/high front vowels, with high jaw and lips unrounded or even wide; low vowels, with open jaw; and mid/high back vowels, with lips more rounded than for other vowels.

Kuhl *et al.* name this phenomenon ‘predominant pitch’. I deviate from their terminology for two major reasons: first, to stress that the imagined vowel is triggered perceptually by the tone, and not that the vowel has an inherent pitch; ‘predominant pitch’ seems to suggest the latter, which may be confusable with the tendency for vowels high in the space to be *produced* with a higher rate of vocal fold vibration. Second, it is important to avoid ascribing tonal *Vokalcharakter* to pitch in the psychoacoustic sense, which I contend is independent of the spectral analysis at the root of the effect. That pitch and spectrum can be perceived from the same tonal stimulus has actually been demonstrated for SWS: Remez and Rubin (1993) show that the acoustic correlate of perceived intonation in SWS sentences is the first formant analogue, which also contributes to the intelligibility of the stimulus. For the remainder of this paper, the term ‘pitch’ is reserved for its psychophysical sense, and the rate of oscillation of a simple tone will always be described as its ‘frequency’.

Similar sounds

Can *Vokalcharakter* be explained entirely by the spectral characteristics of the tone itself?

To answer this question it is helpful to consider experiments on the identification of filtered vowels. Though speech is complex and broadband, filtered speech will approach a pure tone with the narrowing of the passband. If identification of narrowband-filtered vowels matches tone-evoked vowels for that range, it would bolster the intuition that the effect is spectral in nature. (Speech intelligibility under certain filtering conditions has also been studied extensively; see Cunningham [2003] for a review. For the purposes of studying *Vokalcharakter*, I am concerned here specifically with identification of isolated vowels.)

An early study of filtered vowels was conducted by Lehiste and Peterson (1959), who asked listeners to identify low- and high-pass filtered English vowels at cutoff frequencies from 550 to 4800 Hz. With high-pass cutoffs at and above 2100 Hz, vowel were overwhelmingly identified as [i] or, less commonly, the tense front [e^t]. When low-pass filtering the vowels at 540 Hz, nearly all tokens were identified as a back rounded vowel [u], [o], [o^v], or [ɔ]. (Results were similar for low-pass 950 Hz, although [ɑ] was usually identified correctly in this case.) These results match those from TEVs: low tones, especially under 1 kHz, strongly evoke back vowels like [u] and [o^v], while high tones evoke [i], even those tones much higher than the dominant spectral peak of [i]. Shriberg (1992) finds similar confusions for vowels filtered at 1 kHz: with low-pass filtering, front vowels are often identified as back or central vowels, and with high-pass filtering, back vowels are likely to be misidentified as central or front.

Missing so far is an equivalent to mid-range tones, which would be better characterized by band-pass filtering. Chiba and Kajiyama (1958) apply several filtering strategies to Japanese

vowels and make the identification judgments themselves. One of their conclusions is that ‘every vowel turns into **a** or **a^o** with B. P. 900—1600’ (p. 208). Taking these studies together, it appears that the three most common broad TEV categories I noted in Farnsworth’s and Fant’s results—mid/high back vowels, low vowels, and mid/high front vowels—can all be predicted by the phonetic quality of filtered vowels, with the center of the passband roughly corresponding to tone frequency.

Frequency modulation

Prior research on TEVs has been limited to steady tones evoking spectra of single vowels. If the phenomenon is due to the same mechanism underlying speech perception, as it appears to be, then it should also be possible to observe speech percepts associated with frequency-modulated (FM) tones. This has been done extensively with multi-tone complexes in SWS, which usually features three tones continually modulating in both frequency and amplitude to match the frequency and bandwidth of formants in the speech from which it was generated. For longer utterances, particularly those with few obstruents, SWS is highly intelligible. When dropping to a single formant analogue, however, virtually all intelligibility is lost (Remez *et al.*, 1981). When presenting the formants separately, there is evidence that F2 *contributes* the most to intelligibility: when presenting unaltered video of speech with single-formant sinusoidal analogues, Saldaña *et al.* (1996) show that more correct syllables are identified when a sine-wave analogue of F2 is present, but not when either F1, F3, or signal-correlated noise is present.

For FM tones to consistently evoke speech sounds, however, they may have to be designed more deliberately than selecting SWS components. I designed such stimuli and tested their associability to speech using the visual modality, in a paradigm with some similarities to Kuhl *et*

al. (1991). The experiments described in the remainder of this paper extend *Vokalcharakter* to semivowels and investigate the interactions between vowel and semivowel identification within the same syllable.

Experiment 1

If tones with dynamic frequency can have *Vokalcharakter*, then the natural analogical extension is from vowels to semivowels—segments with vowel-like acoustics but with rapid change. An obvious choice for semivowel to test is [w]: its early portion is acoustically virtually identical to [u], which is strongly evoked by low tones; it is extremely visually prominent, as it involves a transition from the lips being unrounded to rounded; and it is in the phonological inventory of American English speakers, who were recruited as subjects.

To present a clear visible [w] with context, video of the CV syllable [wɑ] was filmed; as a similar case without the rounded semivowel, [bɑ] was also filmed. *Vokalcharakter* predicts that [wɑ] should match perceptually with a tone that starts low and rises. I also hypothesized that the rate of FM should impact how well the tone evokes a glide versus a more rapid event, such as formant transitions from a stop closure: generally speaking, it should be expected that the total duration of detectable FM should match the duration of detectable visual modulation—i.e., lip movement. FM tones can be varied in a number of ways, which are discussed in detail below; by asking subjects to choose between the two speech videos as a match for a variety of FM tone types, it is possible to model their choice of visual syllable as a function of the controllable acoustic properties of the tone.

Methods

Subjects

Volunteers were recruited from the undergraduate student population of UC Berkeley. 28 subjects were recruited and split evenly into two groups. Each group performed an identical task with minor differences in the stimuli between the two. No subjects reported any history of language or hearing disorders. Participants were compensated with either cash or extra credit for an introductory linguistics course.

Stimuli

All stimuli were short video clips (1.25 s), with the video and audio tracks generated separately. Videos were unaltered clips of a 27-year-old male native speaker of American English pronouncing CV syllables in isolation. The articulation was exaggerated very slightly for visual clarity. Videos were cropped to a 360x360 resolution and compressed using lossy (MPEG) compression. 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 computer 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 of a person 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 from scratch as a function of time. The instantaneous angular frequency ω of a tone was defined as a logistic function of time, with a minimum approaching the starting frequency of the tone ω_s and a maximum that approached the final frequency ω_f . Equation (1) shows frequency as a function of time t . Also provided are the starting and ending frequencies, the start time of the sweep t_0 , the length of the sweep τ , 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 frequency change over all t . (For these stimuli, α was set at 0.1, meaning that ten elevenths of the frequency change will happen by τ seconds after t_0 .)

