

# Perception of Place of Articulation for Plosives and Fricatives in Noise

Abeer Alwan\*, Jintao Jiang<sup>1,\*</sup>, Willa Chen

*Department of Electrical Engineering, University of California, Los Angeles, California 90095, USA*

---

## Abstract

This study aims at uncovering perceptually-relevant acoustic cues for the labial versus alveolar place of articulation distinction in syllable-initial plosives  $\{/b/,/d/,/p/,/t/\}$  and fricatives  $\{/f/,/s/,/v/,/z/\}$  in noise. Speech materials consisted of naturally-spoken consonant-vowel (CV) syllables from four talkers where the vowel was one of  $\{/a/,/i/,/u/\}$ . Acoustic analyses using logistic regression show that formant frequency measurements, relative spectral amplitude measurements, and burst/noise durations are generally reliable cues for labial/alveolar classification. In a subsequent perceptual experiment, each pair of syllables with the labial/alveolar distinction (e.g., /ba,da/) was presented to listeners in various levels of signal-to-noise-ratio (SNR) in a 2-AFC task. A threshold SNR was obtained for each syllable pair using sigmoid fitting of the percent correct scores. Results show that the perception of the labial/alveolar distinction in noise depends on the manner of articulation, the vowel context, and interaction between voicing and manner of articulation. Correlation analyses of the acoustic measurements and threshold SNRs show that formant frequency measurements (such as F1 and F2 onset frequencies and F2 and F3 frequency changes) become increasingly important for the perception of labial/alveolar distinctions as the SNR degrades.

*Keywords:* speech perception, place of articulation, plosives, fricatives, noise, psychoacoustics

---

Portions of this work were presented at ICSLP 2000 and ICPhS 2003.

\*Corresponding author

*Email addresses:* [alwan@ee.ucla.edu](mailto:alwan@ee.ucla.edu) (Abeer Alwan), [jjt@ee.ucla.edu](mailto:jjt@ee.ucla.edu) (Jintao Jiang)

<sup>1</sup>Now with the George Washington University, Washington, DC

## 1. Introduction

Research on speech perception and human auditory processes, particularly in the presence of background noise, helps to improve and calibrate such practical applications as noise-robust automatic speech recognition systems (e.g., Hermansky, 1990; Strope and Alwan, 1997) and aids for the hearing impaired (e.g., Shannon et al., 1995). The present study examines the contributions of various acoustic characteristics to the perceptual distinction between labial and alveolar places of articulation in syllable-initial plosive and fricative consonant-vowel (CV) syllables in quiet conditions and in the presence of additive white Gaussian noise.

The focus is on labial/alveolar syllable pairs that differ in manner of articulation (plosives  $\{/b,d/,/p,t/\}$  versus fricatives  $\{/v,z/,/f,s/\}$ ) and voicing (voiced  $\{/b,d/,/v,z/\}$  versus voiceless  $\{/p,t/,/f,s/\}$ ) in the vowel contexts  $\{/a/,/i/,/u/\}$ . Plosive consonants are produced by first forming a complete closure in the vocal tract via a constriction at the place of articulation, during which there is generally no sound. The vocal tract is then opened suddenly, releasing the pressure built up behind the constriction; this is characterized acoustically by a transient source and/or a short-duration noise burst (Stevens, 1998). The period between the release and the vowel onset is called the voice onset time (VOT) during which there is silence and/or aspiration noise. In contrast, fricatives are characterized by turbulence in the region of maximum constriction in the vocal tract. The excitation source is noise for voiceless fricatives, while it is noise and a quasi-periodic source for voiced fricatives. Labials and alveolars have noise source energy concentrations at different frequency regions due to differences in the location of the maximum constriction in the vocal tract.

### 1.1. Plosive consonants

Formant frequencies have been examined extensively in acoustic studies of the place of articulation in naturally-spoken plosives (e.g., Potter et al., 1947; Fant, 1973; Kewley-Port, 1982). Fant (1973) analyzed spectrograms of six Swedish plosives in nine vowel contexts and concluded that F2 and F3 formant transitions did not sufficiently reflect place of articulation. Kewley-Port (1982) measured the F1, F2, and F3 transitions for voiced plosives ( $/b,d,g/$ ) in eight vowel contexts and found that F2 and F3 transition onset values were not sufficient to

30 cue place of articulation.

31 Other studies have focused on the characteristics of the noisy burst of the plosives (e.g.,  
32 Zue, 1976; Blumstein and Stevens, 1979; Stevens and Blumstein, 1978). Zue (1976) found  
33 that an alveolar burst had a broad shaped spectral peak in the high frequency region, while  
34 a velar burst had a compact peak in the mid-frequency region when followed by a front vowel  
35 and in the lower frequency region when followed by a back vowel. However, for labials, the  
36 study did not find consistent burst characteristics. Stevens and Blumstein (Blumstein and  
37 Stevens, 1979; Stevens and Blumstein, 1978) defined labial bursts to be “diffused falling”  
38 (widespread spectral energy with a concentration at the low to mid-frequency region), alve-  
39 olar bursts to be “diffused rising” (widespread spectral energy with a concentration at the  
40 high-frequency region), and velars to have a compact mid-frequency spectral peak.

41 More recently, researchers have suggested that the spectral amplitude of the consonant  
42 portion of a CV syllable relative to that of the vowel onset cues place of articulation (e.g.,  
43 Stevens et al., 1999; Suchato, 2004). In (Stevens et al., 1999), three relative spectral quan-  
44 tities were measured, as well as F1 and F2 frequencies. The first relative measure was the  
45 peak spectrum amplitude of the burst in the frequency range above 3500 Hz for female  
46 talkers and 3000 Hz for male talkers (Ahi) relative to the average of spectral peaks in the  
47 F2 and F3 range in the burst (A23), denoted as Ahi-A23, measuring the spectral tilt of  
48 the burst. The second quantity was the spectrum amplitude of the F1 prominence in the  
49 vowel onset (Av) relative to Ahi, denoted as Av-Ahi, measuring the burst amplitude rel-  
50 ative to the vowel amplitude. The third quantity was the difference between Av and the  
51 peak spectrum amplitude in the F2 to F3 range (pA23) in the burst, denoted as Av-pA23,  
52 measuring the mid-frequency spectral prominence. These measurements were performed on  
53 a number of syllable-initial plosives, 15 tokens each, drawn from 100 sentences spoken by  
54 two male and two female talkers. The results showed that Ahi was smaller than A23 for  
55 labials but relatively similar to A23 for alveolars; Av-Ahi measurements indicated that the  
56 labial burst in high frequencies was weaker than the alveolar burst; Av-pA23 was the best  
57 indicator of velars because velars had the most prominent mid-frequency peak; and place  
58 of articulation classification with these measurements showed the effects of talker and vowel  
59 context. Suchato (2004) found that attributes relating to the burst spectrum in relation to

60 that of the vowel were most effective for automatically classifying place of articulation, while  
61 attributes relating to formant transitions were somewhat less effective.

62 Perceptual experiments with synthetic plosives modeling a male adult voice have also  
63 been conducted to find perceptually relevant acoustic cues for the place of articulation (e.g.,  
64 Liberman et al., 1954; Delattre et al., 1955; Ohde and Stevens, 1983). Liberman et al. (1954)  
65 found that the F2 transition cued the place of articulation for plosives. Delattre et al. (1955)  
66 provided further specification of the F2 onset frequencies (720 Hz for /b/ and 1800 Hz for  
67 /d/). Hedrick and colleagues (Hedrick and Jesteadt, 1996; Hedrick et al., 1995) varied the  
68 burst amplitude in the F4-F5 region relative to the vowel onset amplitude and the F2 and  
69 F3 onset frequencies in synthetic voiceless plosive CV syllables. They showed that increasing  
70 the relative presentation level of the burst yielded more alveolar responses, that the increase  
71 in alveolar responses also co-varied with the F2 and F3 onset frequencies, and that burst  
72 amplitude relative to vowel onset amplitude in the F4-F5 region seemed to cue voiceless  
73 labial/alveolar place of articulation.

74 Locus equations have also been examined as cues for place of articulation for plosives  
75 (Sussman et al., 1991, 1993, 1995; Fruchter and Sussman, 1997). These equations are linear  
76 regressions of the F2 onset frequency on F2 vowel frequency (midvowel nucleus) for a single  
77 consonant across a range of vowels. The derived slope and intercept values have been used  
78 as predictors of place of articulation. Sussman et al. (1991) investigated locus equations in  
79 naturally-spoken voiced syllable-initial plosives. The authors found linear regression func-  
80 tions with distinct slopes and intercepts as a function of place. Fruchter and Sussman (1997)  
81 comprehensively sampled the F2 onset-F2 vowel acoustic space in the vicinity of /b,d,g/ lo-  
82 cus equations using synthetic CV stimuli. The authors found that locus equations serve as  
83 important perceptual cues for place of articulation.

84 In summary, the relative spectral amplitudes, formant transitions, and burst character-  
85 istics have been found to be important cues to place of articulation for plosive consonants  
86 in acoustic and perceptual studies.

### 87 *1.2. Fricative consonants*

88 You (1979) found that the duration of frication noise varied with place of articulation.  
89 Shadle and Mair (1996) measured spectral moments, dynamic amplitude, and spectral slope

90 in fricatives with different effort levels and vowel contexts. The authors found that spectral  
91 moments varied significantly by frequency ranges.

92 Perceptual experiments with fricatives were also conducted to find perceptually relevant  
93 acoustic cues for place of articulation. Harris (1958) and Heinz and Stevens (1961) used  
94 natural and synthetic tokens, respectively, and showed that spectral properties of frication  
95 noise were critical perceptual attributes for place of articulation. Heinz and Stevens (1961)  
96 varied the initial frequencies for the fricatives and F2 onset frequencies of the vowel and  
97 then varied the amplitude of the fricative noise relative to the vowel. The results showed  
98 that stimuli with resonance frequencies of 6500 to 8000 Hz usually produced /f/ and /θ/  
99 responses, but these responses only began to emerge when the fricative noise was -15 and -25  
100 dB relative to the vowel. Guerlekian (1981) used several synthesized stimuli with conflicting  
101 cues and found that low and high amplitude of noise relative to the vowel was perceived as  
102 /fa/ and /sa/, respectively, by both Spanish and English listeners. Jongman (1988) edited  
103 the frication noise duration in naturally-spoken CV syllables to include 20 to 70 ms in 10-ms  
104 steps as well as the entire frication noise. Perceptual results indicated that the listeners did  
105 not require the entire fricative-vowel syllable in order to correctly perceive a fricative and  
106 that perception of fricative place of articulation was much more affected by a decrease in  
107 frication duration than perception of voicing or manner of articulation.

108 Other perceptual studies suggested the importance of the amplitude of noise relative to  
109 that of the vowel onset at different frequency regions (Hedrick and Ohde, 1993; Stevens,  
110 1985). Hedrick and Ohde (1993) showed that the amplitude of the frication noise relative  
111 to the vowel in the F3-F5 region affected perception of place across different vowel contexts  
112 and frication durations. Overall, labial and alveolar fricatives seemed to have weaker and  
113 stronger noise relative to the vowel, respectively. However, Behrens and Blumstein (1988)  
114 found that perception of place of articulation of fricatives was generally not influenced by  
115 overall frication amplitude. Nevertheless, the authors suggested that the relevant property  
116 of amplitude may not be the “overall” amplitude of the frication, but rather a change in  
117 amplitude of the fricative noise relative to the vowel in a specific region.

