JSLHR

Article

Roles of Voice Onset Time and F0 in Stop Consonant Voicing Perception: Effects of Masking Noise and Low-Pass Filtering

Matthew B. Winn,[a] Monita Chatterjee,[a] and William J. Idsardi[a]

Purpose: The contributions of voice onset time (VOT) and fundamental frequency (F0) were evaluated for the perception of voicing in syllable-initial stop consonants in words that were low-pass filtered and/or masked by speech-shaped noise. It was expected that listeners would rely less on VOT and more on F0 in these degraded conditions.
Method: Twenty young listeners with normal hearing identified modified natural speech tokens that varied by VOT and F0 in several conditions of low-pass filtering and masking noise. Stimuli included /b/–/p/ and /d/–/t/ continua that were presented in separate blocks. Identification results were modeled using mixed-effects logistic regression.
Results: When speech was filtered and/or masked by noise, listeners' voicing perceptions were driven less by VOT and more by F0. Speech-shaped masking noise exerted greater effects on the /b/–/p/ contrast, while low-pass filtering exerted greater effects on the /d/–/t/ contrast, consistent with the acoustics of these contrasts.
Conclusion: Listeners can adjust their use of acoustic-phonetic cues in a dynamic way that is appropriate for challenging listening conditions; cues that are less influential in ideal conditions can gain priority in challenging conditions.

Key Words: speech perception, noise, voicing contrast, bandwidth

Consonant voicing contrasts are very common in the world's languages (Ladefoged & Maddieson, 1996), and the perception of acoustic cues underlying these contrasts has been explored thoroughly for normal-hearing listeners and other listeners in quiet conditions. Much less is known about how voicing is perceived by individuals who rely on low-frequency hearing (e.g., individuals with hearing impairment [HI]) or individuals listening in background noise. It is clear that perception of voicing remains accurate across a wide range of signal degradations, including high- or low-pass filtering (Miller & Nicely, 1955), masking noise (Miller & Nicely, 1955; Phatak & Allen, 2007; Phatak, Lovitt, & Allen, 2008; Wang & Bilger, 1973), hearing impairment (Bilger & Wang, 1976), spectral degradation (Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995; Xu, Thompson, & Pfingst, 2005), or cochlear implantation (Friesen, Shannon, Başkent, & Wang, 2001). It is thus often stated that the amount of "information transfer" is high for the voicing feature relative to other consonant features. This finding is so consistent that some studies dispense with potential voicing confusions in the very design of the experiment (Dubno & Levitt, 1981). Despite this generalized high level of success, the constraints that face listeners in noise and/or those who rely on low-frequency hearing (e.g., listeners with hearing loss) are likely to change the means by which voicing is perceived—that is, the voice information could be recovered via different cues in varied listening conditions. In the current study, we show that in some degraded conditions, the role of voice onset time (VOT) decreases and the role of fundamental frequency (F0) increases for the perception of word-initial voicing in stop consonants.

[a]University of Maryland, College Park

Correspondence to Matthew B. Winn, who is now at the University of Wisconsin—Madison: mwinn83@gmail.com

Monita Chatterjee is now at the Boystown National Research Hospital, Omaha, NE.

Editor: Sid Bacon
Associate Editor: Marjorie Leek

Received March 18, 2012
Revision received August 30, 2012
Accepted December 28, 2012
DOI: 10.1044/1092-4388(2012/12-0086)

Cues for Voicing in Stop Consonants

Although perception of voicing in word-initial stop consonants has been largely attributed to VOT (i.e., the duration between consonant release and the onset of voicing for the following vowel; Lisker & Abramson, 1964), F0 plays a role as well. F0 is higher after voiceless stops than after voiced stops (House & Fairbanks, 1953), and this difference generally lasts roughly 100 ms into a vowel (Hombert, 1975). Although F0 is not a very potent cue for stop sounds with canonical voiced or voiceless VOTs (Abramson & Lisker,

1985), it can exert potent influence under certain conditions, such as for sounds with ambiguous VOTs (Abramson & Lisker, 1985; Haggard, Ambler, & Callow, 1970). When F0 contour conflicts with VOT, reaction time is slowed (Whalen, Abramson, Lisker, & Mody, 1993), suggesting that listeners are sensitive to F0 information even when identification curves suggest otherwise.

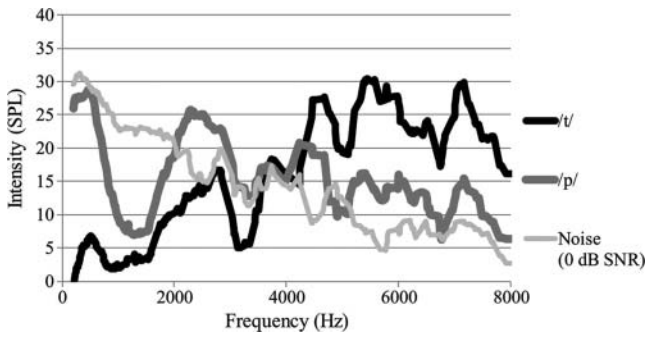
Another acoustic cue that may be useful to voicing perception is F1. Word-initial stop sounds are followed by transitioning formants, the first of which begins at a low frequency (e.g., 300 Hz) and ascends to its target frequency for the following sound. As VOT grows longer, that rising transition becomes increasingly devoiced, resulting in a higher F1 frequency at the onset of voicing; long VOT values yield higher F1 onsets, whereas low F1 onsets are characteristic of short VOT values (Lisker, 1975). An exception to this trend is for high vowels (e.g., /i/, /u/), whose voicing-related F1 perturbations are minimal because F1 begins and ends at low frequencies; F1 is thus not thought to aid the voicing contrast in high vowel environments. In this experiment, we focused on the contributions of VOT and F0 and hence relied on the /i/ vowel to provide an environment in which VOT perception could be measured without confounding changes in F1.

Both low-pass filtering and masking noise should affect VOT and F0 disproportionately. The aspiration noise that characterizes the VOT cue contains considerable energy in the high-frequency regions, particularly for the /t/ consonant (see Figure 1). Eliminating high-frequency energy from the signal would render the aspiration less audible while maintaining lower frequency harmonics in the vowel that drive F0 perception. Competing noise should more effectively mask the VOT than the vowel because the aspirated portion of the word is less intense than the vowel portion; the F0 is recovered from the vowel and should thus be relatively less affected by the noise.

The Effect of Low-Pass Filtering

Exploring the effect of limited bandwidth in terms of a low-pass filter (LPF) has particular relevance for understanding the experience of people with HI. High-frequency hearing loss can render some phonetic cues inaudible, potentially compelling listeners with HI to rely on different cues than those used by people with normal hearing. Furthermore, the lack of access to high-frequency auditory filters is likely to compromise temporal resolution because the wideband high-frequency filters are considerably superior in the temporal domain compared with the narrow-band low-frequency filters (Eddins, Hall, & Grose, 1992). For example, noise gap detection thresholds become smaller as bandwidth grows wider (Eddins & Green, 1995; Eddins et al., 1992; Grose, 1991). For this reason, it is likely that listeners who rely solely on low-frequency energy have poorer ability to use temporal cues (such as VOT) but remain receptive to residual information that should include F0.

Figure 1. Illustration of frequency spectra for /t/ (black) and /p/ (gray) aspiration noises, and the masking noise (light gray) at 0 dB signal-to-noise ratio (SNR).