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

The start and end frequencies always differed by 3.5 Bark. The rate of FM was not warped to conform to critical hearing bands, as this may have interfered with the percept of a steadily sweeping tone. The audio signal itself was defined as the cosine of two pi times the value of the cumulative integral of the logistic frequency function, as is typical procedure for generating FM tones. This function was evaluated for every sample at a sampling rate of 22050 Hz and written to a WAV file.

The first portion of the tone, corresponding to the video consonant, was aligned with the brief period of rapid change in tone frequency (from t_0 to $t_0 + \tau$); the interval over which growth slows rapidly represents the transition from consonant to vowel portion. 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 (i.e., at t_0) 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. This effectively muted any steady tone at starting frequency, leaving behind only a rapid transition from starting frequency ω_s beginning at t_0 followed by a hold at ending frequency ω_f .

Finally, a separate dynamic amplitude envelope was applied to each tone to account for frequency-dependent loudness. This envelope matched an equal loudness contour (for 40 phon, about the level of quiet conversation) at the value of the instantaneous frequency of that tone. As a result, amplitude fluctuated somewhat over the course of the sweep such that no part of the sweep would sound louder than any other part. For WAV output, the tone with the highest peak amplitude was normalized, and all other tones were scaled by the same amount, so differences in maximum amplitude between tones were preserved. Digital audio was not compressed at any point.

Acoustic sweeps varied along three dimensions: direction, frequency range (defined by the ending frequency), and the duration. (Note that, although all stimuli were 210 ms, the amount of time allocated to the sweep changed depending on τ .) 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 that starting frequency is 3.5 Bark from final frequency, so starting frequency is always predictable from direction and ending frequency, and two tones with opposite direction and same final frequency will have starting frequencies 7 Bark apart. The min and max frequencies for each sweep for each group are summarized in Table 1.

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

TABLE 1: Starting and ending frequencies for all audio stimuli.

Duration varied between 30, 50, and 80 ms. Each group, then, was exposed to 12 auditory stimuli types. Each of these sounds was then paired with a video of [ba] and a video of [wa], for 24 unique 1.25-second videos per group.

Procedure

Stimuli were presented over a computer screen and headphones. Audio was set by the experimenter to a comfortable listening volume, similar to speech. Subjects performed a two-interval forced-choice task in which each interval had the same audio but different video (one [wa], one [ba]). The task was to judge which of the two intervals had the ‘best match’ between audio and video; no specific instruction was given as to what criteria should be used to evaluate this match. Every trial was seen three times in random order for 72 total trials (12 audio stimuli * 2 video orderings * 3 repetitions). The entire block took about 8 minutes.

Results

Because the response variable of preferred video is binary, these data can be analyzed using logistic regression. It was arbitrarily decided to consider [wa] the positive response and [ba] the negative. Video preference was modeled as a function of variables related to the auditory stimulus: sweep direction, ending frequency, and duration. The results of the analysis are given in Table 2 for Group 1 and Table 3 for Group 2.

Predictor	β	SE	z	p
(Intercept)	0.37	0.048	7.63	< 0.001
Direction: up	0.19	0.030	6.30	< 0.001
End freq: high	0.018	0.030	0.60	0.55
Duration (ms)	0.0027	0.00073	3.66	< 0.001

TABLE 2: Logistic regression: [wɑ] vs. [ba] for Group 1

Predictor	β	SE	z	p
(Intercept)	0.38	0.047	8.12	< 0.001
Direction: up	0.34	0.030	11.5	< 0.001
End freq: high	-0.073	0.030	-2.5	0.012
Duration (ms)	0.0011	0.00072	1.49	0.14

TABLE 3: Logistic regression: [wɑ] vs. [ba] for Group 2

For both groups, an upward FM sweep was significantly more likely to be identified as [w] than a downward sweep. Only Group 1 showed a significant effect of sweep duration, with a slower/longer sweep predicting more [w], while only Group 2 showed a reliable effect of ending frequency, with the higher predicting more [b] identification.

These data can be visualized simply by plotting the percentage of [w] identification for each type of auditory stimulus. Figures 1 and 2 show the entirety of the responses for Groups 1 and 2; the dotted line marks the point of 50% between [b] and [w] for that stimulus.

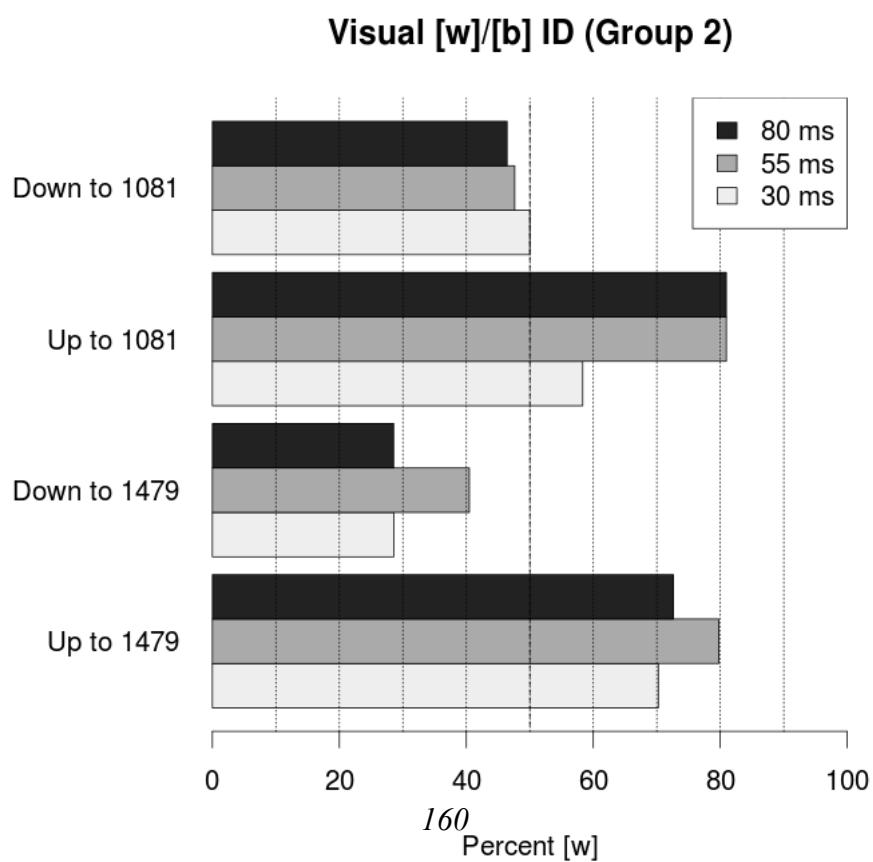
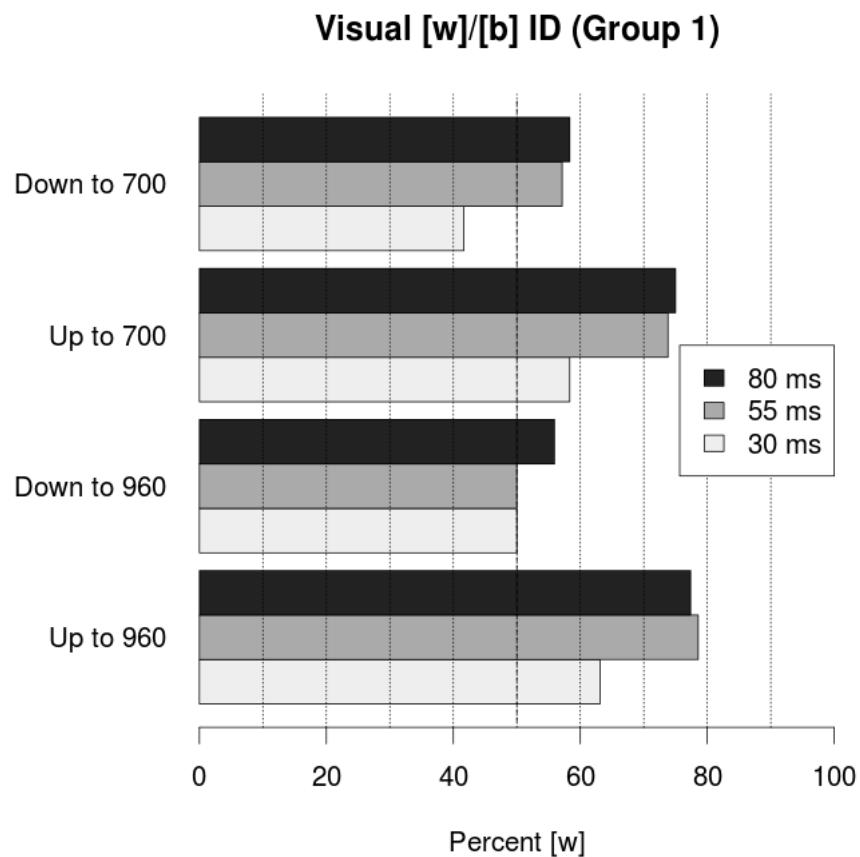