118 Similar to that for plosive consonants, the relative spectral amplitudes have been found  
119 to be important cues to place of articulation for fricative consonants in both acoustic and

120 perceptual studies; and fricative characteristics (especially fricative duration) have also been  
121 found to be effective place cues.

### 122 1.3. *Speech perception in noise*

123 The studies discussed above were all conducted in quiet environments; however, speech is  
124 often heard in the presence of background noise. Finding perceptually salient acoustic cues  
125 in noise has important practical implications for ASR systems and hearing aids.

126 One of the earliest studies on perceptual confusions between consonants in the presence  
127 of noise was conducted by Miller and Nicely (1955). Their study used 200 naturally-spoken  
128 utterances, each consisting of one of 16 consonants followed by the vowel /a/, in varying  
129 levels of white noise and bandpass filtering conditions. The selected consonants varied along  
130 five articulatory features (voicing, nasality, affrication, duration, and place of articulation).  
131 Their work showed that place information was difficult to distinguish at SNRs less than +6  
132 dB. They also found that perception of plosives was much less robust than that of fricatives  
133 in a noisy environment. Other researchers used an information-theoretic approach to model  
134 confusion matrices of speech in noise (Soli and Arabie, 1979; Wang and Bilger, 1973). These  
135 studies attempted to find out which cues account for perceptual results in noise by analyzing  
136 confusion matrices statistically. For example, Soli and Arabie (1979) analyzed the consonant  
137 confusion data from (Miller and Nicely, 1955) and suggested (qualitatively) that consonant  
138 confusion data could be better explained by the acoustic properties of the consonants than  
139 by phonetic features.

140 The perceptual effect of noise on place of articulation cues, however, is not clear and has  
141 not been systematically investigated. Several studies have comprehensively examined physi-  
142 cal measures that could account for the changes in the perception of phonological features in  
143 the presence of background noise for a wide range of consonants (Farar et al., 1987; Hant and  
144 Alwan, 2000, 2003; Jiang et al., 2006; Hedrick and Younger, 2007; Parikh and Loizou, 2005).  
145 Farar et al. (1987) adopted an approach to quantify perceptual confusions in noise by incorpo-  
146 rating speech into psychoacoustic masking models. Using stationary broad-band noises with  
147 spectral shapes resembling certain plosives, the authors measured the discrimination thresh-  
148 olds for different plosive burst pairs as a function of burst duration. The results showed that  
149 discrimination thresholds decreased nearly 20 dB as the bursts' durations increased from 10

150 to 300 ms. However, they were unable to model the data to predict these durational effects.  
151 Alwan (1992) conducted discrimination experiments with synthetic /ba,da/ stimuli while  
152 masking their F2 trajectories with a bandpass noise. The discrimination results suggested  
153 that high-frequency cues (such as relative spectral amplitude differences in the F3 to F4  
154 region) can be used as place cues since subjects were able to identify the consonants when  
155 F2 was completely masked. The (Alwan, 1992) study only examined /ba,da/ syllables. In  
156 (Hant and Alwan, 2000, 2003; Hant, 2000), the authors developed a general, time/frequency  
157 detection model to fit the noise-masked thresholds of bandpass noises which varied in noise  
158 duration, bandwidth, and center-frequency. The model predicted well the discrimination of  
159 synthetic voiced plosive CV syllables in perceptually flat and speech-shaped noise. Their  
160 perceptual experiments and model showed that formant transitions are more perceptually  
161 salient in noise than the plosive burst. Jiang et al. (2006) conducted voicing discrimination  
162 experiments using stimuli consisting of naturally-spoken CV syllables by four talkers in vari-  
163 ous levels of additive white Gaussian noise. Their results indicate that the onset frequency of  
164 the first formant is critical in perceiving voicing in syllable-initial plosives in additive white  
165 Gaussian noise, while the VOT duration is not. Parikh and Loizou (2005) used multi-talker  
166 babble and speech-shaped noise to examine the acoustic and perceptual influence of noise on  
167 plosive consonant cues in VCV syllables. Plosive consonant recognition remained high even  
168 at -5 dB despite the disruption of burst cues due to additive noise. The authors speculated  
169 that listeners must be relying on other cues, perhaps formant transitions, to identify plosives.  
170 The (Parikh and Loizou, 2005) study employed plosive consonant identification rather than  
171 place discrimination, and there was no correlation analyses between identification scores and  
172 acoustic measurements for plosives. Hedrick and Younger (2007) investigated whether there  
173 were different perceptual weightings to cues for the /p,t/ place of articulation in speech-  
174 shaped noise versus reverberant listening conditions. The authors used synthetic /pa/ and  
175 /ta/ stimuli with varying amplitude of the spectral peak in the F4-F5 frequency region of  
176 the burst relative to the adjacent vowel peak amplitude in the same frequency region and  
177 F2/F3 formant transition onset frequencies. Results with normal-hearing listeners showed  
178 that the weightings of relative spectral amplitudes and transition cues depended on the  
179 listening condition (quiet, speech-shaped noise, or reverberation). That is, normal-hearing

180 listeners reduced their weighting of formant transitions in speech-shaped noise, while they  
181 had little difficulty using the formant transition cues in the reverberant listening condition.  
182 The (Hedrick and Younger, 2007) study only examined /pa,ta/ syllables.

183 Noise characteristics influence the perception of speech sounds. Hant and Alwan (2000)  
184 examined the perceptual confusion of synthetic plosives in noise and found that there was a 5  
185 to 10 dB drop in threshold SNRs (for which place of articulation was just perceptually salient)  
186 between speech-shaped noise and perceptually flat noise, suggesting that adult native English  
187 listeners might be using high-frequency cues to discriminate plosives in speech-shaped noise,  
188 while those cues were unavailable in perceptually flat noise. The perceptually flat noise  
189 had equal energy per Equivalent Rectangular Bandwidth of the auditory filter (Glasberg  
190 and Moore, 1990). Nittrouer et al. (2003) showed clear differences in adults' perception of  
191 consonants in white versus speech-shaped noise, while there was no difference in children's  
192 perception. Another type of noise includes background talker[s]. Simpson and Cooke (2005)  
193 demonstrated that a single competing talker or amplitude-modulated noise is a far less  
194 effective masker than multi-talker babble or speech-shaped noise for consonant identification  
195 in VCV syllables and that babble-modulated noise is a less effective masker than natural  
196 babble when there are more than two talkers in the noise. Similar results were found by  
197 Engen and Bradlow (2007) and by Lecumberri and Cooke (2006). Engen and Bradlow  
198 (2007) found that in two-talker babble, native English listeners were more adversely affected  
199 by English babble than by Mandarin Chinese babble for sentence recognition. Lecumberri  
200 and Cooke (2006) showed that English listeners performed better when the competing speech  
201 was Spanish.

202 A number of studies have demonstrated that speech perception in noise depends on the  
203 context information (Benkí, 2003; Bradlow and Alexander, 2007; Cutler et al., 2008). Benkí  
204 (2003) showed that the perception of CVC words in noise depends on the lexical status,  
205 word frequency, and neighborhood density as context effects. Bradlow and Alexander (2007)  
206 examined the semantic and phonetic enhancements for speech perception in noise by native  
207 and non-native listeners. The authors found that non-native listener's final word recognition  
208 improved only when both semantic and acoustic enhancements were available. In contrast,  
209 the native listeners benefited from each source of enhancement separately and in combination.

210 Redford and Diehl (1999) found that initial consonants were significantly more identifiable  
211 than final consonants for CVC syllables embedded in frame sentences.

212 Listener differences have also been investigated (Cutler et al., 2004; Lecumberri and  
213 Cooke, 2006; Bradlow and Alexander, 2007; Cutler et al., 2008; Cooke et al., 2008). Cutler  
214 et al. (2004) examined English phoneme confusions by native and non-native listeners in CV  
215 and VC syllables embedded in multi-talker babble. Although non-native listeners performed  
216 less accurately than native listeners at all noise levels, the effects of language background  
217 and noise did not interact. That is, there were no differential effects of noise on non-native  
218 listening. Lecumberri and Cooke (2006) studied the identification of American English con-  
219 sonants in /aCa/ context with noise being a single competing talker, speech-shaped noise,  
220 or eight-talker babble. The authors showed that non-native listeners were more adversely  
221 affected by noise than native listeners. In a follow-up study, Cutler et al. (2008) presented  
222 the (Lecumberri and Cooke, 2006) experiment to the listeners from the population in (Cutler  
223 et al., 2004) in the quiet and multi-talker babble conditions. Larger noise effects on consonant  
224 identification emerged for non-native listeners than for native listeners, suggesting that task  
225 factors (consonant identification in CV and VC syllables vs. in /aCa/ syllables) rather than  
226 non-native population differences (Dutch vs. Spanish) underlie the discrepancy between the  
227 (Cutler et al., 2004) and (Lecumberri and Cooke, 2006) studies. Cooke et al. (2008) studied  
228 the native and non-native listeners' keywords in English sentences in quiet and masked by  
229 either speech-shaped noise or a competing talker. The authors showed non-native talkers  
230 suffered more from increasing levels of noise.

#### 231 *1.4. The present study*

232 In the present study, we examine the relationship between the acoustic properties of  
233 speech signals and the results from perceptual experiments conducted in the presence of ad-  
234 ditive white Gaussian noise. Our overall goal is to discover the perceptual effect of noise on  
235 acoustic cues for place of articulation and to develop a deeper understanding of the place of  
236 articulation perception in noise. First, measurements of a number of acoustic properties from  
237 a set of CV utterances was made (in quiet) and analyzed for possible place-of-articulation  
238 cues using logistic regression analyses. Second, perceptual experiments were conducted us-  
239 ing the speech tokens mixed with varying amounts of white Gaussian noise. Finally, the

240 acoustic measurements were examined in conjunction with the results from the perceptual  
241 experiments to determine which cues could possibly account for the perception of place of  
242 articulation in noise. This was done by performing correlation analyses between the acoustic  
243 measurements and the place of articulation discrimination threshold SNRs.

244 The present study will contribute to the literature in two ways: (1) it studies a compre-  
245 hensive set of acoustic cues relevant to place of articulation, as reported in various papers,  
246 using a single context (CV syllables) with three vowels, and (2) it examines the perceptual  
247 relevance of these cues in quiet and in noise across a range of consonants (plosives and frica-  
248 tives). Most of the cues were implicated in many separate prior studies, and it is important  
249 to investigate their noise robustness in a single context. The noise robustness of these cues for  
250 place of articulation perception is examined across plosives and fricatives rather than within  
251 each manner of articulation, which could result in more general and consistent results. As a  
252 first step in this research direction, the present study uses naturally-spoken CV syllables for  
253 which higher-level factors such as lexical frequency or contextual information are irrelevant.