Role of F0 in Noise

Previous research has shown that F0 is a useful cue for listening to speech in noise (Brokx & Nooteboom, 1982; McAdams, 1989). When the F0 contours of masked sentences are flattened or inverted around the mean, intelligibility decreases (Binns & Culling, 2007) and self-reported difficulty increases (Laures & Weismer, 1999), particularly when the masker is competing speech. It is likely that the F0 contour can help direct listeners' attention to the timing of target words to aid in intelligibility (Cutler & Foss, 1977). Although the utility of a natural F0 contour is well established at the sentence level, relatively little is known about the contributions of F0 contour to the intelligibility of individual segments or phonetic features. Fogerty and Humes (2012) showed that the flattening of F0 contours or the removal of F0 information (i.e., whisper-like speech) resulted in deficits for both vowels and consonants. Therefore, although vowels are the primary periodic element in the speech signal, consonant sounds stand to benefit from natural F0 contours as well.

Objectives and Hypothesis

The present study was designed to test whether F0 would become a more prominent cue for voicing in word-initial stop consonants in conditions of low-pass filtering and/or masking by speech-shaped noise. Unlike aspiration noise that characterizes the VOT cue, F0 should be perceptible even without high-frequency information. F0 has been previously shown to be a beneficial cue for listening to sentences in noise, but its use at the segmental level has not been fully understood. We hypothesized that in the aforementioned degraded signal conditions, listeners' voicing judgments would be driven more heavily by F0 and less by VOT.

Method

Participants

Participants included 20 adult listeners (15 women; mean age = 24.3 years) with normal hearing, defined as having pure-tone thresholds ≤ 20 dB HL from 250 to 8000 Hz in both ears (American National Standards Institute, 2010).

All participants were native speakers of American English and were screened for self-reported unfamiliarity with tonal languages (e.g., Mandarin, Cantonese, Vietnamese) to ensure that no participant entered with a priori increased bias toward using F0 as a lexical or phonetic cue. Informed consent was obtained for each participant, and the experimental protocol was approved by the institutional review board at the University of Maryland. Participants were reimbursed for their participation.

Stimuli

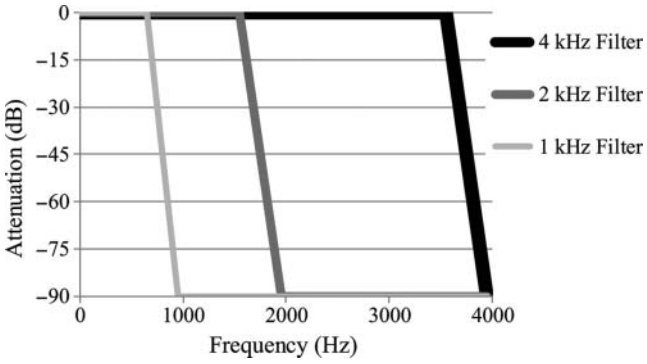
Two sets of stimuli were created using modified natural speech. The words "Pete," "beat," "teen," and "Dean," spoken by a male native speaker of English, were recorded in a double-walled sound-treated room with a 44.1 kHz sampling rate. We sought to measure the contribution of VOT without the added complementary cue of F1, so a high vowel /i/ was used for all consonant environments. Stimuli varied by VOT (in seven or eight steps for /b/–/p/ and /d/–/t/, respectively) and F0 (in eight steps). Following the method used by Andruski, Blumstein, and Burton (1994), onset portions of words with /b/ or /d/ onsets were progressively replaced with equivalently long portions of onset aspiration from /p/ or /t/, respectively, in 10-ms increments from the onsets (bound at the closest zero crossing) to create continua of VOT. Thus, the vowel from each stimulus item came from the /b/- or /d/-initial tokens. For the /d/–/t/ continuum, the VOT range spanned from 0 ms to 70 ms, and the range for the /b/–/p/ continuum spanned from –10 ms (prevoicing) to 50 ms, as suggested by previous studies (Abramson & Lisker, 1970; Lisker & Abramson, 1964).

Intensity of items in the /b/–/p/ series varied within 2.7 dB, and all items in the /d/–/t/ series varied within 1 dB; overall root-mean-square intensity was affected by VOT step because lower intensity aspiration noise progressively replaced higher intensity phonated vowel onsets. Volume was calibrated such that the end-point "voiced" stimulus of each series in the optimal (full-spectrum, in quiet) condition was 65 dBA.

The F0 contour of each stimulus was manipulated using the pitch synchronous overlap-add method in Praat (Boersma & Weenink, 2011). Steps in the F0 continuum were interpolated in eight steps along a log scale ranging from 94 to 142 Hz, which reflects the general range for male speech, as indicated by previous work (Abramson & Lisker, 1985; Ohde, 1984; Whalen et al., 1993). F0 was kept steady over the first two pitch periods of the vowel and fell (or rose) linearly until returning to the original contour at the 100-ms point in the vowel (the time indicated by Hombert, 1975). Following the 100-ms timepoint, all F0 contours were equal within each continuum.

Masking noise. Speech-shaped noise (SSN) was extracted offline from the iCAST program (Fu, 2006). Its spectrum was strongest in the 200–600 Hz region and decreased by roughly 6 dB per octave (see Figure 1). This noise was chosen to reflect the long-term average spectrum (LTAS) of conversational speech rather than the LTAS of our stimuli (the presence of only one vowel in our stimuli would yield a LTAS of only very limited utility with regard to everyday experience). Stimulus timing within the noise was roved so that there was 280–360 ms of noise before the stimulus and 380–450 ms of noise after the stimulus. To present varying signal-to-noise ratios (SNRs), noise levels varied while speech signals were kept at constant amplitude within each condition. Noise levels were set relative to the vowel onset so that SNR was not affected by intensity differences stemming from VOT continuum steps or by syllable type (stop-final or nasal-final). As most studies do not reference one particular point in a syllable to calculate SNR, caution is encouraged when comparing SNR levels in the current study with those from other publications. For example, the SNR at vowel onset for "Pete" is roughly 4 dB greater than that calculated from the entire word, resulting in greater masking to reach equivalent SNR.

Figure 2. Illustration of low-pass filters (LPFs) used in this experiment, implemented using the Praat software.



Low-pass filtering. Stimuli were low-pass filtered using the Hann band filter function in Praat (Boersma & Weenink, 2011), illustrated by the filter responses in Figure 2. Cutoff frequencies of 4 kHz, 2 kHz, and 1 kHz were used to investigate various degrees of residual acoustic information. Filtering was done after the addition of background noise to model the order of signal degradations encountered by a listener with HI. Therefore, the level of noise (and, hence, the overall signal) for conditions at poorer SNRs was more intense than those at better SNRs, and signals with higher LPF cutoffs were more intense that those with lower LPF cutoffs. It should be noted that although the vowels and final consonants in all speech stimuli were degraded by the masking noise and filtering, these segments were already identified by the visual word choices.

Conditions

The levels of low-pass filtering and SNR in this experiment were chosen to measure specific effects highlighted in Table 1.

The arrangement of conditions was inspired by preliminary experiments (Table 1, top row) that suggested that either 0 dB SNR or a 1 kHz LPF permitted use of the VOT

Table 1. Listening conditions tested for each comparison, defined by spectral bandwidth and signal-to-noise ratio (SNR).

Listening condition	Comparison			
Initial exploration of bandwidth and noise effects				
Bandwidth	Full	1000 Hz	Full	1000 Hz
SNR	Quiet	Quiet	0 dB	0 dB
Effect of bandwidth in 0 dB SNR noise				
Bandwidth	Full	4000 Hz	2000 Hz	1000 Hz
SNR	0 dB	0 dB	0 dB	0 dB
Effect of SNR with 1000 Hz LPF				
Bandwidth	1000 Hz	1000 Hz	1000 Hz	1000 Hz
SNR	Quiet	10 dB	5 dB	0 dB

Note. Rows are organized by the specific purpose of comparison. Note that the 1000-Hz, 0-dB SNR condition is present in all three rows. LPF = low-pass filter.

cue, but the combination of these factors promoted the use of F0 nearly exclusively. The combination of the 1 kHz LPF and the 0 dB SNR condition could be improved by either ameliorating the LPF settings or making the SNR more favorable. Thus, questions following this pilot testing included the following: (a) What LPF cutoff is necessary to facilitate the use of VOT when the SNR is 0 dB (middle row)? and (b) What SNR is needed to facilitate the use of VOT when the LPF is 1 kHz (bottom row)? Each of these conditions was tested for the /b/–/p/ stimuli and for the /d/–/t/ stimuli, resulting in a total of 16 conditions (note that some conditions are used for multiple comparisons but were tested just once).

