On the Perception of Voicing for Plosives in Noise

Marcia Chen and Abeer Alwan

Department of Electrical Engineering
University of California, Los Angeles

Abstract

Previous research has shown that the VOT and first formant transition are primary perceptual cues for the voicing distinction for syllable-initial plosives (SLP) in quiet environments. This study seeks to determine which cues are important for the perception of voicing for SLP in the presence of noise. Stimuli for the perceptual experiments consisted of naturally-spoken /CV/ syllables (six plosives in 3 vowel contexts) in varying levels of additive white Gaussian noise. In each experiment, plosives which share the same place of articulation (e.g. /p, b/) were presented to subjects in identification tasks. For each voiced/voiceless pair, a threshold SNR value was calculated. Threshold SNR values were then correlated with measurements of several acoustic parameters of the speech tokens. It was found that the VOT did not appear to influence the perception of voicing in noise as much as the first formant transition.

1. Introduction

Previous studies on the voiced/voiceless distinction in syllable-initial plosives have focused primarily on the VOT. It has been shown that the VOT duration of voiced plosives is significantly shorter than voiceless plosives. Liberman et al. [5] found that the attenuation of the first formant frequency (F1) together with the timing of the onset of voicing (relative to the release) – which was subsequently called the VOT – was a cue for voicing.

The first formant transition has been found to be another acoustic cue for voicing in plosives, of an importance perhaps equal to VOT. According to Stevens and Klatt [10], the VOT boundary between voiced and voiceless consonants isn't stable, and varies depending on the presence or absence of a rapidly changing F1 transition. A rapid spectral change is indicative of voicing.

Lisker [6] conducted further experiments on the F1 transition using synthetic stimuli of {/ga/ and /ka/} and found that while F1 had a significant effect on the voiced/voiceless classification, it was neither equally necessary nor sufficient as the VOT. The VOT duration is also easier to measure. In addition, it wasn't the dynamic quality (rapidly changing) of F1 but the low F1 onset frequency which indicated voicing.

Other cues for voicing include a higher fundamental frequency F0 ([1] and [8]), absence of aspiration [9], a perceptually quieter burst [9], and the presence of a voicebar [4]. Previous literature mostly focused on the /Ca/ context.

This study examines the perception of voicing in the presence of white Gaussian noise. First, measurements of several acoustic features of utterances in a data set were made (in quiet) and analyzed for possible voicing cues. Second, perceptual experiments were conducted using the tokens mixed with varying amounts of background noise. Finally, the acoustic and perceptual data were examined in an attempt to determine which acoustic cues were responsible for the results of the perceptual tests.

2. Methods

2.1. Stimuli

The stimuli used for this study consisted of naturally spoken consonant-vowel utterances (/CV/s) in isolation, comprised of a plosive from the set {/b/, /p/, /d/, /t/, /g/, /k/} followed by a vowel {/a/, /i/, /u/}, as spoken by 2 male and 2 female talkers, 4 repetitions each, of American English. Thus, there were 16 tokens for each of the 18 syllables. The data were sampled at 16 kHz.

2.2. Acoustic Measurements

Temporal and spectral measurements of F1, F2, and F3 were obtained manually from the time waveform, wide-band spectrogram, short-time DFT and LPC spectra using Matlab. Spectral analysis was done by analyzing 20 ms (for male talkers) or 15 ms (for female talkers) frames of speech which were windowed with a Hamming window and overlapped by 2.5 ms. Pre-emphasis was used. The onset of the vowel was defined as the center point of the frame which showed an abrupt change in the waveform and in the spectral features, particularly the introduction of a sharp F1 peak. The end of the transition, 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 frame, and the average rate of change for the next 5 frames was also less than 5 Hz per 2.5 ms [3]. Since locating the transition offset is prone to error, a third point, called the steady-state, was measured at 95 ms after vowel onset. At each of these three points (vowel onset, offset and steady-state), the corresponding time, frequency and amplitude of F1, F2, and F3 were recorded.

The durations and amplitudes of the voicebar, burst, and VOT, and the frequency of F0 at the onset and steady-state of the vowel were obtained manually.

Tokens were classified as voiced or unvoiced based on a single parameter, using different thresholds for each of the nine voiced/unvoiced pairs. Thresholds were chosen such that the highest percentage of correct classification would result. Percentages were therefore calculated using 32 tokens each. Tables 1-4 summarize the results in terms of percent correct voicing classification.