## 254 **2. Acoustic analysis**

### 255 *2.1. Stimuli*

256 Stimuli consisted of isolated, naturally-spoken CV utterances, where C was from the  
257 set  $\{/b/,/d/,/p/,/t/,/f/,/s/,/v/,/z/\}$  and V was from the set  $\{/a/,/i/,/u/\}$ , for a total of  
258 24 syllables. Speech signals were recorded in a sound-attenuating room using a headset  
259 microphone and were sampled at a rate of 16 kHz with a 16 bits per sample representation.  
260 Four talkers (two males, two females; age range 18 to 36 years), all native speakers of  
261 American English, were recorded. Each talker produced eight tokens for each CV, while  
262 only four of them were used for the present study (the first three tokens and the last one  
263 were discarded), resulting in a total of 16 tokens per CV syllable. Syllables were sorted  
264 in labial/alveolar pairs (such as /ba/ and /da/), such that manner of articulation, voicing,  
265 and vowel context were identical, and the two syllables in each pair differed only in the  
266 place-of-articulation dimension. Thus, there were a total of 12 CV pairs (see Table 1).

Table 1: CV pairs used in this study.

voiced		voiceless		
plosives	fricatives	plosives	fricatives	
/a/	/ba,da/	/va,za/	/pa,ta/	/fa,sa/
/i/	/bi,di/	/vi,zi/	/pi,ti/	/fi,si/
/u/	/bu,du/	/vu,zu/	/pu,tu/	/fu,su/

## 2.2. Acoustic measurements

All tokens were normalized such that the peak amplitude of the entire sampled waveform was set to the same level. Acoustic measurements were made for the speech tokens in quiet. The total set of measured properties is described in Table 2.

### 2.2.1. Formant frequency and amplitude measurements

Formant measurements (frequency and amplitude) were made from the time waveforms, wideband spectrograms, LPC (Linear Predictive Coding) spectra, and short-time DFT (Discrete Fourier Transform) spectra using Matlab. To obtain a spectrum, a 20 ms (for tokens from male talkers) or 15 ms (for tokens from female talkers) Hamming window was applied to define an analysis segment.

Each segment was zero-padded for a 1024-point FFT analysis, and the frame shift was half the Hamming window length. For an LPC analysis, no zero padding was applied, the frame shift was 2.5 ms for all talkers, and the LPC order was between 8 and 12 (depending on the variance of the prediction error). Vowel measurements included the first three formants (F1, F2, and F3). The three formants were located by examining the LPC spectra (Fig. 1a) and spectrograms. Three landmark points were defined for each formant: onset, offset, and steady state (Fig. 1c). F1, F2, and F3 onsets, chosen manually, were defined as the center point of the frame that exhibited the following characteristic: a sudden spectral change in the corresponding frequency range, particularly the introduction of a sharp spectral peak. The end of a formant transition (offset), chosen automatically, was defined as the frame during which the rate of change of the formant frequency fell to less than 5 Hz per 2.5 ms, and the average rate of change for the next 12.5 ms was also less than 5 Hz per 2.5 ms (Kewley-Port, 1982, see Fig. 1d). The steady-state point was centered at 95 ms after the onset, and the

290 steady-state measurements were averaged over five frames. At the formant transition onset,  
291 offset, and steady-state points, formant frequency (in Hz), and formant amplitude (in dB)  
292 were recorded based on the LPC spectrum. From these measurements, formant frequency  
293 and amplitude changes were measured between the formant transition onset and steady  
294 state. Formant transition duration was defined as the time difference between the formant  
295 transition offset and onset.

### 296 *2.2.2. Duration and relative spectral amplitude measurements*

297 The burst, frication noise, and VOT measurements were made by visually inspecting the  
298 time waveforms and wideband spectrograms of the tokens using the software CoolEdit Pro.  
299 Wideband spectrograms were calculated using a 6.4 ms Hamming window with a frame shift  
300 of one sample. The burst was defined as the short segment characterized by a sudden, sharp  
301 vertical line in the spectrogram. If multiple bursts were present, the burst duration was  
302 measured (in ms) from the beginning of the first burst to the end of the last. The spectrum  
303 of the combined transient and burst was estimated using Welch's averaged periodogram  
304 method (Stevens et al., 1999). That is, the signal was divided into overlapping sections of  
305 specified window length. If the burst duration was shorter than 9 ms, then a 3 ms window  
306 with 1.5 ms overlap was used; otherwise, a 6 ms window with a 3 ms overlap was used.  
307 The spectrum was obtained using a 256 point FFT (Fast Fourier Transform) method. VOT  
308 duration in plosives was measured from the end of the burst to the beginning of the vowel,  
309 which was also the beginning of the first waveform period. VOT duration in fricatives was  
310 measured from the consonant release to the beginning of the vowel, including noise duration  
311 and aspiration.

312 Ahi represents the peak amplitude of the burst/noise spectrum in the frequency range  
313 above 3500 Hz for female talkers and 3000 Hz for male talkers. A23 and A45 are the average  
314 amplitudes of the burst/noise spectrum in the F2-F3 and F4-F5 regions, respectively. Av  
315 and Av4 represent the peak amplitudes of the vowel spectrum at the F1 and F4 prominence,  
316 respectively. The pA23 and pA45 measures represent the peak amplitudes of the burst/noise  
317 spectrum in the F2-F3 and F4-F5 regions, respectively. Am and Avm are the average  
318 amplitudes of the burst and vowel onset spectrum at mid frequencies (between 3200 Hz and  
319 4800 Hz), respectively. Ans represents the average amplitude of the entire noise spectrum.

320 All these measures were in dB. For the vowels /a/ and /u/, F2-F3 and F4-F5 formant  
321 frequency regions are 1000-3000 and 3000-5000 Hz, respectively. For vowel /i/, F2-F3 and F4-  
322 F5 formant frequency regions are 1500-3500 and 4000-6000 Hz, respectively. The definitions  
323 of Ahi, A23, and pA23 are illustrated in Fig. 1b.

324 From these measurements, a set of relative spectral amplitude measures was constructed:  
325 (1) Ahi-A23 characterizes the spectral tilt of the burst/noise; (2) Av-Ahi is the high-  
326 frequency burst/noise spectral amplitude relative to F1 amplitude in the vowel; (3) Av-pA23  
327 is calculated only for plosives (Stevens et al., 1999) to determine a mid-frequency spectral  
328 prominence; (4) Av4-A45 characterizes the relative spectral amplitude of the vowel versus  
329 the burst/noise in the F4-F5 region; (5) Av4-pA45 is very similar to Av4-A45 except that  
330 the peak amplitude of the burst/noise in the F4-F5 region is calculated; (6) Am-Avm char-  
331 acterizes the difference between burst and vowel spectral amplitude at the mid-frequency  
332 range for plosives (Stevens et al., 1999); and (7) Av-Ans quantifies the overall amplitude  
333 of the noise relative to spectral amplitude of the vowel at the F1 prominence for fricatives  
334 (Hedrick and Ohde, 1993).

335 The measurements Av-Ahi, Ahi-A23, and Av-pA23 were inspired by Stevens et al. (1999),  
336 except for two differences in calculating these noise measures. First, the burst segment in  
337 Stevens et al. (1999) did not include aspiration in voiceless plosives. Second, Stevens et al.  
338 (1999) used the same window length for both the vowel onset and the burst segment. In  
339 addition, the present study also examined the noise properties in the F4 to F5 regions  
340 that had been suggested for place of articulation distinction (Hedrick and Jesteadt, 1996;  
341 Hedrick et al., 1995). Furthermore, the average of the entire noise spectrum was measured  
342 for fricatives.

343 \_\_\_\_\_  
344 Table 2 about here (actual table on Page 35)

345 \_\_\_\_\_  
346

### 347 *2.3. Place-of-articulation classification based on acoustic measurements*

348 The acoustic measurements were analyzed using logistic regression, where the quiet  
349 speech tokens were classified as either labial or alveolar according to a single acoustic property

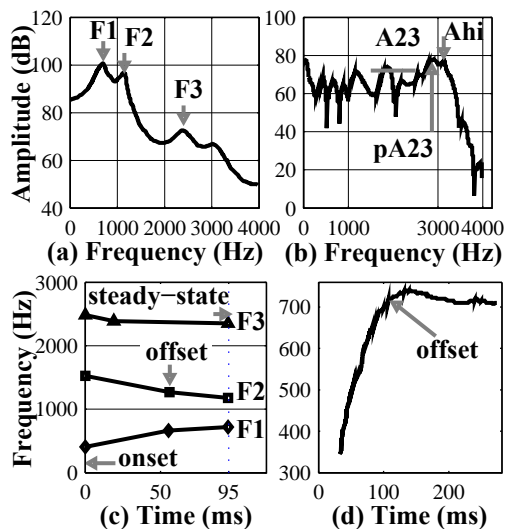


Figure 1: (a) LPC spectrum of a /ta/ token during the vowel, (b) DFT spectrum of a /ta/ token during the burst, (c) formant transition measurements, and (d) illustration of the determination of formant transition offset (in this case, F1 frequencies obtained using LPC analyses) when the change in frequency drops below 5 Hz per 2.5 ms.

350 measured without the addition of the white Gaussian noise. A separate logistic regression  
 351 model was applied to each acoustic variable for each CV pair,

$$\log[\text{prob}/(1 - \text{prob})] = \alpha + \beta \cdot \text{Mea} + e \quad (1)$$

352 where *prob* is the probability of a token being labial,  $\alpha$  is a constant,  $\beta$  is a weighting  
 353 coefficient, *Mea* is one acoustic feature (measurement), and *e* is the error term. For each  
 354 token, the consonant was either labial or alveolar, and thus *prob* was either 0 or 1. After  
 355 logistic regression,  $\alpha + \beta \cdot \text{Mea} = 0$  was used for classification, and results were compared  
 356 against ideal classification to obtain the percent correct scores. Table 3 lists the results  
 357 in terms of percent correct classification based on logistic regression using the tokens from  
 358 all talkers. Only acoustic measures with 79% or higher correct classification are listed and  
 359 sorted (from high to low) for each CV pair.

360 \_\_\_\_\_

361 Table 3 about here (actual table on Page 36)

362 \_\_\_\_\_

363

364 Of the 37 recorded measurements in Table 2, a number of acoustic measurements do not  
365 appear in Table 3. That is, these acoustic measurements were not prominent in classifying the  
366 labial/alveolar place-of-articulation distinction. These non-prominent measurements include  
367 formant steady-state frequencies and amplitudes, formant offset amplitudes, F3 onset and  
368 offset frequencies, F3 onset amplitude, F1 and F3 amplitude change, F1 and F3 transition  
369 duration, and Av-pA23. Several other acoustic measurements, although they appear in Table  
370 3, produced moderate place of articulation classification performance for only one or two CV  
371 pairs. Such measurements include F1 offset frequency (F1e, 81% for /bu,du/ and /vu,zu/),  
372 F1 frequency change (F1df, 81% for /fa,sa/), F1 and F2 onset amplitude and F2 amplitude  
373 change (F1bA, F2bA, and F2dA, 84% for /bi,di/), F2 transition duration (F2D, 84% for  
374 /ba,da/), VOT duration (votD, 84% for /vi,zi), Av4-A45 (84% for /pu,tu/), and Am-Avm  
375 (84% for /bi,di/). A first generalization from these non-prominent acoustic measures is that  
376 formant amplitudes, steady-state frequencies, and offset frequencies were not discriminative  
377 for labial/alveolar place of articulation classification. An exception is that the F2 offset  
378 frequency (F2e) yielded moderate classification performance for several CV pairs (84% for  
379 /bu,du/, 84% for /pu,tu/, 81% for /va,za/, and 81% for /fu,su/). A second generalization is  
380 that the voicing feature measurements (e.g., VOT) were not reliable cues for labial/alveolar  
381 place of articulation except for the noise/burst duration measurements.