Procedure

All testing was conducted in a double-walled sound-treated booth. Stimuli were presented in the free field through a single Tannoy Reveal studio monitor loudspeaker (frequency response: 65 Hz–20 kHz, ±3 dB) at a distance of 1–2 ft, placed in front of the listener at eye level. Listeners responded to these stimuli by clicking a button on a computer screen labeled with two word choices (either *Pete–beat* or *teen–Dean*). There was no time limit on their response, and they were permitted to enable stimulus repetitions up to three times; stimulus repetitions were very rare.

Conditions were defined by SNR and LPF cutoff (e.g., [0 dB SNR, 4 kHz], [10 dB SNR, 1 kHz]; see Table 1). There were 56 items (7 VOT × 8 F0) for the /b/–/p/ blocks and 64 items (8 VOT × 8 F0) for the /d/–/t/ blocks; stimuli within each block were presented in random order. Each block was presented five times, which resulted in well-defined psychometric functions along the stimulus parameters. Each participant began with at least one block of the optimal (quiet, no LPF) condition before hearing any masked–LPF conditions. Because of the large number of conditions, participants generally did not volunteer enough time to complete five blocks of each condition; condition selection and ordering was constrained within participants' scheduling availability. Listeners heard a variable subset of the conditions (that were not necessarily limited to one contrast), depending on their scheduling availability; most completed between 5 and 10 different conditions. The final data set included at least 10 listeners for each condition, for a total of over 800 tested blocks. Each repetition of a single block took roughly 3–5 min.

Before performing the group analyses, we used Sigmaplot 9.01 (Systat, 2004) to initially fit individual listeners' response functions to a simple logistic model. When listeners' data for a particular condition did not reach satisfactory convergence to the model, one or two more repetitions of that condition were conducted to smooth the function to allow a better fit. This was done for 5 of 20 listeners in some of the more challenging conditions (i.e., those where signal degradations were harsh enough to inhibit consistent use of the cues).

Analysis

Listeners' responses were fit using generalized linear (logistic) mixed-effects models (GLMMs). This was done in the R software interface (R Development Core Team, 2010), using the lme4 package (Bates & Maechler, 2010). A random effect of participant was used, and the fixed effects were the stimulus factors described above (Consonant Place, VOT, F0, LPF, SNR). The binomial family link function was used. The models included each main factor and all possible interactions (the four-way interactions were significant, necessitating the inclusion of all nested factors and interactions). The goal of these models was similar to that used by Peng, Lu, and Chatterjee (2009) and by Winn, Chatterjee, and Idsardi (2012); the models tested whether the coefficient of the resulting parameter estimate for an acoustic cue was different from zero and, crucially, whether the coefficient was different across conditions of LPF and SNR levels. Changes in this coefficient represent changes in the log odds of voiceless perceptions resulting from the condition or cue level change. Following previous studies (Morrison & Kondaurova, 2009; Winn et al., 2012), we interpreted the factor estimate from the GLMM as an indication of the strength of the factor (i.e., a higher estimate indicates higher perceptual weight).

Results

Averaged group responses to the continua of VOT and F0 are displayed in the tiled grids in Figures 3, 4, and 5. Grayscale intensity represents the proportion of voiced responses. Listeners who solely used VOT to distinguish these contrasts would yield grids with rows of different grayscale intensity, whereas listeners who solely used F0 would yield grids with columns of different grayscale intensity; each grid reflects the use of both cues in varying proportions. Sharper grayscale contrasts in successive rows and columns are akin to steeper psychometric functions.

Three GLMMs were used to describe listeners' responses for each of the three planned comparisons. Model terms along with the intercept and parameter estimates are given in Tables 2, 3, and 4. Simplified parameter estimates

Figure 3. Tiled grid showing the proportion of voiced or voiceless responses at each level of the voice onset time (VOT) and F0 continua for the first comparison (initial exploration of filtering and noise).

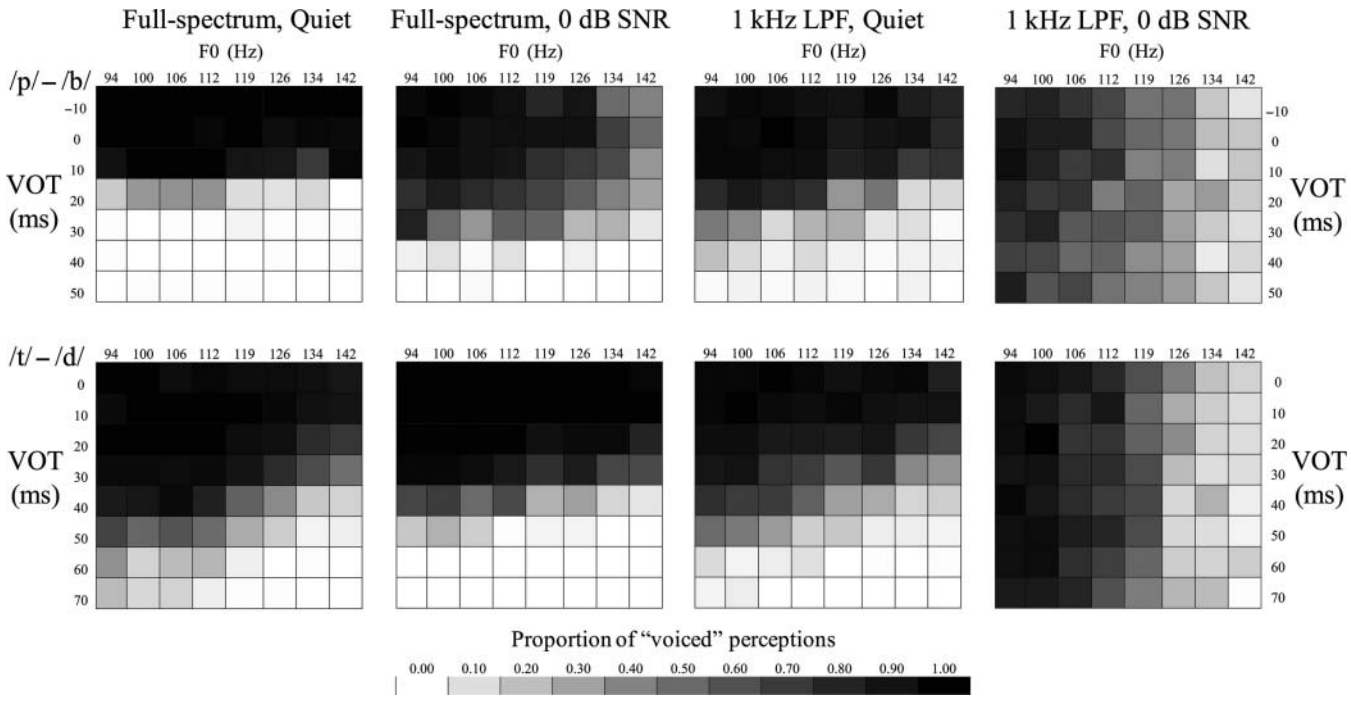


Figure 4. Tiled grid showing the proportion of voiced or voiceless responses at each level of the VOT and F0 continua for the second comparison (exploration of LPF effects in 0 dB SNR noise).