FIGURE 1: Percentage of visual syllable [wɑ] chosen for each of 12 FM tone types, Group 1.

FIGURE 2: Percentage of visual syllable [wɑ] chosen for each of 12 FM tone types, Group 2.

Trends found to be significant in the logistic models are clearly visible in these plots: sets of bars for the upward direction are generally higher than for downward. Note that the observed significance of sweep duration for Group 1 seems to be driven by the 30 ms stimuli, which more favor [b], with little difference between 50 ms and 80 ms. For Group 2, a similar pattern shows up only for the rising low tone—that is, the tone whose *Vokalcharakter* would be predicted to be most like [w].

Discussion

The observed effects of stimulus type on [w] identification reflect the acoustic properties of [w]. It was assumed *a priori* that a rising tone sweep would be especially evocative of [w] given the spectral similarity between [u] and the pre-transitional part of [w]. The second part of the syllable should also be a good match for English [ɑ]: the *Vokalcharakter* of a tone near 1 kHz should resemble this vowel, and the upward direction and the visual opening mouth both suggest movement away from a high back vowel sound to some other open sound. In terms of total counts, the data favor [w] over [b]. This bias suggests that the observed differences are driven more by enthusiasm for [w] than for [b]. At the same time, subjects did not *overwhelmingly* choose [w]; however, they should not be expected to overwhelmingly choose one or the other given the constant binary choice, which probably led them to consider that [b] should be the preferred answer for at least some of the stimuli.

The tested differences in ending frequency were not significant for Group 1. Both were rather low for this group. Starting frequencies for the upward sweep were either 300 Hz or 484

Hz, both of which are near the center of the low spectral pole formed by the first two formants of [u]. In Group 2, however, the higher upsweep stimuli *started* at 838 Hz, generally above both formants for [u]. Recall that Farnsworth (1937) found that [u] had a median of 500 Hz and [ɑ] 825 Hz. The low upsweeps for Group 2 started at 569 Hz, near Farnsworth's median [o] value. As both [o] and [w] feature rounded lips, which are visually salient and block out other articulators, the visual difference between them is subtle, if not unnoticeable.

Modulation rate seemed to not to be significant for Group 2, who heard higher tones higher in frequency but otherwise identical to Group 1. The spectral unsuitability of Group 2's tones to the visual speech may have prevented them from recruiting temporal cues to help decide. Nevertheless, the results from Group 1 alone are strong enough evidence that temporal similarity between the FM tone and the phone type can bias identification.

The fact that acoustic parameters of the FM tone affect how it evokes speech sounds confirms that listeners can make use of dynamic spectral information for this type of speech-nonspeech processing. However, although the temporal characteristics of the stimulus in this experiment were somewhat complex, there was little variation in the visual syllables, especially in terms of the spectra that listeners would associate with them. To show a wider variety of effects in the speech perception of FM tones, a second experiment was conducted that gave listeners different choices of visual syllables that would have more spectrally diverse hypothetical FM correlates.

Experiment 2

As noted earlier, previous work has determined that filtered vowels and tones are most strongly evocative at the extremes: low-pass spectra tend to associate with back rounded vowels

and high-pass spectra with front vowels, most notably [i]. As [i] analogizes to the glide [j], the set of visual stimuli was expanded to include this consonant. To further explore the spectral effects of these stimuli, the vowels were also varied between rounded and unrounded.

A control condition for this experiment was also included that used steady tones rather than sweeps. For clarity, the stimuli and results of that condition are discussed separately as ‘Experiment 2B’.

Methods

Subjects

The same subjects from Experiment 1 participated, with the same division into two groups.

Stimuli

Stimuli were generated in the same manner as Experiment 1 but with different visual syllables. Four were available: [wɑ], [wo^o], [jɑ], and [jo^o]. As before, labial articulation was slightly exaggerated for maximum clarity.

The only difference in audio stimuli from Experiment 1 was the exclusion of a sweep length contrast: tones varied only in direction and range (τ was always set to 80 ms), so each group heard four distinct tone types: two sweep directions for each of two ending frequencies.

Procedure

Stimuli were presented in the same manner as Experiment 1. However, with only two intervals and four available videos, the pairings of videos changed between trials. Every permutation of two videos for every possible audio stimulus was presented three times in random order, for a total of 144 trials. The entire block took about 12 minutes.