382 Several formant frequency measurements, F1 and F2 onset frequencies and F2 and F3  
383 frequency changes (F1b, F2b, F2df, and F3df), were mostly distinctive for the /a/-context  
384 labial/alveolar pairs (11 out of 16 cases), moderately for the /u/-context ones (4 out of  
385 16 cases), but not for the /i/-context ones. Labials had a higher F1 onset frequency than  
386 alveolars except for /pi,ti/ and /pu,tu/. F2 onset frequency was lower for labials than for  
387 the alveolars by 200-400 Hz except for the /i/ context where the onsets were approximately  
388 the same. The F2 frequency change was smaller in amplitude for labials than for alveolars,  
389 and this difference was the most prominent for the /a/-context pairs and least prominent  
390 for the /i/-context ones (see Fig. 2). The F3 frequency change was a distinctive cue for  
391 labial/alveolar place of articulation for plosives in the /a/ and /i/ contexts but not in the  
392 /u/ context.

393 The relative spectral amplitude measurements, Ahi-A23, Av-Ahi, and Av4-pA45, were

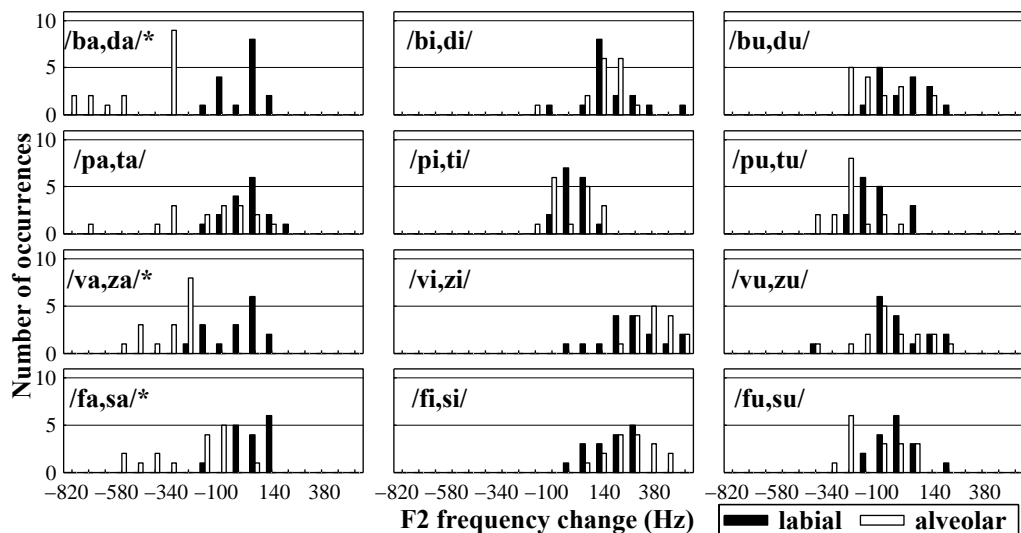


Figure 2: Histograms of F2 frequency change (F2df) for the 12 labial/alveolar pairs with the labial and alveolar tokens counted separately. The histogram bin centers ranges from -820 to 540 Hz with a 80 Hz step. F2df of less than -860 Hz and of more than 580 Hz is counted into the -820 Hz center and 540 Hz center regions, respectively. F2df was a reliable cue for the vowel /a/ pairs except for /pa,ta/. Asterisks are added next to the CV pair name to indicate 79% or above correct classification of place of articulation.

394 reliable cues for labial/alveolar place of articulation for both plosives and fricatives to varying  
 395 degrees. Ahi-A23 was higher in alveolars than in labials for plosives (by about 4 dB for voiced  
 396 plosives and 14 dB for voiceless plosives), but approximately the same for fricatives (with  
 397 a difference of about 2 dB). However, in the classification analyses, the measure was only  
 398 reliable for voiceless plosives. Av-Ahi in labials were, on average, about 23 dB and 4 dB  
 399 higher than those in alveolars for plosives and fricatives, respectively. Therefore, Av-Ahi  
 400 produced relatively high labial/alveolar place of articulation classification for plosives (e.g.,  
 401 100% for /pu,tu/) except for /pa,ta/. Six out of the 12 pairs were reliably classified by Av4-  
 402 pA45 whose values in labials were, on average, about 16 dB and 30 dB higher than those in  
 403 alveolars for plosives and fricatives, respectively (see Fig. 3). In fact, for voiceless plosives  
 404 and fricatives in the /i/ and /u/ contexts, classification was above 80% correct using only  
 405 Av4-pA45. Av-Ans appeared to be the most reliable cue for classifying place of articulation  
 406 for five out of six fricative pairs. That is, it resulted in 100% classification for /vi,zi/ and  
 407 /fu,su/, 94% for /va,za/, and 91% for /vu,zu/ and /fa,sa/. Av-Ans measurements were  
 408 higher for labial fricatives than for alveolar ones (by about 17 dB on the average). The

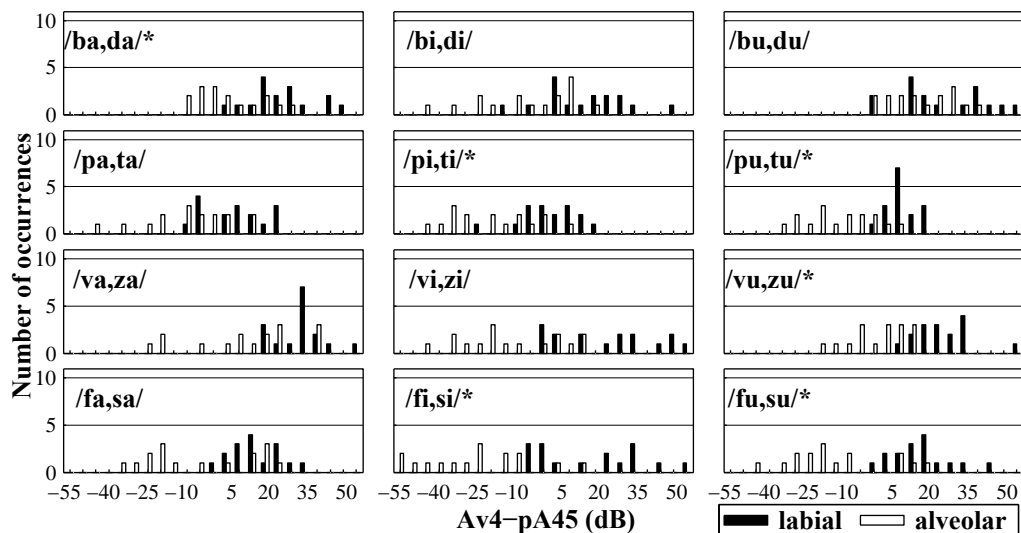


Figure 3: Histograms of Av4-pA45 for the 12 labial/alveolar pairs with the labial and alveolar tokens counted separately. The histogram bin centers ranges from -55 to 55 dB with a 5 dB step. Av4-pA45 of less than -57.5 dB and of more than 57.5 dB is counted into the -55 dB center and 55 dB center regions, respectively. Av4-pA45 was distinctive for /ba,da/, /pi,ti/, /pu,tu/, /vu,zu/, /fi,si/, and /fu,su/. Asterisks are added next to the CV pair name to indicate 79% or above correct classification of place of articulation.

409 place-of-articulation distinctions appeared to be more prominent in the higher frequency  
 410 ranges (i.e., Av4-A45, Av4-pA45, and Av-Ans) for fricatives than for plosives. In plosives,  
 411 burst duration (bstD) signaled labial/alveolar place of articulation for /ba,da/, /pa,ta/, and  
 412 /pi,ti/. In fricatives, noise duration (nD) appeared to be a cue for labial/alveolar place of  
 413 articulation for /va,za/ and /vi,zi/. The noise duration was about 40 ms longer for alveolars  
 414 than for labials.

415 In summary, formant amplitudes, steady-state frequencies, offset frequencies (except F2  
 416 offset frequency), and voicing feature measurements (except noise/burst duration measure-  
 417 ments) were generally not discriminative for labial/alveolar place of articulation classifica-  
 418 tion. Several formant frequency measurements (F1 and F2 onset frequencies and F2 and F3  
 419 frequency changes) were somewhat distinctive for labial/alveolar place of articulation classi-  
 420 fication (mostly in the /a/ context and moderately in the /i/ context). The relative spectral  
 421 amplitude measurements in the higher frequency ranges, Ahi-A23, Av-Ahi, and Av4-pA45,  
 422 were reliable cues for labial/alveolar place of articulation for both plosives and fricatives to  
 423 varying degrees, consistent with the results in (Stevens et al., 1999) on Av-Ahi and Ahi-A23

424 and the results in (Hedrick et al., 1995) on Av4-pA45. Av-Ans was the most reliable cue for  
425 classifying place of articulation for fricatives. The burst and noise duration measurements,  
426 bstD and nD, appeared to be a moderate cue for labial/alveolar place of articulation for  
427 plosives and fricatives, respectively. A summary of the relevance of several acoustic prop-  
428 erties in classifying labial/alveolar place of articulation for plosives and fricatives is shown  
429 in Table 4. (Threshold SNRs from the perception experiment are also given; see Section  
430 3.4.) Generally speaking, formant frequency measurements (F1b, F2b, F2df, F3df), relative  
431 spectral amplitude measurements (Ahi-A23, Av4-pA45, Av-Ahi, Av-Ans), and noise/burst  
432 duration were reliable cues for labial/alveolar place of articulation (marked by asterisks in  
433 Table 4).

434 \_\_\_\_\_  
435 Table 4 about here (actual table on Page 37)

436 \_\_\_\_\_  
437

### 438 3. Perceptual study

#### 439 3.1. Stimuli

440 All 384 CV tokens described in Sec. 2.1 were used as stimuli for the perceptual study.  
441 The masking noise used in the perceptual experiments was a 1250-ms segment of white  
442 Gaussian noise. At the beginning of each experimental session, 32 Gaussian noise sources  
443 were generated. During the presentation of each stimulus, a noise masker was randomly  
444 selected from the 32 Gaussian noise sources. The SNR was defined as the ratio of the  
445 maximum root mean square (RMS) value in the CV to the RMS value of the noise token  
446  $[20 \log_{10}(max\_RMS_{CV}) - 20 \log_{10}(RMS_{noise})]$ . A post-hoc examination of the acoustic  
447 portions with maximum RMS energy indicated that the first term occurred in the vowel  
448 part for most of the CV tokens. The maximum RMS energy of a token was computed using  
449 a 30 ms rectangular window so as to exclude acoustic spikes. The use of maximum RMS  
450 energy was consistent with the approach in (Miller and Nicely, 1955) where SNR was set  
451 based on the peak deflection of the VU needle. The RMS energy of the noise was based

452 on the entire noise segment. Hence, the SNR did not depend on the duration of the speech  
453 token.