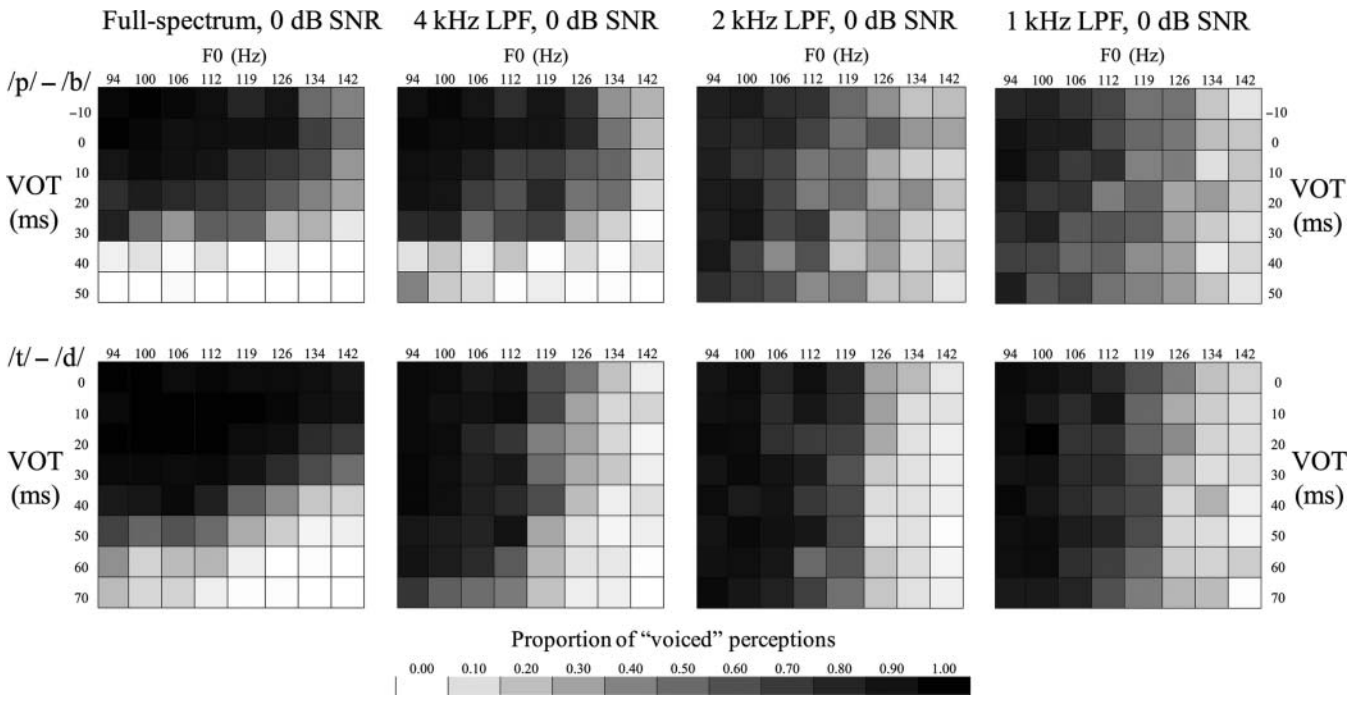
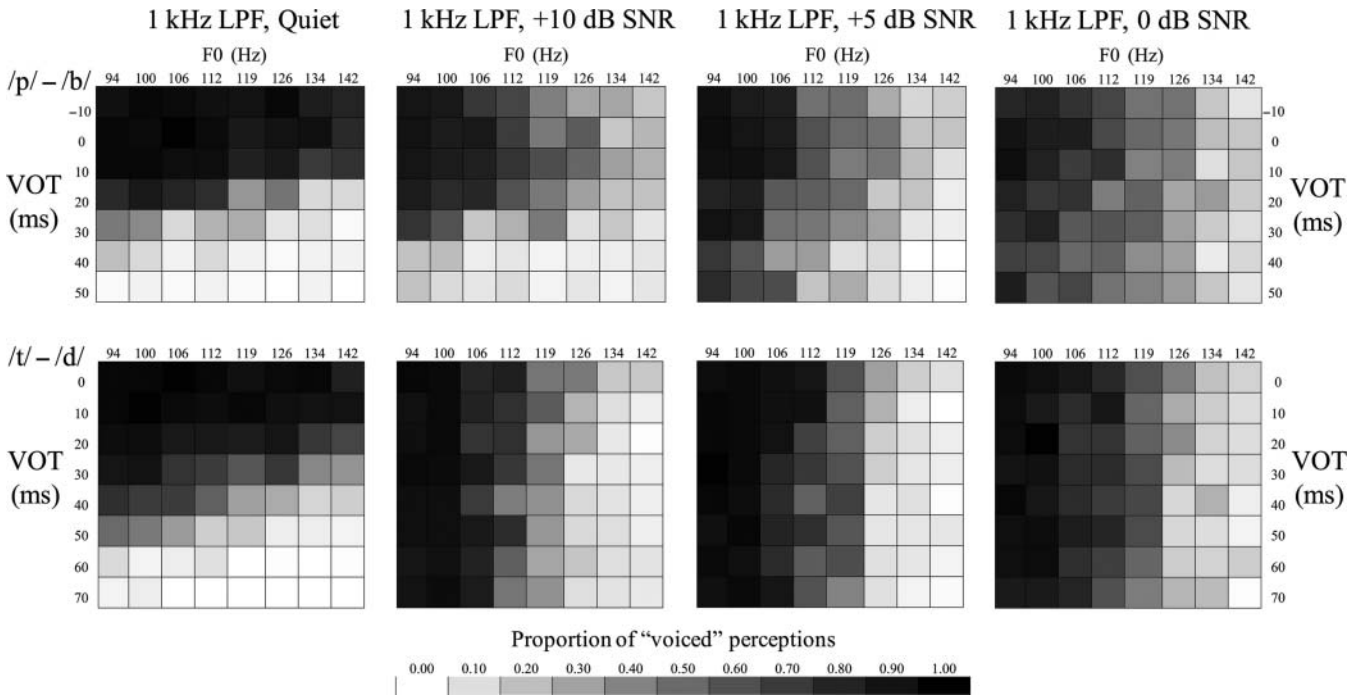


Figure 5. Tiled grid showing the proportion of voiced or voiceless responses at each level of the VOT and F0 continua for the third comparison (exploration of SNR effects with 1 kHz LPF).



from the GLMMs are summarized in the supplemental table of coefficients (see online supplemental materials) and are illustrated in Figure 6.

In the first comparison (initial exploration of LPF and noise; Figure 3; Table 2), there were significant main effects of VOT, F0, SNR, and LPF (all *p*s < .001). Alveolar consonants were less likely to be heard as voiced (consistent with acoustics of these consonants; Lisker & Abramson, 1964). F0 was a stronger cue for the /d/–/t/ contrast ($p < .001$). The effect of VOT was significantly reduced when the signal was either low-passed ($p < .001$ for both contrasts) or in noise ($p < .001$ for both contrasts) and significantly reduced further in the presence of both filtering and noise ($p < .001$). The effect of F0 was significantly stronger for both contrasts when the signals were masked by noise ($p < .001$); when signals were low-passed, F0 became stronger only for the /b/–/p/ contrast. When the signal was both low-passed and in noise, the use of F0 significantly increased for both contrasts compared with either degradation alone ($p < .001$).

In the second comparison (effects of LPF in 0 dB SNR noise; Figure 4; Table 3), the effect of VOT significantly decreased for both contrasts with each successive reduction in LPF cutoff ($p < .001$), with the exception of the 1 kHz condition, which was not significantly different from the 2 kHz condition. The use of VOT for the /d/–/t/ contrast was greater than that for the /b/–/p/ contrast when the full spectrum was available, but in any LPF condition, there was a significant advantage for the /b/–/p/ contrast for the use of VOT (all *p*s < .001). The effect of F0 was significantly stronger for the /d/–/t/ contrast in all low-pass filtered conditions ($p < .001$ for 4 kHz and 2 kHz; $p < .05$ for 1 kHz). The effect of F0 for the /b/–/p/ contrast did not significantly increase in any LPF condition.