Results

Unlike Experiment 1, responses are not exactly binary, as there are four videos available. One way to model responses might be to use one-versus-all multinomial logistic regression. Such an analysis is difficult to interpret, however, because it does not separate the effects of acoustic parameters on visual consonant selection from those on vowel selection. Consonant and vowel selection *are* binary, so one straightforward way to model them is with separate logistic regressions for consonant and vowel. (Note that this means that two statistical tests are being conducted on the same data, increasing the possibility of type 1 error. One way to correct for multiple comparisons is by adjusting the significance threshold α ; as a conservative correction for two comparisons, α was halved from the standard 0.05 to 0.025.)

Each model considered only those trials that had a contrast in the variable in question; that is, trials with video options of [w α] and [wo^o] or [j α] and [jo^o] were excluded from the consonant model, and those with [w α] and [j α] or [wo^o] and [jo^o] were excluded from the vowel model. Results of both of these models for Group 1 are given in Tables 4 and 5. Note that [w] and [o^o] are coded as 1 for the purposes of the model, and [j] and [α] as 0—positive coefficients favor rounded lips.

Predictor	β	SE	z	p
(Intercept)	0.13	0.099	1.28	0.20
Direction: up	1.53	0.13	12.0	< 0.001
End freq: high	-0.24	0.12	-1.96	0.050

TABLE 4: Logistic regression: [w] vs. [j] for Group 1

Predictor	β	SE	z	p
(Intercept)	0.48	0.079	6.13	< 0.001
Direction: up	-0.32	0.090	-3.58	< 0.001
End freq: high	-0.33	0.090	-3.67	< 0.001

TABLE 5: Logistic regression: [o^ø] vs. [ɑ] for Group 1

The consonant results are similar to those found in Experiment 1 when considering the role of [w] in both: an upward modulation strongly predicts [w], but the difference in frequency range is not significant. Recall in the prior discussion that both the low and high upsweeps for Group 1 should be expected to evoke movement away from a rounded back vowel quality.

Both acoustic parameters have reliable effects on vowel identification, with tones that are lower—even if only very locally so, i.e., following a downswEEP—being predictive of [o^ø] identification.

As in Experiment 1, the proportions of each visual stimulus chosen can be visualized using the barplot in Figure 3. Note that these plots are drawn differently than those for Experiment 1, which showed percentage [w] identification. Because decisions are no longer binary, each of the four video types gets its own bar, and these four are grouped together for every type of stimulus. Every group of four adds to 1 (bars are displayed side-by-side for clarity).

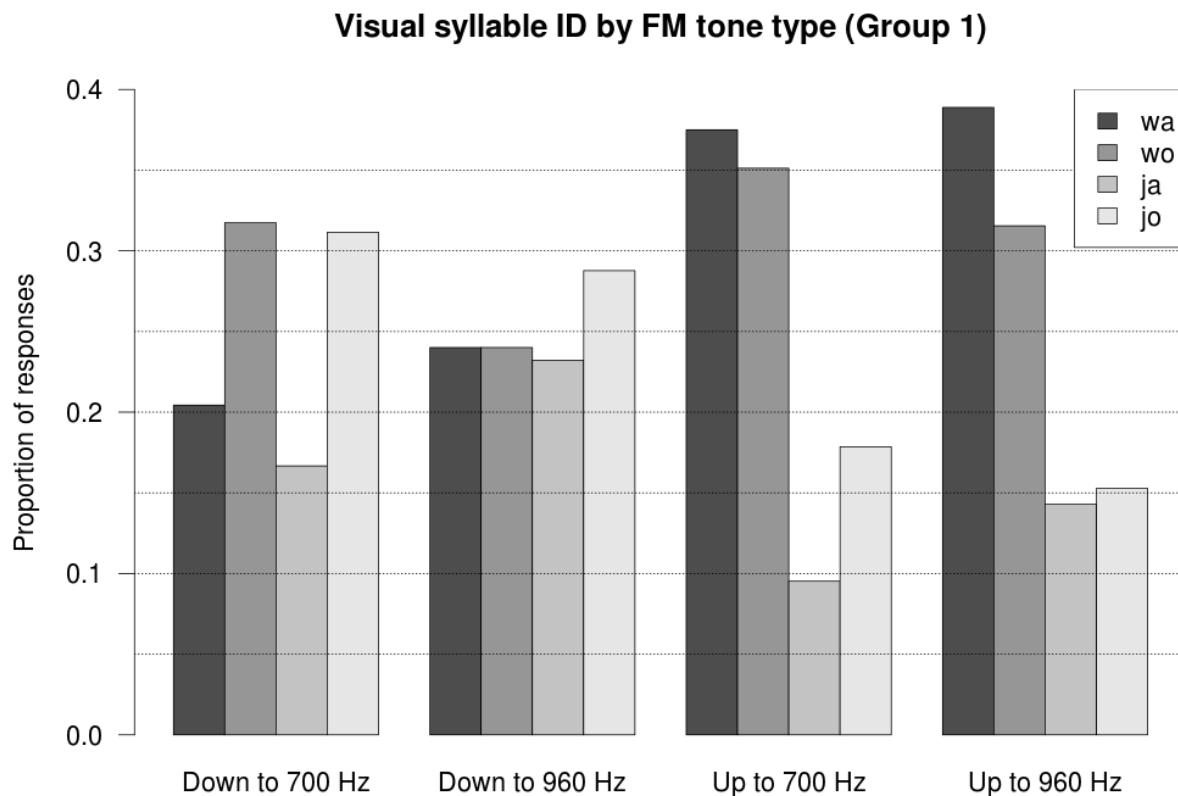


FIGURE 3: Percentage choice of all four visual syllables for each of four FM tone types, Grp. 1.

This plot illustrates at least one aspect of consonant choice that is not clear from the regression model. Judging from the proportions of responses, [w] is vastly preferred for upward tones, although there is no clear preference 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^o] identifications for the lower range than the higher.

The same techniques were applied to the data for Group 2. The models are summarized in Tables 6 and 7.

Predictor	β	SE	z	p
(Intercept)	2.20	0.14	15.6	< 0.001
Direction: up	2.68	0.14	18.8	< 0.001
End freq: high	-0.74	0.14	-5.40	< 0.001

TABLE 6: Logistic regression: [w] vs. [j] for Group 2

Predictor	β	SE	z	p
(Intercept)	0.26	0.078	3.28	0.0011
Direction: up	-0.17	0.090	-1.90	0.058
End freq: high	-0.67	0.090	-7.45	< 0.001

TABLE 7: Logistic regression: [o^v] vs. [a] for Group 2

As in Experiment 1, the gap between Group 2's low and high tones makes frequency a significant predictor. In this case, the negative coefficient indicates that sweeps ending at the higher target predict [j], while those at the lower target predict [w]. For the vowels model, ending frequency is still a strong predictor, but the effect of sweep direction seen in Group 1 is much less reliable.