### 454 *3.2. Participants*

455 Listening experiments were conducted with four participants (two males, two females;  
456 age range 18 to 36 years; different from the speakers), all native speakers of American English  
457 who passed a hearing test (i.e., their hearing thresholds were equal to or below 10 dB HL,  
458 sound pressure level, from 250 Hz to 8 kHz).

### 459 *3.3. Procedure*

460 Perceptual testing took place in a sound-attenuating room. Digital speech stimuli were  
461 played via an Ariel Pro Port 656 board digital-to-analog converter (16 bits at a rate of 16  
462 kHz). The resulting analog waveforms were amplified by a Sony 59ES DAT recorder and were  
463 then presented binaurally via Telephonics TDH49P headphones. The system was calibrated  
464 within 0.5 dB (from 125 to 7500 Hz at third octave intervals) using a 6-cc coupler and a  
465 Larson Davis 800B sound level meter (with the “A” weighting scale and a slow response)  
466 prior to each experiment.

467 Each signal (without noise) was played at 60 dB SPL, and the accompanying noise level  
468 was adjusted. The SPL of the speech signals were set based on their maximum RMS energy  
469 in a 30 ms rectangular window around the maximum level of the CV. The SPL of the white  
470 Gaussian noise was adjusted based on its RMS energy to result in different SNRs. The  
471 speech signal was added to a 1250 ms noise (or silence) segment such that it was centered  
472 in the middle of the segment.

473 Participants made two-alternative forced choices (2-AFC). Utterances were played in  
474 blocks of 64 tokens of a single CV pair (32 tokens x 2 presentations). When an utterance  
475 was played, subjects were asked to label the sound heard as either the labial or alveolar  
476 consonant (e.g., /b/ or /d/). A computer program was developed to record participants’  
477 responses from their keyboard inputs. No feedback was given at any time. The test was  
478 then repeated at different SNR levels. The order of SNR conditions was: quiet, 10 dB, 5  
479 dB, 0 dB, -5 dB, -10 dB, and -15 dB (same order for all listeners). The CV pairs were  
480 presented in the order of /ba,da/, /bi,di/, /bu,du/, /pa,ta/, /pi,ti/, /pu,tu/, /fa,sa/, /fi,si/,

481 /fu,su/, /va,za/, /vi,zi/, and /vu,zu/. To counterbalance the effects of talker and token  
482 order, the order of presenting the 64 tokens within each CV pair was pseudo-randomized.  
483 Participants were forced to take a break after each CV pair and were instructed to take at  
484 least one break every hour. Also, they were allowed to take voluntary breaks if they felt  
485 tired while listening to each CV pair. Each session lasted about one hour but no longer than  
486 two hours to prevent fatigue. On Day 1, each participant had a one-hour training session.  
487 During training, the experimenter explained and demonstrated the experimental procedure  
488 to the participants; participant read a written instruction; and participants then had a set  
489 of practice trials during which they could ask the experiment questions.

### 490 *3.4. Results*

#### 491 *3.4.1. Percent correct classification and threshold SNRs for place of articulation in noise*

492 \_\_\_\_\_

493 Table 5 about here (actual table on Page 38)

494 \_\_\_\_\_

495

496 The percentage of correct place of articulation judgments was computed and listed as  
497 a function of SNR, manner of articulation, voicing, and vowel context (see Table 5). The  
498 percent correct values were calculated using all the data collected from the perceptual exper-  
499 iments, including all listeners and all talkers. Each data entry thus represents 256 responses  
500 from four listeners for a CV pair at a specific SNR condition (4 talkers x 4 listeners x 4  
501 tokens x 2 presentations x 2 consonants).

502 Most of the 12 CV pairs had close to 100% correct place of articulation judgments in the  
503 absence of noise. The listeners appeared to have had a particularly difficult time classifying  
504 the /pa,ta/ pair (with 81% correct place of articulation judgments even when the SNR was  
505 10 dB). However, for CV pairs other than /pa,ta/, the percent correct was 92% or above  
506 when the SNR was 10 dB. Among the 12 CV pairs, the /f,s/ pairs appeared to have the  
507 best place of articulation judgments (93-96% correct) when the SNR was 0 dB. For SNRs of  
508 -10 dB and below, place of articulation judgments for all 12 pairs were dramatically affected  
509 by noise (below 70% correct). When the SNR was -15 dB, the percent correct of place of  
510 articulation judgments was about 50%, which is chance performance.

511 In order to analyze how the acoustic properties account for the perceptual results, a single  
 512 SNR value for each CV pair was needed to represent the robustness of that CV pair in the  
 513 presence of noise. That value, or threshold (in dB), was computed along the SNR continuum  
 514 at which the percent correct of responses is 79% (Levitt, 1971). The perceptual results for  
 515 the 12 CV pairs were arranged into plots as shown in Fig. 4 where percent correct is plotted  
 516 versus SNR. For the quiet conditions, the SNR was estimated as 21 dB. A sigmoid was then  
 517 fitted to each plot and described by the following equation:

$$y = c + \frac{d - c}{2} \left( 1 - \frac{1 - e^{(x-b)/a}}{1 + e^{(x-b)/a}} \right) \quad (2)$$

518 where  $x$  and  $y$  represent SNR and percent correct, respectively;  $d$  and  $c$  are the maximum  
 519 and minimum values of percent correct, respectively;  $a$  and  $b$  are parameters to adjust the  
 520 slope and position of the transition of the sigmoid function between the top and bottom  
 521 flat areas. In this study,  $c$  was set to 50%, the chance performance, and  $d$  was set to 100%.  
 522 Therefore,  $a$  and  $b$  varied systematically to obtain the best fit sigmoid by minimizing the  
 523 mean squared error. From the best fit sigmoid, the threshold SNR level corresponding to  
 524 79% correct responses was obtained. Thus, a single threshold SNR value for each of the 12  
 525 pairs of labial/alveolar CV syllables was calculated to represent the perceptual robustness  
 526 of that pair. A lower threshold SNR corresponded to better perceptual results (more robust  
 527 to noise).

528 The sigmoid fitting was applied to perceptual results for each CV pair and each lis-  
 529 tener. The obtained threshold SNRs were submitted to an omnibus repeated measures  
 530 analysis of variance with the manner of articulation (2), voicing (2), and vowel (3) as within-  
 531 subjects factors. The only reliable interaction was between manner of articulation and  
 532 voicing [ $F(1,3)=23.9, p=.016$ ]. That is, for plosives, voiced CV syllables yielded better place  
 533 of articulation classification than voiceless ones (by about 2.4 dB on average with the /i/  
 534 context as an exception), while the opposite was true for fricatives (by about 1.5 dB on aver-  
 535 age). Note that in (Miller and Nicely, 1955), the authors suggested the relative independence  
 536 between the perception of manner of articulation and that of voicing. This inconsistency  
 537 might result from the task difference between the two studies (i.e., open-set identification  
 538 vs. 2-AFC on place of articulation. The main effect of manner of articulation was reliable

539 [ $F(1,3)=61.1, p=.004$ ], with fricatives (mean threshold SNR = -3.9 dB) being more robust  
 540 than plosives (mean threshold SNR = 0.9 dB), agreeing with the results from (Miller and  
 541 Nicely, 1955). A possible reason may be due to the differences in noise spectra between  
 542 labial and alveolar CV syllables for plosives and fricatives. Generally speaking, fricatives  
 543 have longer duration, and thus their noise spectral differences for place of articulation can  
 544 be more easily perceived than those for plosives. The main effects of vowel context was  
 545 marginally significant [ $F(2,2)=16.8, p=.056$ ]. The vowel /i/ context yielded high threshold  
 546 SNRs (less robust) than the /a/ [ $F(1,3)=9.8, p=.052$ ] and /u/ [ $F(1,3)=29.0, p=.013$ ] con-  
 547 texts, but there was no significant difference between the /a/ and /u/ contexts [ $F(1,3)=0.7,$   
 548  $p=.452$ ]. This agrees with (Hant, 2000), where /bi,di/ was the least robust while /ba,ga/  
 549 was the most robust. This vowel effect on threshold SNRs may result from the fact that  
 550 formant frequency measurements were distinctive in /a/ and /u/ contexts in quiet (except  
 551 for /pa,ta/), but not for /i/ contexts (see Sec. 2.3). The mean threshold SNRs were -1.9  
 552 dB, 0 dB, and -2.6 dB for the /a/, /i/, and /u/ pairs, respectively. The /pa,ta/ pair was an  
 553 exception for the vowel effect. The main effect of voicing was not significant [ $F(1,3)=1.6,$   
 554  $p=.297$ ]. As a demonstration, the threshold SNR levels at 79% correct for all CV pairs are  
 555 shown in Fig. 4, where percent correct scores were averaged over all talkers and all listeners.

556 Table 4 lists the threshold SNRs and the relevance of several acoustic properties in classi-  
 557 fying labial/alveolar place of articulation for the 12 CV pairs in quiet conditions. For plosives,  
 558 three pairs (/ba,da/, /bu,du/, and /pu,tu/), which had formant frequencies in addition to  
 559 noise measurements as cues, were relatively more robust in noise for the labial/alveolar place  
 560 of articulation distinctions. The other three plosive pairs (/pi,ti/, /bi,di/, and /pa,ta/) did  
 561 not have formant frequency cues, and correspondingly their threshold SNRs are 0 dB or  
 562 above. Fricative pairs in general have lower threshold SNRs (less than -1 dB) compared to  
 563 the plosive ones. Similarly for fricative pairs, both the formant frequencies and the relative  
 564 spectral amplitude measurements appeared to be responsible for the low threshold SNRs.

565 In summary, the perception of labial/alveolar place of articulation in noise depended on  
 566 the interaction between voicing and manner of articulation, manner of articulation, and vowel  
 567 context. Fricatives were generally more robust than plosives. The labial/alveolar distinction  
 568 was not robust in the vowel /i/ context.

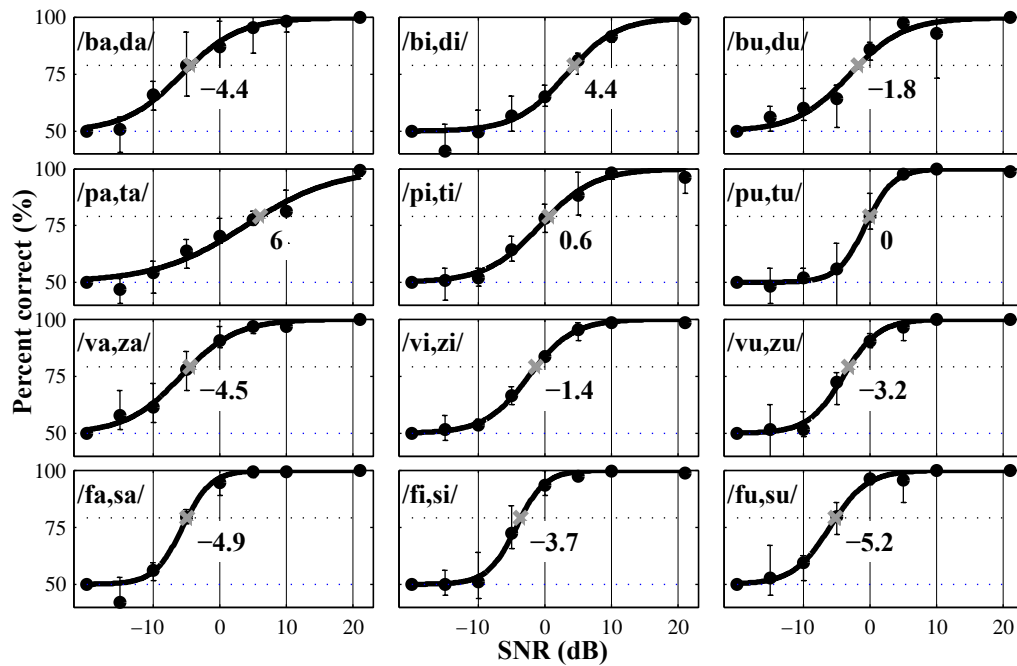


Figure 4: A sigmoid fitting (solid line) of percent correct scores as a function of SNR (dB) for the 12 labial/alveolar pairs. For each pair, the 79% threshold line is drawn, and the threshold SNR value is labeled. The average percent correct scores (from the four listeners) are in circles. The error bars represent the minimum and maximum numbers among the four listeners.