In the third comparison (effects of SNR with a 1 kHz LPF; Figure 5; Table 4), the use of VOT significantly decreased with each successive reduction in SNR for both contrasts (all *p*s < .001). The use of F0 increased for the /b/–/p/ contrast for 5 and 0 dB SNR (both *p*s < .001), while it increased for the /d/–/t/ contrast in all conditions with noise (all *p*s < .001).

A question that remains from the results presented thus far is whether a listener can achieve the same or similar level of accuracy for voicing recognition via F0 as with VOT. Because the analyses presented thus far do not speak to correctness per se, a final analysis was conducted to evaluate the identification of stimuli where both the VOT and F0 cues cooperated at typical "voiceless" or "voiced" values. Figure 7 illustrates performance levels for these end-point stimuli by listeners in all conditions. Voicing was correctly identified with 80% accuracy or greater in all conditions except for /b/ in the 1 kHz LPF with 0 dB SNR noise condition. The cue estimate for VOT showed a significant positive correlation with accuracy for end-point accuracy ($r = .76$ for /b/–/p/, $r = .74$ for /d/–/t/; $p < .05$ for each). The cue estimate for F0 showed a significant negative correlation with end-point accuracy for the /d/–/t/ contrast ($r = –.68$, $p < .05$) but did not show a significant correlation with accuracy for the /b/–/p/ contrast. Thus, although

Table 2. Generalized linear (logistic) mixed-effects model (GLMM) results for the first comparison (initial exploration of bandwidth and noise).

Effect	Estimate	*z*
Intercept	1.109	11.15***
/b/–/p/		
LPF	–1.184	–10.07***
0 dB SNR	–1.338	–11.31***
LPF:0 dB SNR	1.476	10.40***
/d/–/t/		
POA	–1.788	–13.33***
POA:LPF	1.756	10.95***
POA:0 dB SNR	0.954	5.76***
POA:LPF:0 dB SNR	–1.262	–6.38***
VOT	0.302	21.94***
/b/–/p/		
VOT:LPF	–0.169	–11.62***
VOT:0 dB SNR	–0.191	–13.22***
VOT:LPF:0 dB SNR	0.070	4.55***
/d/–/t/		
VOT:POA	–0.067	–3.98***
VOT:POA:LPF	0.054	3.02**
VOT:POA:0 dB SNR	0.081	4.53***
VOT:POA:LPF:0 dB SNR	–0.072	–3.75***
F0	0.027	4.18***
/b/–/p/		
F0:LPF	0.028	3.69***
F0:0 dB SNR	0.037	4.85***
F0:LPF:0 dB SNR	–0.019	–2.02*
/d/–/t/		
F0:POA	0.044	5.19***
F0:POA:LPF	–0.039	–3.81***
F0:POA:0 dB SNR	–0.023	–2.16*
F0:POA:LPF:0 dB SNR	0.058	4.49***
VOT:F0 interaction	0.000	–0.50[a]
/b/–/p/		
VOT:F0:LPF	0.000	–0.06[a]
VOT:F0:0 dB SNR	–0.002	–2.15*
VOT:F0:LPF:0 dB SNR	0.002	2.26*
/d/–/t/		
VOT:F0:POA	0.000	–0.33[a]
VOT:F0:POA:LPF	0.001	0.84[a]
VOT:F0:POA:0 dB SNR	0.002	2.16*
VOT:F0:POA:LPF:0 dB SNR	–0.003	–2.24*

Note. The default (intercept) condition was full spectrum in quiet. POA refers to place of articulation; corresponding numbers reflect the difference between factor estimates for the /b/–/p/ and /d/–/t/ series.

*$p < .05$. **$p < .01$. ***$p < .001$.

[a]Not significant.

listeners were able to mostly compensate for the degraded VOT cue via the F0 cue, they performed better when relying on VOT.

Discussion

In this experiment, listeners perceived voicing contrasts in signals that were low-pass filtered and/or masked by SSN. These two signal degradations generally yielded a decline in the use of VOT that was accompanied by an increase in the use of F0. Thus, as acoustic degradations compromised the availability of the VOT cue, listeners did not simply guess at the words—they recruited appropriate information from a different acoustic cue. The one condition common to all three comparisons (0 dB SNR, 1 kHz LPF; the most challenging condition) was theoretically the most challenging in the experiment. It was distinct because while the use of VOT decreased, the use of F0 was less than the use of F0 for some other, more favorable masked and/or filtered conditions. For the most challenging condition, it appears that the acoustic signal is so degraded that even F0 is difficult to perceive in a useful way.

Speech-shaped masking noise and low-pass filtering had disproportionate effects on the /b/–/p/ and /d/–/t/ contrasts. To the extent that voicing perception can be framed

Table 3. GLMM results for the second comparison (effects of bandwidth in 0 dB SNR noise).

Effect	Estimate	*z*
Intercept (full spectrum)	–0.503	–3.99***
/b/–/p/		
4 kHz	0.267	3.04**
2 kHz	0.553	6.57***
1 kHz	0.562	6.59***
POA	–0.708	–7.01***
/d/–/t/		
POA:4 kHz	0.774	6.20***
POA:2 kHz	0.419	3.45***
POA:1 kHz	0.415	3.42***
VOT	0.114	25.08***
/b/–/p/		
VOT:4 kHz	–0.040	–7.40***
VOT:2 kHz	–0.100	–19.92***
VOT:1 kHz	–0.102	–20.15***
VOT:POA	0.017	2.68**
/d/–/t/		
VOT:POA:4 kHz	–0.061	–8.46***
VOT:POA:2 kHz	–0.021	–3.03**
VOT:POA:1 kHz	–0.020	–2.93**
F0	0.065	15.59***
/b/–/p/		
F0:4 kHz	0.009	1.64[a]
F0:2 kHz	0.005	1.06[a]
F0:1 kHz	0.010	1.82†
F0:POA	0.023	3.75***
/d/–/t/		
F0:POA:4 kHz	0.034	4.08***
F0:POA:2 kHz	0.039	4.80***
F0:POA:1 kHz	0.020	2.47*
VOT:F0 interaction	–0.002	–9.30***
/b/–/p/		
VOT:F0:4 kHz	0.001	3.10**
VOT:F0:2 kHz	0.003	8.46***
VOT:F0:1 kHz	0.002	7.00***
VOT:F0:POA	0.002	5.81***
/d/–/t/		
VOT:F0:POA:4 kHz	–0.001	–2.38*
VOT:F0:POA:2 kHz	–0.002	–4.32***
VOT:F0:POA:1 kHz	–0.002	–4.14***

Note. The default (intercept) condition was full spectrum in 0 dB SNR noise.

*$p < .05$. **$p < .01$. ***$p < .001$. †$p < .1$.

[a]Not significant.

Table 4. GLMM results for the third comparison (effects of SNR using 1 kHz LPF).