Proportions of each video response are given in Figure 4. Note that the figure highlights a difference between Groups 1 and 2 that is not entirely clear from the regression model: Group 2 enthusiastically chooses [j] for the downsweeps, especially the higher of the two, while Group 1 seems to show no clear glide preference for these stimuli ([w] versus [j] counts on downsweep between groups: $\chi^2 = 34.6, p < 0.001$). Only for Group 2 do the downsweeps partially traverse the frequency range with *Vokalcharakter* typical of high front vowels and the palatal glide.

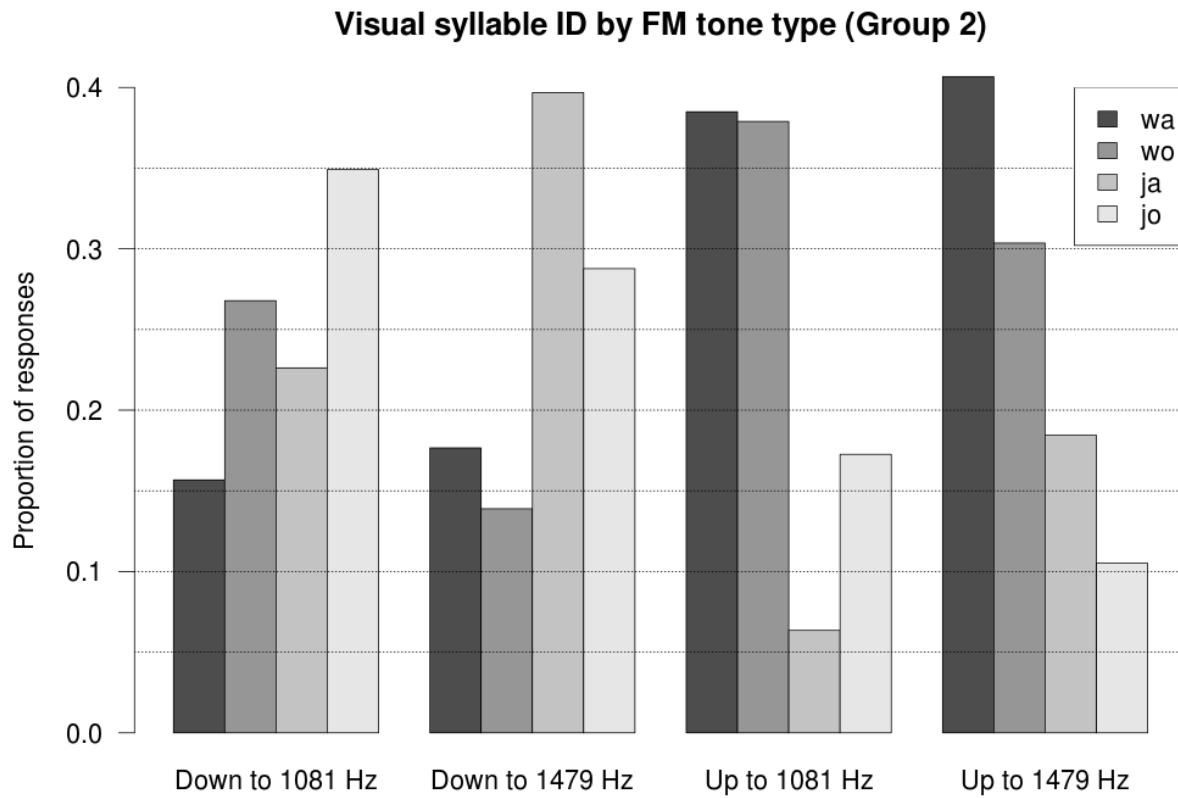


FIGURE 4: Percentage choice of all four visual syllables for each of four FM tone types, Grp. 2.

Discussion

Whereas Experiment 1 tested the suitability of FM tones to a glide versus a stop, including the effect of modulation rate, this one asked subjects to choose between sounds with different spectral but identical temporal characteristics. When varying the direction and absolute frequency of the tone sweep, clear associations between tones and glides emerged. Frequency range only mattered for glide identification in Group 2, who saw sweeps starting as high as 2500 Hz, well within the range found to evoke [i]. Group 1's highest downsweep started at 1664 Hz, which is not a particularly good match for the semivowel [j] suggested by the visual, and results bore this out. Overall, the effects of sweep direction and range for semivowel selection are

entirely consistent with the TEVs documented in previous work, and the present results show that these associations can be straightforwardly generalized to temporally modulated stimuli.

Vowel choice also seemed to depend on aspects of the sweep, which was not concurrent with the visual vowel, indicating that identification is influenced by temporally proximal tones. I will return to this aspect of the results in the general discussion; before doing so, it is helpful to consider the simple case of vowel identification when no FM is present, which was measured in Experiment 2B.

Experiment 2B

For this short condition, audio stimuli differed from Experiment 2 in that the tones were not modulated in frequency. Stimuli were generated using the same method but with starting and ending frequencies being equivalent. All other aspects of the stimuli were the same. With duration and direction both rendered irrelevant, only two tone types, high and low, were available to each group. All possible combinations of one tone and two videos were generated, and two repetitions of each trial were presented, for a total of 48 trials. Subjects completed this session quickly, in about 4 minutes.

Results

For both groups, steady tone frequency had a significant impact on the numbers of responses of each vowel (Group 1 $\chi^2 = 28.8, p < 0.001$; Group 2 $\chi^2 = 18.2, p < 0.001$), but not on consonant responses ($\chi^2 = 1.52, p = 0.22$; $\chi^2 = 0.02, p = 0.88$). Response counts can be visualized similarly to Experiment 2 and are shown in Figures 5 and 6.