569 **4. Correlations between threshold SNR values and absolute acoustic differences**  
570 **of the means**

571 Correlations were computed between the 12 threshold SNR values from the perceptual  
572 experiments and the absolute differences of the mean values of each measured acoustic prop-  
573 erty for the 12 labial/alveolar pairs. The mean value of each acoustic measurement for every  
574 CV syllable was calculated from 16 tokens (4 talkers x 4 tokens of the same syllable). The  
575 correlation is defined as

$$r = \text{corr}(\|\overline{Mea}_{la} - \overline{Mea}_{al}\|, 10 - SNR_t) \quad (3)$$

576 where *corr* represents the Pearson correlation function, *la* represents labial tokens, *al* repre-  
577 sents alveolar tokens, *Mea* represents one type of acoustic measurement, the bar over *Mea*  
578 represents the mean operation,  $SNR_t$  represents the threshold SNR values, and  $10 - SNR_t$   
579 indicates how much the threshold SNRs were below 10 dB. The assumption is that if an acous-  
580 tic property is an important cue for place of articulation, then a larger absolute difference  
581 between the means would correspond to better performance (a lower threshold SNR), while  
582 a smaller absolute difference between the means would correspond to poorer performance  
583 (a higher threshold SNR). A negative correlation coefficient indicates that larger differences  
584 correspond to higher (worse) threshold SNRs, which is opposite to a normal psychoacoustic  
585 relationship. Also, to evaluate how the correlations vary with perceptual performance levels,  
586 the threshold SNRs were re-estimated at a number of perceptual thresholds between 71%  
587 and 84% using Equation 2.

588 Pearson product correlation coefficients were obtained only for those acoustic properties  
589 that appear in Table 4. Figure 5 shows the results of correlating threshold SNRs with  
590 the absolute differences of the means of several acoustic properties for the 12 CV pairs  
591 along different perceptual performance levels. Those acoustic properties that had negative  
592 correlation coefficients were not displayed. The correlation coefficients shown in Fig. 5  
593 were not significant after Bonferroni correction because of the low N. They were therefore  
594 not definitive in and of themselves (because of the risk of false positives), but they were  
595 considered to be useful aids for the interpretation of the core results as given in Secs. 2.3  
596 and 3.4. For this reason, the correlations for each acoustic property were examined across

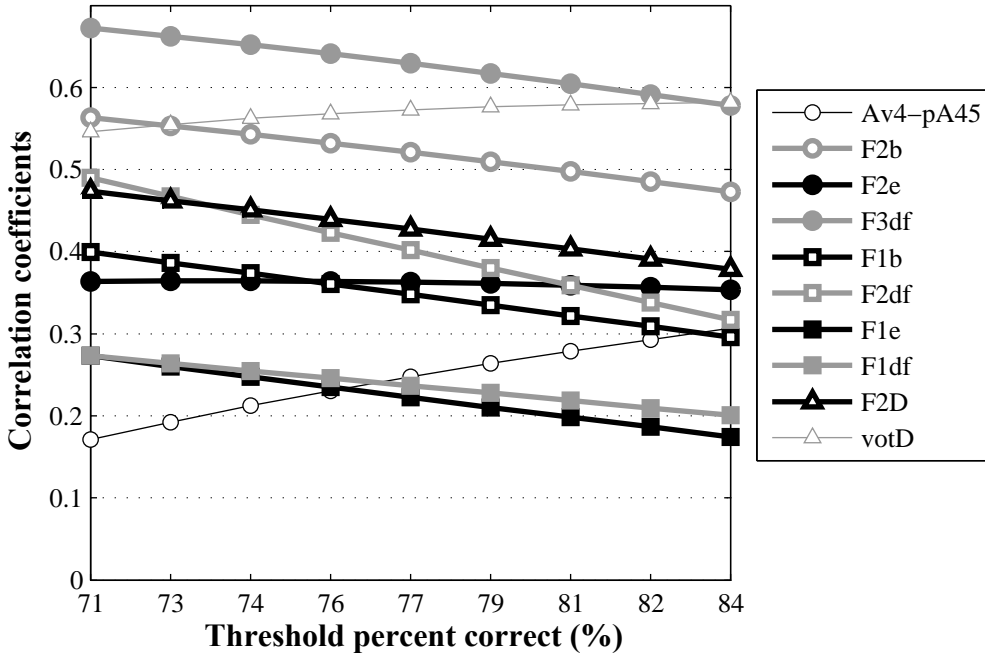


Figure 5: Correlation coefficients between threshold SNRs and acoustic measures (distances between means) across all talkers as a function of the threshold percent correct (71%-84%). Acoustic measures that produced negative correlations were not displayed.

597 different threshold percent corrects and were examined in terms their pattern (i.e., increasing  
 598 or decreasing with threshold percent corrects).

599 There were 10 acoustic properties that had positive correlations with threshold SNRs.  
 600 They are F1b, F1e, F1df, F2b, F2e, F2df, F2D, F3df, votD, and Av4-pA45. Their correla-  
 601 tions with threshold SNRs were between 0.15 and 0.70. Among the 10 acoustic properties,  
 602 the correlation for votD and Av4-pA45 became lower with decreasing perceptual perfor-  
 603 mance. That is, votD and Av4-pA45 became less effective when there was more noise (lower  
 604 performance levels). At all SNRs, votD was more effective than Av4-pA45. In contrast, the  
 605 eight formant measures became more effective when there was more noise (lower performance  
 606 levels). The order of these formant measures in terms of correlations (from high to low) was:  
 607 F3df, F2b, F2D, F2df, F2e, F1b, F1df, and F1e. The correlations for F3df, F2b, and votD  
 608 were at about the same level.

609 Consistent with the results in Sec. 2.3, formant amplitudes produced negative correla-  
 610 tions, indicating that formant amplitudes did not contribute to lowering the threshold SNRs  
 611 of labial/alveolar distinction in noise. In Sec. 2.3, F1 and F2 onset frequencies (F1b and

612 F2b) and F2 and F3 frequency changes (F2df and F3df) resulted in a high percentage of  
613 labial/alveolar classification for several quiet CV pairs. These formant properties also con-  
614 tributed to the place of articulation distinction in noise. For example, F3df, F1b, F2b, and  
615 F2df yielded relatively high correlations between the perceptual SNR thresholds and acous-  
616 tic measures. Other formant properties (e.g., F1e and F2e frequencies) also showed positive  
617 correlations, although they did not classify labial/alveolar well in quiet conditions.

618 Relative spectral amplitude measurements (e.g., Ahi-A23, Av-Ahi, Av-Ans, bstD, and  
619 Am-Avm), which were found in Sec. 2.3 to be acoustically distinctive in terms of place  
620 of articulation, usually produced negative correlations with the perceptual measures. This  
621 might be due to the relative spectral amplitude measurements being easily corrupted in the  
622 presence of additive noise. The acoustic property Av4-pA45 produced positive but low corre-  
623 lations. Although the VOT duration was not a distinctive acoustic feature for labial/alveolar  
624 place of articulation in quiet conditions, it resulted in relatively high correlations between  
625 acoustic and perceptual measures. The VOT duration was longer for fricatives than for  
626 plosives, while fricatives usually had lower threshold SNRs than plosives. Therefore, the  
627 relatively high correlations for VOT duration mainly resulted from the differences between  
628 fricatives and plosives.

629 In summary, formant frequency properties were more noise robust than the relative spec-  
630 tral amplitude measurements. Because the /a/ context resulted in larger absolute differences  
631 in the F3df, F2b, F2df, and F1b measurements between the labial and alveolar pairs than  
632 the /i/ and /u/ contexts, the place of articulation judgments in the /a/ context were more  
633 robust than those in the /i/ and /u/ contexts.

## 634 5. General discussion

635 The present study examines the acoustic correlates and perception in noise of place of ar-  
636 ticulation in naturally-spoken syllable-initial plosive and fricative consonants. Both formant  
637 frequency and relative spectral amplitude measurements were the cues most predictive of  
638 place of articulation decisions in quiet conditions, but relative spectral amplitude measure-  
639 ments appeared to be masked at low SNRs, with a contrasting result that formant frequency  
640 measurements were better place of articulation cues at low SNRs. Specifically, in quiet

641 conditions, all of the 12 CV pairs were correctly classified at or near 100% with formant  
642 frequency or relative spectral amplitude measurements. Nevertheless, no single cue showed  
643 high classification for both fricatives and plosives across all vowel contexts.

644 In the presence of noise, listeners could still make correct labial/alveolar place of artic-  
645 ulation judgments even when the SNR level was -5 dB. However, for an SNR of -15 dB,  
646 listeners' responses were equivalent to random guesses (chance performance). The present  
647 study showed that fricatives, in general, had lower threshold SNRs than plosives, agreeing  
648 with (Miller and Nicely, 1955). Similar to that in (Miller and Nicely, 1955), this study  
649 showed that voiceless fricatives, in particular, were slightly more robust than the voiced  
650 ones.

651 For place of articulation classification in noise, vowel effect was significant in the sense that  
652 the vowel /a/ context, except for /pa,ta/, yielded lower threshold SNRs than the /u/ context,  
653 which was more robust than the /i/ context. The reason might be that the distinctive  
654 acoustic features (e.g., F1b, F2b, F2df, and F3df) were most prominent for the /a/-context  
655 pairs and least prominent for the /i/-context pairs (see Sec. 2.3), and this was confirmed  
656 with the correlation analyses for which the high correlation coefficients usually resulted from  
657 the vowel differences in the formant frequency measurements. The vowel effect is consistent  
658 with the (Parikh and Loizou, 2005) study for which acoustic analyses indicated that F1 was  
659 detected more reliably than F2 and correlation analyses indicated that vowel identification  
660 scores were highly correlated with acoustic parameter values at a SNR of -5 dB. Interestingly,  
661 the formant frequency measurements for /pa,ta/ were not discriminative compared to other  
662 /a/-context pairs. The effect of manner of articulation was also reliable, which could be  
663 attributed to the noise durations in plosives and fricatives (Jongman, 1988).