Effect	Estimate	*z*
Intercept	–0.075	–1.16[a]
/b/–/p/		
10 dB SNR	0.604	7.52***
5 dB SNR	0.442	5.49***
0 dB SNR	0.138	1.76[†]
/d/–/t/		
POA	–0.032	–0.36[a]
POA:10 dB SNR	–0.451	–4.06***
POA:5 dB SNR	–0.525	–4.61***
POA:0 dB SNR	–0.309	–2.84**
VOT	0.133	27.96***
/b/–/p/		
VOT:10 dB SNR	–0.083	–15.40***
VOT:5 dB SNR	–0.105	–19.61***
VOT:0 dB SNR	–0.121	–23.14***
/d/–/t/		
VOT:POA	–0.013	–2.07*
VOT:POA:10 dB SNR	–0.025	–3.54***
VOT:POA:5 dB SNR	–0.002	–0.33[a]
VOT:POA:0 dB SNR	0.009	1.36[a]
F0	0.055	13.24***
/b/–/p/		
F0:10 dB SNR	0.006	1.23[a]
F0:5 dB SNR	0.041	7.43***
F0:0 dB SNR	0.018	3.51***
/d/–/t/		
F0:POA	0.005	0.93[a]
F0:POA:10 dB SNR	0.066	8.55***
F0:POA:5 dB SNR	0.051	6.14***
F0:POA:0 dB SNR	0.035	4.71***
VOT:F0	0.000	–1.70[†]
/b/–/p/		
VOT:F0:10 dB SNR	–0.001	–2.12*
VOT:F0:5 dB SNR	0.001	1.72[†]
VOT:F0:0 dB SNR	0.000	0.69[a]
/d/–/t/		
VOT:F0:POA	0.001	1.59[a]
VOT:F0:POA:10 dB SNR	0.001	1.89[†]
VOT:F0:POA:5 dB SNR	–0.001	–1.51[a]
VOT:F0:POA:0 dB SNR	0.000	–0.63[a]

Note. The default (intercept) condition was 1 kHz low-pass filtered in 0 dB SNR noise.

*$p < .05$. **$p < .01$. ***$p < .001$. [†]$p < .1$.

[a]Not significant.

as detection of aspiration noise, these disproportionate effects can be explained by the acoustics of labial and alveolar stop sounds in this study. The use of cues for labial sounds was influenced more by the level of masking noise, presumably because the spectrum of SSN competes more directly with the /p/ burst and aspiration (see Figure 1). Low-pass filtering the sounds had little effect on the use of cues for labial sounds, presumably because the spectrum of the /p/ aspiration contains sufficient low-frequency components that remain after filtering. Conversely, the use of VOT for the /d/–/t/ contrast was heavily reduced by low-pass filtering. Consistent with earlier literature on the acoustics and perception of /t/ (Régnier & Allen, 2008), the audibility of energy above 4 kHz is essential for the perception of /t/ aspiration; all conditions that used a LPF of 4 kHz or lower saw dramatic reductions in VOT use along with increased use of F0 for alveolar sounds, even for modest SNRs. Table 5 shows that the SNR advantage of /p/ compared with /t/ is evident in the lower frequency regions (i.e., from 0 to 4 kHz), whereas in higher frequency regions (i.e., between 4 and 8 kHz), /t/ has an advantage (all relative to the masking noise used in this experiment). The pronounced asymmetry in energy at the upper and lower regions for labial and alveolar aspiration noise helps to explain differences in cue weighting across these two places of articulation and accords with previous reports of individual consonant advantages in SSN (Phatak & Allen, 2007).

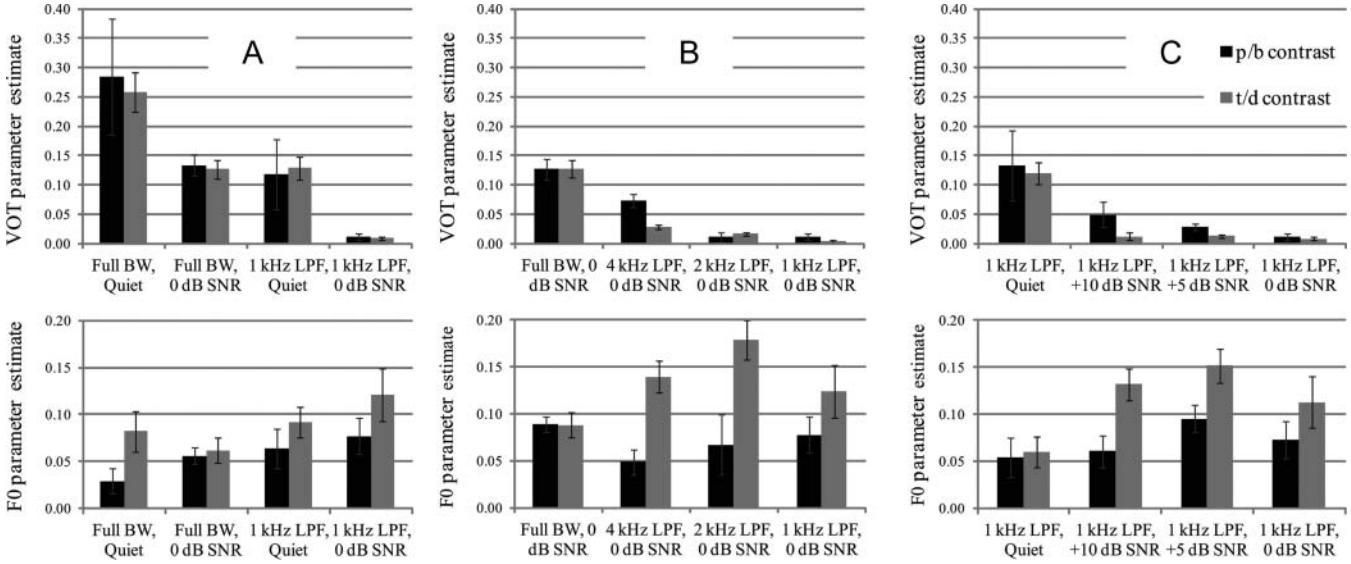
Other types of noise may have different effects than those shown in this study. With the entire spectrum available and audible, the SNR of stimuli in white noise may affect the /t/–/d/ contrast more heavily than the /p/–/b/ contrast. Amplitude-modulated noise or competing speech might momentarily provide a favorable SNR for these cues via a dip in amplitude concurrent with the onset of the target words; there is no reason to think that modulation would differentially affect /p/ versus /t/. It should be noted that listeners with hearing loss are less able to capitalize on short-term valleys of a masker (Carhart & Tillman, 1970; Festen & Plomp, 1990) and, thus, may not recover segmental information in such conditions.

Although listeners were able to reliably identify the voicing of the most natural (i.e., "end-point") stimuli (Figure 7), it is not yet known whether the signal degradation requires listeners to use greater cognitive resources to perceive voicing. Given listeners' general tendency to rely on VOT rather than on F0 (Abramson & Lisker, 1985), conditions driven by F0 could have been more difficult than those driven by VOT, despite similar accuracy scores. Accuracy for end-point stimuli in this experiment was significantly correlated to the strength (i.e., factor estimate) of VOT. Thus, scores in phoneme identification tasks may tell only part of the story; similar scores could have arisen because of different perceptual strategies. Pupillometry during speech perception tasks suggests that extra effort is required to maintain equal intelligibility of speech in the presence of different types of maskers (Koelewijn, Zekveld, Festen, & Kramer, 2012) or if listeners have hearing loss (Kramer, Kapteyn, Festen, & Kuik, 1997). It is not yet known whether alternative phonetic cue-weighting strategies would elicit similar signs of increased listening effort.

In this experiment, the role of F1 was minimized via the use of the high vowel /i/. Jiang, Chen, and Alwan (2006) showed that F1 can play a role in the perception of voicing in noise for non-high vowels, confirming a prediction by Hillenbrand, Ingrisano, Smith, and Flege (1984). It is not yet known whether F1 or F0 is more dominant in compensating for degradations of VOT in masked and/or filtered conditions.