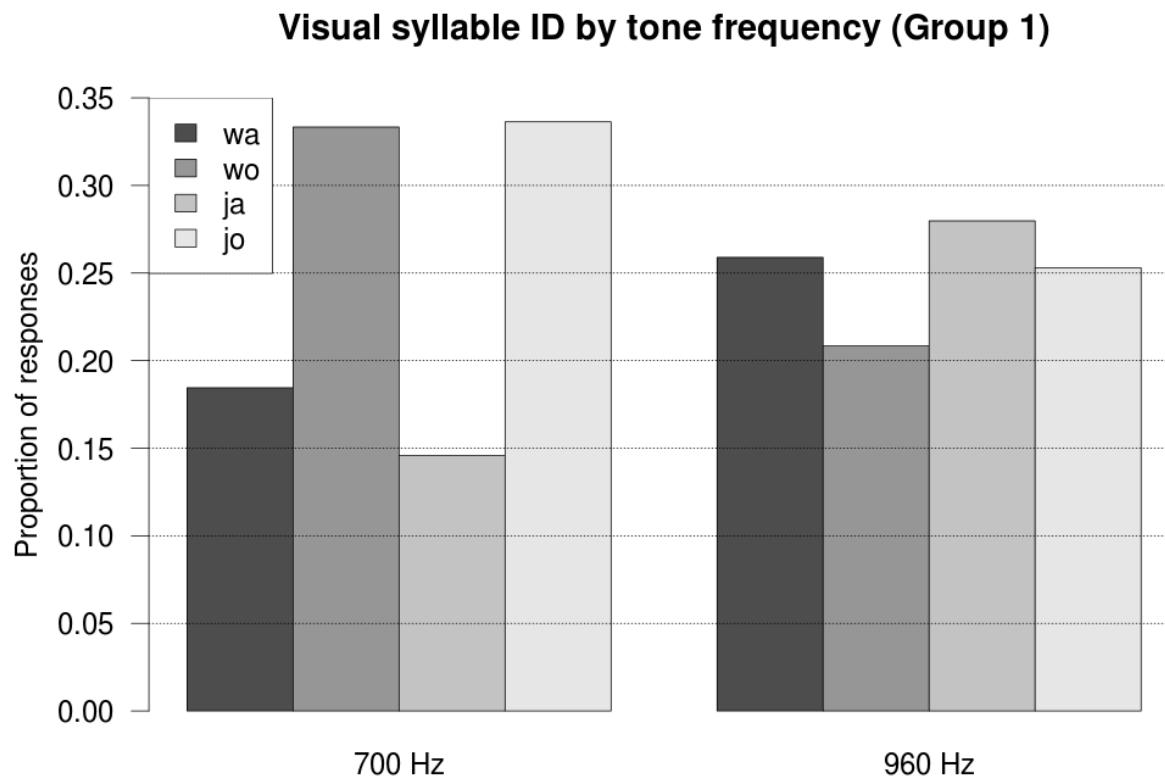


FIGURE 5: Percentage choice of all four visual syllables for both steady tones, Group 1.

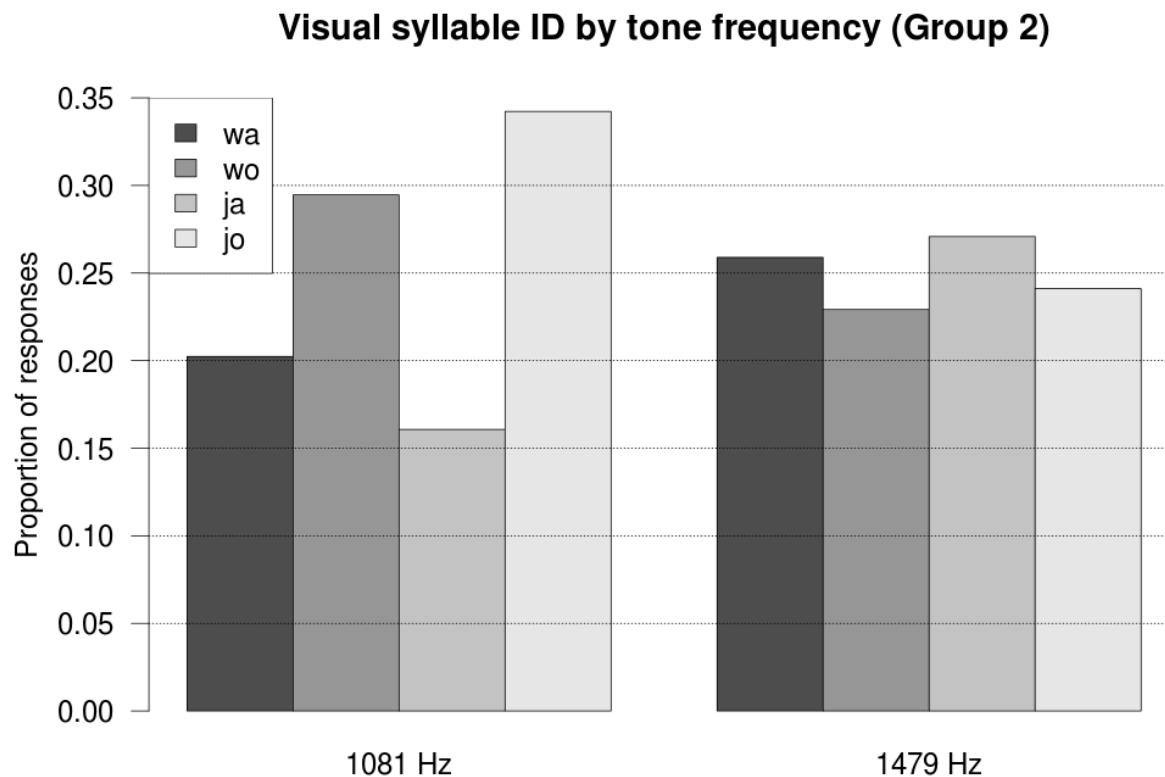


FIGURE 6: Percentage choice of all four visual syllables for both steady tones, Group 2.

The results from this condition are simpler to interpret: when asked to choose both a vowel and consonant but are given only a simple tone, listeners extract a vowel percept before a glide percept. There is no evidence that either glide is evoked when FM is not present. As is consistent with prior research, when only two tone types are heard in the same block, [o^v] is a more popular choice for the lower of the two.

An interesting difference does emerge between the two groups for this condition. Both show similar identification patterns between their respective high and low tones, with [o^v] weakly preferred for the lower tone and no clear preference for the higher tone. What is of interest here is the fact that the low tone for Group 2 is just slightly *higher* than the high tone for

Group 1, yet the pattern of behavior seems relative to the other tone heard in the same block.

Presenting two stimulus types in the same block has been found to exaggerate effects of their differences (Johnson, 1990); in the context of this experiment, the observed adaptation to other stimuli in the block may follow from long-term auditory adaptation to statistical properties of frequency, as demonstrated by Holt (2005).

General discussion

The experiments in this study demonstrate that listeners can extract dynamic spectral cues from a minimally spectral dynamic stimulus—an FM tone. Observations of *Vokalcharakter* and the frequency ranges typically associated with various vowels can be cleanly generalized to semivowels, at modulation rates similar to speech. This extension is consistent with the view that the tone-vowel association is due to the same mechanisms that process speech. Although single FM tones are certainly not intelligible as speech, under controlled circumstances they have clear associations with speech sounds. I further discuss three points here. First, I address the theoretical significance of nonspeech-as-speech processing. I then delve further into the effects of relative frequency on vowel identification noted in my experiments. Finally, I offer some perspectives on the spectral features likely to be at the root of tone-evoked speech.