664 Relative spectral amplitude measurements, although acoustically distinctive in quiet con-  
665 ditions, usually had low or negative correlations with the threshold SNRs (except for votD  
666 and Av4-pA45). These results indicate that the formant frequency measurements were more  
667 important for the perception of place of articulation at low SNRs than the relative spec-  
668 tral amplitude measurements, agreeing with acoustic analysis and perceptual results from  
669 (Parikh and Loizou, 2005). Compared to formant frequency measurements, relative spectral  
670 amplitude measurements are easily corrupted by noise, especially broadband noise. The

671 higher correlations between threshold SNR values and formant measurements at lower per-  
672 cent correct threshold are consistent with the glimpsing model of speech perception in noise  
673 for which listeners use spectro-temporal regions in which the target signal is least affected  
674 by the background for speech perception and integration (Hant and Alwan, 2003; Cooke,  
675 2006; Li and Loizou, 2007). That is, listeners use whatever cues are available, and those  
676 cues crucially depend on the nature of the noise masker. Therefore, the effect of the type of  
677 noise masker should also be taken into account. One speculation is that speech-shaped or  
678 multi-talker babble noise might also corrupt the formant frequency measurements. For ex-  
679 ample, in (Hedrick and Younger, 2007), the authors showed that for the perception of place  
680 of articulation in plosive consonants /p,t/, normal hearing listeners reduced their weight-  
681 ing of formant transitions and relied more on the relative spectral amplitude cues in the  
682 speech-shape noise than in the quiet condition. If the present study were carried out with  
683 speech-shaped or multi-talker babble noise, the correlations between threshold SNR values  
684 and relative spectral amplitude measurements at lower percent correct threshold might be  
685 higher.

686 A limitation in the correlation analyses is that the within- and between-talker variations  
687 were not examined due to the limited number of tokens and perceptual responses. One possi-  
688 bility is that the correlations were driven by data from one talker (between-talker variation)  
689 or some specific tokens within one talker (within-talker variation). This in turn would limit  
690 the generalization of results from the present study. However, the distributions of all acous-  
691 tic properties (e.g., Figs. 2 and 3) were examined and were found not to be not abnormal.  
692 Nevertheless, evaluating the within- and between-talker variations is an interesting future  
693 topic.

694 In summary, for white Gaussian noise, the formant frequency measurements are more  
695 dominant cues for labial/alveolar place of articulation than relative spectral amplitude mea-  
696 surements; place of articulation perception is dependent on the interaction of voicing and  
697 manner of articulation, manner of articulation, and vowel context; and no single acous-  
698 tic feature could cue perception of place of articulation. These results could eventually be  
699 useful for hearing aids, cochlear implant processing algorithms, and noise-robust automatic  
700 speech recognition. For example, for hearing aids, better noise reduction algorithms could

701 be designed by enhancing noise-level-dependent salient acoustic cues.

702 In future studies, experiments will be conducted using a larger dataset and synthetic  
703 stimuli to construct acoustic continua and to control interactions between a limited number  
704 of acoustic properties (e.g., independently vary the F2 onset frequency, the F2 frequency  
705 change, and the F3 frequency change). Perceptual experiments can be expanded by masking  
706 the stimuli with different types of noise maskers (e.g., perceptually flat noise, speech-shaped  
707 noise, multi-talker babble, car noise, etc.) In addition, perceptual experiments can be per-  
708 formed with listeners with cochlear implants so as to help understand which speech cues  
709 they rely on in a noisy environment.

## 710 **Acknowledgments**

711 This work was supported in part by NIH-NIDCD grant 1R29-DC02033-01A1, the NSF,  
712 and a Fellowship from the Radcliffe Institute to Abeer Alwan. We thank Marcia Chen for  
713 her help in data analysis and Wendy Espeland, Marwa Elshakry, and Christine Stansell  
714 for commenting on an earlier version of this manuscript. Thanks also to Steven Lulich for  
715 constructive comments.

## 716 **References**

717 Alwan, A., 1992. The role of F3 and F4 in identifying the place of articulation for stop con-  
718 sonants. In: Proceedings of the International Conference on Spoken Language Processing.  
719 Banff, Canada, pp. 1063–1066.

720 Behrens, S., Blumstein, S. E., 1988. On the role of the amplitude of the fricative noise in the  
721 perception of place of articulation in voiceless fricative consonants. *J. Acoust. Soc. Am.*  
722 84 (3), 861–867.

723 Benkí, J. R., 2003. Quantitative evaluation of lexical status, word frequency, and neighbor-  
724 hood density as context effects in spoken word recognition. *J. Acoust. Soc. Am.* 113 (3),  
725 1689–1705.

- 726 Blumstein, S. E., Stevens, K. N., 1979. Acoustic invariance in speech production: Evidence  
727 from measurements of the spectral characteristics of stop consonants. *J. Acoust. Soc. Am.*  
728 66 (4), 1001–1017.
- 729 Bradlow, A. R., Alexander, J. A., 2007. Semantic and phonetic enhancements for speech-in-  
730 noise recognition by native and non-native listeners. *J. Acoust. Soc. Am.* 121 (4), 2339–  
731 2349.
- 732 Cooke, M., 2006. A glimpsing model of speech perception in noise. *J. Acoust. Soc. Am.*  
733 119 (3), 1562–1573.
- 734 Cooke, M., Lecumberri, M. L. G., Barker, J., 2008. The foreign language cocktail party  
735 problem: Energetic and informational masking effects in non-native speech perception. *J.*  
736 *Acoust. Soc. Am.* 123 (1), 414–427.
- 737 Cutler, A., Lecumberri, M. L. G., Cooke, M., 2008. Consonant identification in noise by  
738 native and non-native listeners: Effects of local context. *J. Acoust. Soc. Am.* 124 (2),  
739 1264–1268.
- 740 Cutler, A., Weber, A., Smits, R., Cooper, N., 2004. Patterns of English phoneme confusions  
741 by native and non-native listeners. *J. Acoust. Soc. Am.* 116 (6), 3668–3678.
- 742 Delattre, P. C., Liberman, A. M., Cooper, F. S., 1955. Acoustic loci and transitional cues  
743 for consonants. *J. Acoust. Soc. Am.* 27 (4), 769–773.
- 744 Engen, K. J. V., Bradlow, A. R., 2007. Sentence recognition in native- and foreign-language  
745 multi-talker background noise. *J. Acoust. Soc. Am.* 121 (1), 519526.
- 746 Fant, G., 1973. Stops in cv-syllables. In: Fant, G. (Ed.), *Speech Sounds and Features*. MIT,  
747 Cambridge, MA, pp. 110–139.
- 748 Farar, C. L., Reed, C. M., Ito, Y., Durlach, N. I., Delhorne, L. A., Zurek, P. M., Braidia,  
749 L. D., 1987. Spectral-shape discrimination. i. results from normal-hearing listeners for  
750 stationary broadband noises. *J. Acoust. Soc. Am.* 81, 1085–1092.

- 751 Fruchter, D., Sussman, H. M., 1997. The perceptual relevance of locus equations. *J. Acoust.*  
752 *Soc. Am.* 102 (5), 2997–3008.
- 753 Glasberg, B. R., Moore, B. C., 1990. Derivation of auditory filter shapes from notched noise  
754 data. *Hear. Res.* 47 (1-2), 103–138.
- 755 Guerlekian, J. A., 1981. Recognition of the spanish fricatives /s/ and /f/. *J. Acoust. Soc.*  
756 *Am.* 70, 1624–1627.
- 757 Hant, J. J., 2000. A computational model to predict human perception of speech in noise.  
758 Ph.D. dissertation, University of California, Los Angeles, CA.
- 759 Hant, J. J., Alwan, A., 2000. Predicting the perceptual confusion of synthetic plosive con-  
760 sonants in noise. In: *Proceedings of the International Conference on Spoken Language*  
761 *Processing*. Beijing, China, pp. 941–944.
- 762 Hant, J. J., Alwan, A., 2003. A psychoacoustic-masking model to predict the perception of  
763 speech-like stimuli in noise. *Speech Commun.* 40 (3), 291–313.
- 764 Harris, K. S., 1958. Cues for discrimination of American English fricatives in spoken syllables.  
765 *Lang. Speech* 1, 1–17.
- 766 Hedrick, M., Ohde, R. N., 1993. Effect of relative amplitude of frication on perception of  
767 place of articulation. *J. Acoust. Soc. Am.* 94 (4), 2005–2026.
- 768 Hedrick, M. S., Jesteadt, W., 1996. Effect of relative amplitude, presentation level and  
769 vowel duration on perception of voiceless stop consonants by normal and hearing impaired  
770 listeners. *J. Acoust. Soc. Am.* 100 (5), 3398–3407.
- 771 Hedrick, M. S., Schulte, L., Jesteadt, W., 1995. Effect of relative and overall amplitude on  
772 perception of voiceless stop consonants by listeners with normal and impaired hearing. *J.*  
773 *Acoust. Soc. Am.* 98 (3), 1292–1303.
- 774 Hedrick, M. S., Younger, M. S., 2007. Perceptual weighting of stop consonant cues by normal  
775 and impaired listeners in reverberation versus noise. *J. Speech Lang. Hear. Res.* 50 (2),  
776 254–269.

- 777 Heinz, J. M., Stevens, K. N., 1961. On the properties of voiceless fricative consonants. *J.*  
778 *Acoust. Soc. Am.* 33, 589–596.
- 779 Hermansky, H., 1990. Perceptual linear predictive (PLP) analysis of speech. *J. Acoust. Soc.*  
780 *Am.* 87 (4), 1738–1752.
- 781 Jiang, J., Chen, M., Alwan, A., 2006. On the perception of voicing in syllable-initial plosives  
782 in noise. *J. Acoust. Soc. Am.* 119 (2), 1092–1105.
- 783 Jongman, A., 1988. Duration of frication noise required for identification of English fricatives.  
784 *J. Acoust. Soc. Am.* 85 (4), 1718–1725.
- 785 Kewley-Port, D., 1982. Measurement of formant transitions in naturally produced stop  
786 consonant-vowel syllables. *J. Acoust. Soc. Am.* 72 (2), 379–389.
- 787 Lecumberri, M. L. G., Cooke, M., 2006. Effect of masker type on native and non-native  
788 consonant perception in noise. *J. Acoust. Soc. Am.* 119 (4), 2445–2454.
- 789 Levitt, H., 1971. Transformed up-down methods in psychoacoustics. *J. Acoust. Soc. Am.*  
790 49 (2B), 467–477.
- 791 Li, N., Loizou, P. C., 2007. Factors influencing glimpsing of speech in noise. *J. Acoust. Soc.*  
792 *Am.* 122 (2), 1165–1172.
- 793 Liberman, A. M., Delattre, P. C., Cooper, F. S., Gerstman, L. J., 1954. The role of consonant-  
794 vowel transitions in the perception of the stop and nasal consonants. *Psychol. Mono.* 68 (8),  
795 1–13.
- 796 Miller, G. A., Nicely, P. E., 1955. An analysis of perceptual confusions among some English  
797 consonants. *J. Acoust. Soc. Am.* 27 (2), 338–352.
- 798 Nittrouer, S., Wilhelmsen, M., Shapley, K., Bodily, K., Creutz, T., 2003. Two reasons not  
799 to bring your children to cocktail parties. *J. Acoust. Soc. Am.* 113, 2254.
- 800 Ohde, R. N., Stevens, K. N., 1983. Effect of burst amplitude on the perception of stop  
801 consonant place of articulation. *J. Acoust. Soc. Am.* 74, 706–714.