The motivation for this experiment was to model potential listening strategies that could arise when a person experiences HI. Because HI is more complex than a simple LPF, the results of this study should be interpreted with

Figure 6. Parameter estimates (coefficients) for the logistic models for the three comparisons: (A) initial exploration of LPF and noise effects; (B) effects of LPF with 0 dB SNR; and (C) effects of SNR with 1 kHz LPF. Greater bar height indicates greater influence of the cue in the model. Black and gray bars represent estimates for the /b/–/p/ and /d/–/t/ contrasts, respectively. The upper and lower panels illustrate estimates for the VOT and F0 cues, respectively. Error bars reflect ±1 standard error of the mean of the coefficients across participants. BW = bandwidth.



caution. There are suprathreshold deficits in the spectral and temporal domains that might limit a listener's ability to utilize either of the acoustic cues explored in this study. These deficits are frequently attributed to poor frequency resolution and/or temporal fine structure coding (Bernstein & Oxenham, 2006; Lorenzi, Gilbert, Carn, Garnier, & Moore, 2006). Turner and Brus (2001) showed that although the amplification of low-frequency energy (including F0) provided benefit for listeners with HI, this benefit was smaller than that observed for those with normal hearing. Grant (1987) suggested that listeners with HI are not able to detect subtle F0 contrasts and would therefore benefit from F0 contours only if they were exaggerated by roughly 1.5–6 times those observed in natural speech. Thus, it remains unclear whether listeners with cochlear hearing loss can capitalize on F0 cues in noise to the same extent as the participants in this study. Further difficulties might be experienced by older listeners, who have been shown to experience deficiencies in auditory temporal

Figure 7. Mean accuracy in identification of consonant voicing at continuum end points by different listener groups. Voiceless and voiced items for this analysis were limited to those where both the VOT and F0 cues cooperated appropriately (i.e., long VOT and high F0 or short VOT and low F0) at continuum end points to signal the same feature.

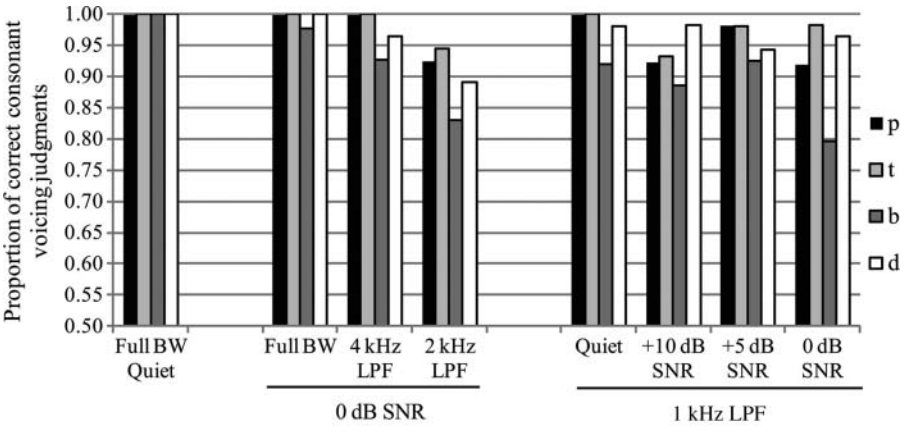


Table 5. SNRs for /p/and /t/aspiration relative to masking noise in the 0 dB SNR condition, in different frequency bands.

Low frequency (Hz)	High frequency (Hz)	/p/ SNR (dB)	/t/ SNR (dB)	Relative advantage	
				dB	Stronger consonant
0	1000	–10.2	–20.9	10.7	/p/
0	2000	–14.6	–21.0	6.4	/p/
0	4000	–9.6	–13.9	4.3	/p/
0	8000	–6.0	–0.8	5.2	/t/
4000	8000	–2.4	12.3	14.7	/t/

processing in basic psychophysical tasks (Gordon-Salant & Fitzgibbons, 1993, 1999) and tasks involving perception of temporal phonetic cues in isolated words (Gordon-Salant, Yeni-Komshian, Fitzgibbons, & Barrett, 2006) and in sentence contexts (Gordon-Salant, Yeni-Komshian, & Fitzgibbons, 2008).

The use of F0 as a segmental cue in this study may partly explain the benefit of a natural F0 contour of sentences presented in noise (Binns & Culling, 2007; Laures & Weismer, 1999; Miller, Schlauch, & Watson, 2010). It is not known whether the segmental use of F0 cues in this study would generalize to longer utterance contexts, where F0 contrast is likely constrained by other sources of variability (e.g., intonation) and phonetic reduction. It should be noted that the current experiment used a two-alternative forced-choice task that assessed only voicing perception in single words; it is likely that these stimuli would be confused with other consonants (but perhaps not with consonants of different voicing) if a larger response set were used. Conversely, the influence of sentence context and other top-down factors may compensate for the added difficulty of an open response set (McClelland, Mirman, & Holt, 2006).

Emergent models of phonetic perception and categorization increasingly acknowledge the integration of multiple covarying acoustic cues in the speech signal (McMurray & Jongman, 2011; McMurray, Tanenhaus, & Aslin, 2002; Toscano & McMurray, 2010). The presence of multiple cues for voicing can at least partly explain why voicing is such a robust feature in phoneme identification tasks in adverse listening conditions like low-pass filtering and masking noise. Listeners are capable, without any explicit instructions, of increasing reliance on residual cues in a speech signal when otherwise stronger cues have been degraded.

Acknowledgments

This work was supported by National Institutes of Health (NIH) Grants R01 DC 004786 (awarded to Monita Chatterjee) and 7R01DC005660-07 (awarded to David Poeppel and William J. Idsardi). Matthew B. Winn was supported by the University of Maryland Center for Comparative and Evolutionary Biology of Hearing Training Grant (NIH Grant T32 DC000046-17; principal investigator Arthur N. Popper). We are grateful to Ewan Dunbar for his assistance with the statistical analysis.

References

Abramson, A., & Lisker, L. (1970). Discriminability along the voicing continuum: Cross language tests. In Xxxxx Xxxxxx (Ed.), *Proceedings of the 6th International Congress of Phonetic Sciences, Prague* (pp. 569–573). Prague, Czech Republic: Academia.

Abramson, A., & Lisker, L. (1985). Relative power of cues: F0 shift versus voice timing. In V. A. Fromkin (Ed.), *Phonetic linguistics: Essays in honor of Peter Ladefoged* (pp. 25–33). New York, NY: Academic Press.

American National Standards Institute. (2010). *Specifications for audiometers* (ANSI S3.6-2010). New York, NY: Author.

Andruski, J., Blumstein, S., & Burton, M. (1994). The effect of subphonetic differences on lexical access. *Cognition, 52,* 173–187.

Bates, D., & Maechler, M. (2010). lme4: Linear mixed-effects models using S4 classes (R package Version 0.999375-37) [Software package]. Available from http://CRAN.R-project.org/package=lme4

Bernstein, J., & Oxenham, A. (2006). The relationship between frequency selectivity and pitch discrimination: Sensorineural hearing loss. *The Journal of the Acoustical Society of America, 120,* 3929–3945.

Bilger, R., & Wang, M. (1976). Consonant confusions in patients with sensorineural hearing loss. *Journal of Speech and Hearing Research, 19,* 738–748.

Binns, C., & Culling, J. (2007). The role of fundamental frequency contours in the perception of speech against interfering speech. *The Journal of the Acoustical Society of America, 122,* 1765–1776.

Boersma, P., & Weenink, D. (2011). Praat: Doing phonetics by computer (Version 5.1.23) [Computer software]. Available from www.praat.org

Brokx, J., & Nooteboom, S. (1982). Intonation and the perceptual separation of simultaneous voices. *Journal of Phonetics, 10,* 23–36.

Carhart, R., & Tillman, T. W. (1970). Interaction of competing speech signals with hearing losses. *Archives of Otolaryngology, 91,* 273–279.

Cutler, A., & Foss, D. (1977). On the role of sentence stress in sentence processing. *Language and Speech, 20,* 1–10.

Dubno, J., & Levitt, H. (1981). Predicting consonant confusions from acoustic analysis. *The Journal of the Acoustical Society of America, 69,* 249–261.

Eddins, D., & Green, D. (1995). Temporal integration and temporal resolution. In B. Moore (Ed.), *Hearing* (pp. 207–242). San Diego, CA: Academic Press.