Speech processing of tones

Why are tones receiving the speech treatment? Past research has illustrated many cases of unmistakably phonetic or linguistic processing being applied to nonspeech: language experience (Iverson *et al.*, 2011); phonetic context effects (Finley, 2012); and even universal and language-specific phonology (Berent *et al.*, 2010). More generally, there is ample evidence of top-down linguistic/phonetic influences upon auditory perception (e.g., Davis & Johnsrude,

2007). It is reasonable and economical to posit that these tones are processed as speech would be, and phonetic identification follows automatically. This processing is probably at least partially motivated by the fact that the tones are concurrent with visual articulation, inducing the expectation of speech; there is behavioral and neuroimaging evidence that it is possible to control, by exploiting experience or expectation, whether auditory input is processed in a ‘speech mode’ (Repp, 1982; Remez *et al.*, 2001; Liebenthal *et al.*, 2003; Möttönen *et al.*, 2006).

Processing artificial nonspeech as speech could have, paradoxically, ecological motivations. Real-world listening conditions are virtually never ideal—reverberation, acoustic filtering, noise, etc. are all possible obstacles to intelligibility. Speech perception needs to remain robust to these destructive effects; it would not be well served by the exclusion of hypothetically degraded inputs. The identification functions of pure tones do differ from those of speech in at least one aspect: the lack of a clear categorical boundary between two phones. Although the tones *evoke* speech sounds, the cues are generally insufficient to make a firm judgment. This aspect of the data suggests that, although listeners eagerly interpret phonetics from any kind of auditory input, the system remains agnostic when that input only partially matches known speech patterns.

Relative effects on vowel

The experiments presented here agree with past work on the approximate frequency ranges associated with broad vowel quality categories. As mentioned, there are no sharp boundaries in identification—neither in prior work, in which there are large bands of overlap between adjacent tone-evoked vowels, nor here, where a particular vowel or glide is usually not overwhelmingly chosen over the other option. The vagueness of the boundary suggests that it may be malleable,

but no previous study has attempted to induce boundary shifts for these kinds of stimuli.

Although that was not the stated intent of these experiments, I did find that vowel identification depended on both the frequency of the vowel-synced tone and the direction (and thus frequency range) of the preceding sweep. The sweep had minimal temporal overlap with the visual vowel, yet vowel preference showed a clear effect of direction: the vowel [o^ø] was more common when the vowel-synced portion was low in frequency *relative* to the consonant portion. Though reliable *Vokalcharakter* boundaries are difficult to measure, there is evidence here that the boundaries can be shifted by context.

The simple nature of the shift in this case—tones are modulated either up or down, with frequency change and rate held constant—makes it difficult to distinguish whether the boundary shift is best explained as phonetic or auditory in nature. If the effect is phonetic, the explanation would lie in compensating for the preceding glide, which would putatively move the vowel either forward or back. Consider these same experiments but with speech stimuli: nearly this same condition, ambiguous vowels in [w-w] or [j-j] contexts, was found by Lindblom and Studdert-Kennedy (1967) to induce compensatory phonetic effects. On the other hand, the effect could be considered purely auditory, as would be predicted by the framework of spectral contrast (Lotto *et al.*, 1997; Lotto and Kluender, 1998). Under this view, the spectral distance between points of high energy would perceptually exaggerated using cognitively general contrast mechanisms. The phonetic and auditory viewpoints make equivalent predictions for the simple case demonstrated here, and neither should be disregarded as a possible explanation.

Spectral features of tone-evoked speech

Finally, I would like to return to the question of exactly what it is about pure tones that makes them evocative of certain vowels. A naïve approach of matching the tone to vocal tract resonances actually finds striking parity between tone and F2 frequency. (It could also be said, for the back and low vowels, that the tone corresponds to the single spectral bump created by the first two formants.) However, an explanation of *Vokalcharakter* that rests only on F2 is problematic because it does not account for the predominance of low vowels, as opposed to mid or high central vowels, at around 1 kHz. Note that Fant (1973) shows very little identification of tones in this frequency range as the Swedish high central vowel, whereas an account relying on F2 only would predict this vowel would be as popular a choice as [ɑ]. The Swedish data reinforce that we need to know not only where the vowels fall along the continuum, but also which vowels are most strongly evoked. For English, these vowels do appear to be those that have a somewhat band-limited characteristic: [u] and [o] are largely devoid of high frequencies; [ɑ], slightly less so, but the high F1 and low F2 do leave significant gaps outside of a narrow mid-frequency band; [i], excepting its very low F1, is essentially high-pass, as is [y]. Although none of these vowels are exactly narrowband, all feature broad bands of very low spectral energy near their characteristic peaks. Tone spectra exhibit this same property to an extreme degree. Again, identification of filtered vowels bears out the same predictions: when introducing spectral zeros through filtering, misidentification tends to favor [i] for very high-pass sounds, and back rounded vowels for very low-pass sounds (Lehiste & Peterson, 1959).

That these spectral zeros are a property that is clearly shared between tones and the vowels they commonly evoke suggests that they are important perceptual cues. As another piece of

evidence, consider that recognition accuracy of spectrally gapped speech can be enhanced by adding noise in the empty bands (Shriberg, 1992; Warren *et al.*, 1997; Bashford *et al.*, 2005; McDermott & Oxenham, 2008; Carlyon *et al.*, 2002). If the filtering of speech creates artificial zero cues that listeners attempt to use, added noise effectively removes these spurious cues and forces the listener to remain agnostic to those bands. It also creates a much more ecologically plausible stimulus, as natural background noise is much more common and plausible than natural sharp bandpass filters.

The tone-evoked speech results suggest that a model of speech sound recognition considering only poles in the vocal tract's transfer function—i.e., formants—is insufficient to predict how these stimuli will be classified. Although formant frequencies constitute a useful, low-dimensional representation of speech acoustics, competing models that take the entire spectrum into account have been successful as well, and are more plausible given our knowledge of psychoacoustics and auditory physiology (Bladon & Lindblom, 1981; Bladon, 1982; Ito *et al.*, 2001; Molis 2005). This is a major debate in the literature and should take into account results from nonspeech stimuli. Even cases of auditory stimuli generating a very subtle speech percept, as with *Vokalcharakter*, should not be considered irrelevant to speech perception; on the contrary, these very controlled stimuli offer a unique perspective on how to reverse-engineer the cognitive machinery.

Human speech processing allows a wide variety of possible inputs, and models of spectral recognition need to make correct predictions for *any* stimulus with speech associations. The present work has demonstrated the possibility for stimuli of minimal spectral complexity to evoke dynamic speech and related this behavior directly to their spectral properties. Other findings of the experiments, such as the relative effects on vowel association over the length of a

tone or an entire experimental block, can be related to known auditory and phonetic phenomena. I also suggested a research direction towards an explanation for tonal *Vokalcharakter* and noted ways in which current models of spectral perception fall short. These findings should reinforce the usefulness of nonspeech work in phonetics and highlight the extraordinary ability of human listeners to find speech in difficult conditions, including the absence of speech altogether.