- 802 Parikh, G., Loizou, P. C., 2005. The influence of noise on vowel and consonant cues. *J.*  
803 *Acoust. Soc. Am.* 118 (6), 3874–3888.
- 804 Potter, R. K., Kopp, G. A., Green, H., 1947. *Visible Speech*. Van Nostrand, Princeton, NJ.
- 805 Redford, M. A., Diehl, R. L., 1999. The relative perceptual distinctiveness of initial and final  
806 consonants in CVC syllables. *J. Acoust. Soc. Am.* 106 (3), 1555–1565.
- 807 Shadle, C. H., Mair, S. J., 1996. Quantifying spectral characteristics of fricatives. In: *Pro-*  
808 *ceedings of the International Conference on Spoken Language Processing*. Philadelphia,  
809 PA, pp. 1521–1524.
- 810 Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., Ekelid, M., 1995. Speech recognition  
811 with primarily temporal cues. *Science* 270 (5234), 303–304.
- 812 Simpson, S. A., Cooke, M., 2005. Consonant identification in n-talker babble is a nonmono-  
813 tonic function of n. *J. Acoust. Soc. Am.* 118 (5), 2775–2778.
- 814 Soli, S. D., Arabie, P., 1979. Auditory versus phonetic accounts of observed confusions  
815 between consonant phonemes. *J. Acoust. Soc. Am.* 66 (1), 46–59.
- 816 Stevens, K. N., 1985. Evidence for the role of acoustic boundaries in the perception of  
817 speech sounds. In: Fromkin, V. A. (Ed.), *Phonetic Linguistics: Essays in Honor of Peter*  
818 *Ladefoged*. Academic Press, New York, NY, pp. 243–255.
- 819 Stevens, K. N., 1998. *Acoustic Phonetics*. MIT Press, Cambridge, MA.
- 820 Stevens, K. N., Blumstein, S. E., 1978. Invariant cues for place of articulation in stop con-  
821 sonants. *J. Acoust. Soc. Am.* 64, 1358–1368.
- 822 Stevens, K. N., Manuel, S. Y., Metthies, M., 1999. Revisiting place of articulation measures  
823 for stop consonants: Implications for models of consonant production. In: *Proceedings of*  
824 *the International Congress of Phonetic Sciences*. San Francisco, CA, pp. 1117–1120.
- 825 Strope, B., Alwan, A., 1997. A model of dynamic auditory perception and its application to  
826 robust word recognition. *IEEE Trans. Speech Audio Process.* 5, 451–464.

- 827 Suchato, A., 2004. Classification of stop consonant place of articulation. Ph.D. dissertation,  
828 Massachusetts Institute of Technology, Cambridge, MA.
- 829 Sussman, H. M., Fruchter, D., Cable, A., 1995. Locus equations derived from compensatory  
830 articulation. *J. Acoust. Soc. Am.* 97 (5), 3112–3124.
- 831 Sussman, H. M., Hoemeke, K. A., Ahmed, F. S., 1993. A cross-linguistic investigation of  
832 locus equations as a phonetic descriptor for place of articulation. *J. Acoust. Soc. Am.*  
833 94 (3), 1256–1268.
- 834 Sussman, H. M., McCaffrey, H. A., Matthews, S. A., 1991. An investigation of locus equations  
835 as a source of relational invariance for stop place categorization. *J. Acoust. Soc. Am.* 90 (3),  
836 1309–1325.
- 837 Wang, M. D., Bilger, R. C., 1973. Consonant confusion in noise: A study of perceptual  
838 features. *J. Acoust. Soc. Am.* 54, 1248–1265.
- 839 You, H. Y., 1979. An acoustical and perceptual study of English fricatives. Master thesis,  
840 University of Edmonton, Edmonton, Canada.
- 841 Zue, V., 1976. Acoustic characteristics of stop consonants: A controlled study. Ph.D. disser-  
842 tation, Massachusetts Institute of Technology, Cambridge, MA.

Table 2: Acoustic measurements.  $F_i$  can be F1, F2, or F3. Superscripts  $f$  or  $p$  indicate that the measures were made only for fricatives or plosives, respectively. The “v” letter indicates that measures were made for the vowel spectrum. Those without asterisks were intermediate measures that were used to make relative spectral amplitude measurements.

Name	Description
* $F_i$ b/ $F_i$ e/ $F_i$ s/ $F_i$ D	$F_i$ onset/offset/steady-state frequency/transition duration
* $F_i$ bA/ $F_i$ eA/ $F_i$ sA	$F_i$ onset/offset/steady-state amplitude
* $F_i$ df/ $F_i$ dA	$F_i$ frequency/amplitude change
*votD/bstD <sup>p</sup> /nD <sup>f</sup>	VOT/burst/noise duration
Ahi	Peak amplitude of burst/noise spectrum in high frequencies (female: above 3.5 kHz; male: above 3 kHz)
Av/Av4	Peak amplitude of vowel spectrum at the F1/F4 prominence
A23/A45	Average amplitude of burst/noise spectrum in F2-F3/F4-F5
pA23/pA45	Peak amplitude of burst/noise spectrum in F2-F3/F4-F5
Am/Avm	Average amplitude of burst/vowel onset spectrum at mid-frequencies (3.2-4.8 kHz)
Ans	Average amplitude of the entire noise spectrum
*Ahi-A23	Spectral tilt of the burst/noise
*Av-Ahi	Peak spectral amplitude of burst/noise in high frequencies relative to that of vowel at F1
*Av4-A45	Peak spectral amplitude of vowel at F4 relative to the average spectral amplitude of burst/noise in F4-F5
*Av4-pA45	Peak spectral amplitude of vowel at F4 relative to that of burst/noise in F4-F5
*Av-pA23 <sup>p</sup>	Mid-frequency spectral prominence for plosives
*Am-Avm <sup>p</sup>	Difference between burst and vowel spectral amplitude at mid-frequencies
*Av-Ans <sup>f</sup>	Average spectral amplitude of noise relative to the peak of vowel at F1

Table 3: Percent correct classification (shown as a superscript) of the quiet speech tokens (from all talkers) based on a single acoustic property measured without the addition of the white Gaussian noise.

/ba,da/	/bi,di/	/bu,du/	/pa,ta/	/pi,ti/	/pu,tu/	/va,za/	/vi,zi/	/vu,zu/	/fa,sa/	/fi,si/	/fu,su/
F2b <sup>100</sup>	Av-Ahi <sup>94</sup>	F3df <sup>91</sup>	bstD <sup>97</sup>	Ahi-A23 <sup>94</sup>	Av-Ahi <sup>100</sup>	nD <sup>100</sup>	Av-Ans <sup>100</sup>	Av-Ans <sup>91</sup>	F2df <sup>94</sup>	Av4-pA45 <sup>88</sup>	Av-Ans <sup>100</sup>
F2df <sup>100</sup>	F1bA <sup>84</sup>	Av-Ahi <sup>91</sup>	Ahi-A23 <sup>84</sup>	Av-Ahi <sup>81</sup>	Ahi-A23 <sup>91</sup>	F2df <sup>97</sup>	votD <sup>84</sup>	Av4-pA45 <sup>84</sup>	F3df <sup>94</sup>		F2e <sup>81</sup>
F1b <sup>94</sup>	F2bA <sup>84</sup>	F2b <sup>88</sup>		Av4-pA45 <sup>81</sup>	Av4-pA45 <sup>91</sup>	Av-Ans <sup>94</sup>	nD <sup>84</sup>	F1e <sup>81</sup>	F1b <sup>91</sup>		Av4-pA45 <sup>81</sup>
F3df <sup>94</sup>	Am-Avm <sup>84</sup>	F2e <sup>84</sup>		bstD <sup>81</sup>	F2e <sup>88</sup>	F1b <sup>88</sup>		F3df <sup>81</sup>	Av-Ans <sup>91</sup>		
Av-Ahi <sup>88</sup>	F2dA <sup>84</sup>	F1e <sup>81</sup>			F2b <sup>84</sup>	F2b <sup>81</sup>			F2b <sup>84</sup>		
F2D <sup>84</sup>					Av4-A45 <sup>84</sup>	F2e <sup>81</sup>			F1df <sup>81</sup>		
Av4-pA45 <sup>81</sup>											
bstD <sup>81</sup>											

Table 4: A summary of acoustic features that each yielded 79% or above correct classification of place of articulation. Threshold SNRs are listed beneath each CV pair. Asterisks are added next to the measures that were discussed at the end of Sec. 2.3.

	/ba,da/	/bu,du/	/pu,tu/	/pi,ti/	/bi,di/	/pa,ta/	/fu,su/	/fa,sa/	/va,za/	/fi,si/	/vu,zu/	/vi,zi/
	-4.4	-1.8	0	0.6	4.4	6.0	-5.2	-4.9	-4.5	-3.7	-3.2	-1.4
*F1b	✓							✓	✓			
F1e		✓									✓	
F1df								✓				
F1bA					✓							
*F2b	✓	✓	✓					✓	✓			
F2e		✓	✓				✓		✓			
*F2df	✓							✓	✓			
F2D	✓											
F2bA					✓							
F2dA					✓							
*F3df	✓	✓						✓			✓	
*Av-Ans							✓	✓	✓		✓	✓
*Av-Ahi	✓	✓	✓	✓	✓							
*Av4-pA45	✓		✓	✓			✓			✓	✓	
Av4-A45			✓									
*Ahi-A23			✓	✓		✓						
Am-Avm					✓							
*bstD	✓			✓		✓						
votD												✓
*nD								✓				✓

Table 5: Percent correct judgments as a function of SNR (dB), manner of articulation, voicing, and vowel context (data averaged across all talkers and all listeners).

SNR	/b,d/				/p,t/				/v,z/				/f,s/			
	/a/	/i/	/u/	<b>/a,i,u/</b>	/a/	/i/	/u/	<b>/a,i,u/</b>	/a/	/i/	/u/	<b>/a,i,u/</b>	/a/	/i/	/u/	<b>/a,i,u/</b>
21	100	100	100	<b>100</b>	99	96	99	<b>98</b>	100	98	100	<b>99</b>	100	99	100	<b>100</b>
10	98	92	93	<b>94</b>	81	98	100	<b>93</b>	97	98	100	<b>98</b>	99	100	100	<b>100</b>
5	96	81	98	<b>92</b>	78	88	98	<b>88</b>	97	95	96	<b>96</b>	99	97	96	<b>97</b>
0	87	65	86	<b>79</b>	70	78	79	<b>76</b>	91	84	90	<b>88</b>	95	93	96	<b>95</b>
-5	79	57	64	<b>67</b>	64	64	56	<b>61</b>	78	66	72	<b>72</b>	79	72	80	<b>77</b>
-10	66	50	60	<b>59</b>	54	52	52	<b>53</b>	61	54	52	<b>56</b>	56	51	59	<b>55</b>
-15	51	41	56	<b>49</b>	47	51	48	<b>49</b>	58	52	52	<b>54</b>	42	50	53	<b>48</b>