Eddins, D., Hall, J., & Grose, J. (1992). The detection of temporal gaps as a function of frequency region and absolute noise bandwidth. *The Journal of the Acoustical Society of America, 91,* 1069–1077.

Festen, J., & Plomp, R. (1990). Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing. *The Journal of the Acoustical Society of America, 88,* 1725–1736.

Fogerty, D., & Humes, L. (2012). The role of vowel and consonant fundamental frequency, envelope, and temporal fine structure cues to the intelligibility of words and sentences. *The Journal of the Acoustical Society of America, 131,* 1490–1501.

Friesen, L., Shannon, R., Başkent, D., & Wang, X. (2001). Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants. *The Journal of the Acoustical Society of America, 110,* 1150–1163.

Fu, Q.-J. (2006). Internet-Based Computer-Assisted Speech Training (iCAST) by TigerSpeech Technology (Version 5.04.02)

[Computer program]. Available from www.tigerspeech.com/tst_icast.html

Gordon-Salant, S., & Fitzgibbons, P. (1993). Temporal factors and speech recognition performance in young and elderly listeners. *Journal of Speech and Hearing Research, 36,* 1276–1285.

Gordon-Salant, S., & Fitzgibbons, P. (1999). Profile of auditory temporal processing in older listeners. *Journal of Speech, Language, and Hearing Research, 42,* 300–311.

Gordon-Salant, S., Yeni-Komshian, G., & Fitzgibbons, P. (2008). The role of temporal cues in word identification by younger and older adults: Effects of sentence context. *The Journal of the Acoustical Society of America, 124,* 3249–3260.

Gordon-Salant, S., Yeni-Komshian, G., Fitzgibbons, P., & Barrett, J. (2006). Age-related differences in identification and discrimination of temporal cues in speech segments. *The Journal of the Acoustical Society of America, 119,* 2455–2466.

Grant, K. (1987). Identification of intonation contours by normally hearing and profoundly hearing-impaired listeners. *The Journal of the Acoustical Society of America, 82,* 1172–1178.

Grose, J. (1991). Gap detection in multiple narrow bands of noise as a function of spectral configuration. *The Journal of the Acoustical Society of America, 90,* 3061–3068.

Haggard, M., Ambler, A., & Callow, M. (1970). Pitch as a voicing cue. *The Journal of the Acoustical Society of America, 47,* 613–617.

Hillenbrand, J., Ingrisano, D., Smith, B., & Flege, J. (1984). Perception of the voiced–voiceless contrast in syllable-final stops. *The Journal of the Acoustical Society of America, 76,* 18–26.

Hombert, J. (1975). *Towards a theory of tonogenesis: An empirical, physiologically and perceptually-based account of the development of tonal contrasts in language* (Unpublished doctoral dissertation). University of California, Berkeley.

House, A., & Fairbanks, G. (1953). The influence of consonant environment upon the secondary acoustical characteristics of vowels. *The Journal of the Acoustical Society of America, 25,* 105–113.

Jiang, J., Chen, M., & Alwan, A. (2006). On the perception of voicing in syllable-initial plosives in noise. *The Journal of the Acoustical Society of America, 119,* 1092–1105.

Koelewijn, T., Zekveld, A., Festen, J., & Kramer, S. (2012). Pupil dilation uncovers extra listening effort in the presence of a single-talker masker. *Ear and Hearing, 33,* 291–300.

Kramer, S., Kapteyn, D., Festen, J., & Kuik, D. (1997). Assessing aspects of auditory handicap by means of pupil dilatation. *Audiology, 36,* 155–164.

Ladefoged, P., & Maddieson, I. (1996). *The sounds of the world's languages*. Oxford, England: Blackwell.

Laures, J., & Weismer, G. (1999). The effects of a flattened fundamental frequency on intelligibility at the sentence level. *Journal of Speech, Language, and Hearing Research, 42,* 1148–1156.

Lisker, L. (1975). Is it VOT or a first-formant transition detector? *The Journal of the Acoustical Society of America, 57,* 1547–1551.

Lisker, L., & Abramson, A. (1964). A cross-language study of voicing in stops: Acoustical measurements. *Word, 20,* 384–422.

Lorenzi, C., Gilbert, G., Carn, H., Garnier, S., & Moore, B. (2006). Speech perception problems of the hearing impaired reflect inability to use temporal fine structure. *Proceedings of the National Academy of Sciences of the United States of America, 103,* 18866–18869.

McAdams, S. (1989). Segregation of concurrent sounds: I. Effects of frequency modulation coherence. *The Journal of the Acoustical Society of America, 86,* 2148–2159.

McClelland, J., Mirman, D., & Holt, L. (2006). Are there interactive processes in speech perception? *Trends in Cognitive Science, 10,* 363–369.

McMurray, B., & Jongman, A. (2011). What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological Review, 118,* 219–246.

McMurray, B., Tanenhaus, M., & Aslin, R. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition, 86,* B33–B42.

Miller, G., & Nicely, P. (1955). An analysis of perceptual confusions among some English consonants. *The Journal of the Acoustical Society of America, 27,* 338–352.

Miller, S., Schlauch, R., & Watson, P. (2010). The effects of fundamental frequency contour manipulations on speech intelligibility in background noise. *The Journal of the Acoustical Society of America, 128,* 435–443.

Morrison, G., & Kondaurova, M. (2009). Analysis of categorical response data: Use logistic regression rather than endpoint-difference scores or discriminant analysis. *The Journal of the Acoustical Society of America, 126,* 2159–2162.

Ohde, R. (1984). Fundamental frequency as an acoustic correlate of stop consonant voicing. *The Journal of the Acoustical Society of America, 75,* 224–230.

Peng, S.-C., Lu, N., & Chatterjee, M. (2009). Effects of cooperating and conflicting cues on speech intonation recognition by cochlear implant users and normal hearing listeners. *Audiology and Neurotology, 14,* 327–337.

Phatak, S., & Allen, J. (2007). Consonant and vowel confusions in speech-weighted noise. *The Journal of the Acoustical Society of America, 121,* 2312–2316.

Phatak, S., Lovitt, A., & Allen, J. (2008). Consonant confusions in white noise. *The Journal of the Acoustical Society of America, 124,* 1220–1233.

R Development Core Team. (2010). R: A language and environment for statistical computing [Computer software]. Available from www.R-project.org

Régnier, M., & Allen, J. (2008). A method to identify noise-robust perceptual features: Application for consonant /t/. *The Journal of the Acoustical Society of America, 123,* 2801–2814.

Shannon, R., Zeng, F.-G., Kamath, V., Wygonski, J., & Ekelid, M. (1995, October 13). Speech recognition with primarily temporal cues. *Science, 270,* 303–304.

Systat. (2004). SigmaPlot (Version 9.01) [Computer program]. San Jose, CA: Systat.

Toscano, J., & McMurray, B. (2010). Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive Science, 34,* 434–464.

Turner, C., & Brus, S. (2001). Providing low- and mid-frequency speech information to listeners with sensorineural hearing loss. *The Journal of the Acoustical Society of America, 109,* 2999–3006.

Wang, M., & Bilger, R. C. (1973). Consonant confusions in noise: A study of perceptual features. *The Journal of the Acoustical Society of America, 54,* 1248–1266.

Whalen, D., Abramson, A., Lisker, L., & Mody, M. (1993). F0 gives voicing information even with unambiguous voice onset times. *The Journal of the Acoustical Society of America, 93,* 2152–2159.

Winn, M., Chatterjee, M., & Idsardi, W. (2012). The use of acoustic cues for phonetic identification: Effects of spectral degradation and electric hearing. *The Journal of the Acoustical Society of America, 131,* 1465–1479.

Xu, L., Thompson, K., & Pfingst, B. (2005). Relative contributions of spectral and temporal cues for phoneme recognition. *The Journal of the Acoustical Society of America, 117,* 3255–3267.