References

- Bashford, J. A., Warren, R. M., & Lenz, P. W. (2005). Enhancing intelligibility of narrowband speech with out-of-band noise: Evidence for lateral suppression at high-normal intensity. *J. Acoust. Soc. Am.* 117(1): 365–369.
- Berent, I., Balaban, E., Lennertz, T., & Vaknin-Nusbaum, V. (2010). Phonological Universals Constrain the Processing of Nonspeech Stimuli. *J. Exp. Psychol.* 139(3): 418–435.
- Bladon, R. A. W., and Lindblom, B. (1981). Modeling the judgment of vowel quality differences. *J. Acoust. Soc. Am.* 69(5), 1414–1422.
- Bladon, A. (1982). Arguments against formants in the auditory representation of speech. *The representation of speech in the peripheral auditory system*, 95–102. Elsevier Biomedical Press.
- Carlyon, R. P., Deeks, J., Norris, D., & Butterfield, S. (2002). The Continuity Illusion and Vowel Identification. *Acta Acustica united with Acustica* 88, 408–415.
- Chiba, T., & Kajiyama, M. (1958). *The Vowel: Its Nature and Structure*. Phonetic Society of Japan.
- Cunningham, S. (2003). *Modelling the recognition of band-pass filtered speech*. Doctoral dissertation, University of Sheffield.
- Davis, M. H., & Johnsrude, I. S. (2007). Hearing speech sounds: Top-down influences on the interface between audition and speech perception. *Hear. Res.* 229, 132–147.
- Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., & McGettigan, C. (2005). Lexical Information Drives Perceptual Learning of Distorted Speech: Evidence From the Comprehension of Noise-Vocoded Sentences. *J. Exp. Psychol.* 134(2), 222–241.

- Delattre, P., Liberman, A. M., Cooper, F. S., & Gerstman, L. J. (1952). An experimental study of the acoustic determinants of vowel color: observations on one- and two-formant vowels synthesized from spectrographic patterns. *Word*.
- Engelhardt, V., & Gehrcke, E. (1930). *Vokalstudien: eine akustisch-psychologische Experimentaluntersuchung über Vokale, Worte und Sätze*. JA Barth.
- Fant, G. (1973). *Speech sounds and features*. Cambridge: MIT Press.
- Farnsworth, P. R. (1937). An Approach to the Study of Vocal Resonance. *J. Acoust. Soc. Am.* 9, 152–155.
- Finley, G. P. (2012). Partial effects of perceptual compensation need not be auditorily driven. *UC Berkeley Phonology Lab Annual Report*, 169–188.
- Helmholtz, H. L. F. (1954). *On the sensations of tone as a physiological basis for the theory of music* (A. J. Ellis, Trans.). New York: Dover. (Original work published 1877)
- Holt, L. L. (2005). Temporally non-adjacent non-linguistic sounds affect speech categorization. *Psychol. Sci.* 16, 305–312.
- Iverson, P., Wagner, A., Pinet, M., & Rosen, S. (2011). Cross-language specialization in phonetic processing: English and Hindi perception of /w/-/v/ speech and nonspeech. *J. Acoust. Soc. Am.* 130(5), EL297–EL303.
- Ito, M., Tsuchida, J., & Yano, M. (2001). On the effectiveness of whole spectral shape for vowel perception. *J. Acoust. Soc. Am.* 110(2), 1141–1149.
- Johnson, K. (1990). The role of perceived speaker identity in F0 normalization of vowels. *J. Acoust. Soc. Am.* 88, 642–654.
- Köhler, W. (1910). Akustische Untersuchungen I. *Zeitsch. f. Psychologie* 54, 241.

- Kuhl, P. K., Williams, K. A., & Meltzoff, A. N. (1991). Cross-Modal Speech Perception in Adults and Infants Using Nonspeech Auditory Stimuli. *J. Exp. Psychol.* 17(3), 829–840.
- Ladefoged, P. (1967). *Three areas of experimental phonetics*. Oxford University Press.
- Lehiste, I., & Peterson, G. E. (1959). The Identification of Filtered Vowels. *Phonetica* 4, 161–177.
- Liebenthal, E., Binder, J. R., Piorkowski, R. L., & Remez, R. E. (2003). Short-Term Reorganization of Auditory Analysis Induced by Phonetic Experience. *J. Cog. Neurosci.* 15(4), 549–558.
- Lindblom, B. E. F., & Studdert-Kennedy, M. (1967). On the Rôle of Formant Transitions in Vowel Recognition. *J. Acoust. Soc. Am.* 42, 830–843.
- Lotto, A. J., & Kluender, K. R. (1998). General contrast effects of speech perception: Effect of preceding liquid on stop consonant identification. *Percep. & Psychophys.* 60(4), 602–619.
- Lotto, A. J., Kluender, K. R., & Holt, L. L. (1997). Perceptual compensation for coarticulation by Japanese quail (*Coturnix coturnix japonica*). *J. Acoust. Soc. Am.* 102(2), 1134–1140.
- Molis, M. R. (2005). Evaluating models of vowel perception. *J. Acoust. Soc. Am.* 118(2), 1062–1071.
- Möttönen, R., Calvert, G. A., Jääskeläinen, I. P., Matthews, P. M., Thesen, T., Tuomainen, J., & Sams, M. (2006). Perceiving identical sounds as speech or non-speech modulates activity in the left posterior temporal sulcus. *Neuroimage* 30, 563–569.
- Remez, R. E., & Rubin, P. E. (1993). On the intonation of sinusoidal sentences: Contour and pitch height. *J. Acoust. Soc. Am.* 94(4), 1983–1988.

- Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. (1981). Speech Perception Without Traditional Speech Cues. *Science* 212, 947–949.
- Remez, R. E., Pardo, J. S., Piorkowski, R. L., & Rubin, P. E. (2001). On the bistability of sine wave analogues of speech. *Psychol. Sci.* 12(1), 24–29.
- Repp, B. H. (1982). Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. *Psychological Bulletin* 92(1), 81.
- Saldaña, H. M., Pisoni, D. B., Fellowes, J. M., & Remez, R. E. (1996). Audio-visual speech perception without speech cues. *ICSLP Proceedings* 4, 2187–2190.
- Shannon, R. V., Zeng, F., Kamath, V., Wygonski, J., Ekelid, M. (1995). Speech Recognition with Primarily Temporal Cues. *Science* 270, 303–304.
- Shriberg, E. E. (1992). Perceptual restoration of filtered vowels with added noise. *Language and Speech* 35(1, 2), 127–136.
- Weiss, A. P. (1920). The Vowel Character of Fork Tones. *Am. J. Psychol.* 31(2), 166–193.