While the presence of a voicebar is almost certainly indicative of a voiced utterance (only 1 in 144 unvoiced tokens exhibited signs of a voicebar), the converse doesn't hold. Only about half of the voiced tokens contained a voicebar, predominantly the tokens produced by the two female talkers. Predictably, voicing classification based solely on the duration of the voicebar did not produce high

percentages of correct classification results (Table 1).

The burst peak amplitude was about 2.3 dB higher on average for voiceless tokens than for their voiced counterparts. However, classification based solely on either burst amplitude or duration did not produce accurate results (Table 1).

	voicebar duration	burst duration	burst peak amp
bapa	71.9%	68.8%	62.5%
data	75.0%	75.0%	65.6%
gaka	65.6%	68.8%	62.5%
bipi	75.0%	62.5%	65.6%
diti	75.0%	65.6%	75.0%
giki	75.0%	56.3%	62.5%
bupu	68.8%	62.5%	59.4%
dutu	75.0%	59.4%	59.4%
guku	78.1%	59.4%	68.8%

Table 1: Percent correct voicing classification using voicebar duration, burst duration, and burst amplitude.

	F0 freq change	VOT duration	VOT peak amp
bapa	68.8%	100.0%	75.0%
data	71.9%	100.0%	87.5%
gaka	65.6%	100.0%	90.6%
bipi	75.0%	100.0%	78.1%
diti	78.1%	100.0%	100.0%
giki	93.8%	100.0%	81.3%
bupu	78.1%	100.0%	78.1%
dutu	78.1%	100.0%	87.5%
guku	62.5%	100.0%	75.0%

Table 2: Percent correct voicing classification using F0 frequency change, VOT duration, and VOT amplitude.

	F1 duration	F1 onset freq	F1 freq change
bapa	75.0%	100.0%	100.0%
data	93.8%	100.0%	100.0%
gaka	81.3%	100.0%	100.0%
bipi	75.0%	90.6%	84.4%
diti	84.4%	68.8%	81.3%
giki	65.6%	62.5%	59.4%
bupu	84.4%	84.4%	75.0%
dutu	68.8%	65.6%	65.6%
guku	71.9%	59.4%	62.5%

Table 3: Percent correct voicing classification using F1 duration, onset frequency, and frequency change.

Voiceless tokens had on average a 15.7 Hz greater F0 frequency change than their voiced counterparts. Voicing classification based solely on the frequency change in F0 using all tokens produced a high percentage of correct results only for /gi-ki/ (Table 2).

The VOT duration proved to be the single best acoustic feature for classification for voicing (Table 2). All tokens could be correctly classified based only on their VOT duration. All tokens with a VOT duration greater than 30 ms were unvoiced, and less than 30 ms were voiced, except for the /gi-ki/ pair, which had a threshold of 40 ms. Most voiced tokens had VOT durations under 20 ms, while unvoiced tokens were usually over 45 ms, and very few tokens were in between.

	F2 duration	F2 onset freq	F2 freq change
bapa	65.6%	65.6%	62.5%
data	81.3%	81.3%	90.6%
gaka	87.5%	100.0%	100.0%
bipi	62.5%	71.9%	90.6%
diti	71.9%	75.0%	84.4%
giki	59.4%	62.5%	71.9%
bupu	62.5%	62.5%	78.1%
dutu	59.4%	62.5%	75.0%
guku	56.3%	71.9%	65.6%

Table 4: Percent correct voicing classification using F2 duration, onset frequency, and frequency change.

The F1 transition was an excellent classifier for voicing only in the /a/ context (Table 3). All /Ca/ tokens could be correctly classified based solely on their F1 onset frequency or F1 frequency change (between onset of the vowel and steady-state). The F1 onset frequency is higher for voiceless tokens than for voiced tokens in the /a/ context. Voiceless tokens usually showed an F1 onset frequency higher than 600 Hz, while voiced tokens were below. For tokens in the /i/ and /u/ contexts, however, the F1 onset frequency was in about the same range for voiced and voiceless. Accordingly, the F1 frequency change is larger for voiced tokens, and smaller for voiceless tokens in the /a/ context, and about the same for voiced and voiceless tokens in the /i/ and /u/ context. In fact, for both voiced and voiceless tokens in the /i/ and /u/ context, the F1 frequency changed relatively little between onset of the vowel and the steady state, usually under 200 Hz.

F2 and F3 measurements were generally poor classifiers for voicing. Notable exceptions were F2 measurements for /da-ta/ and /ga-ka/ (Table 4).

2.3. Perceptual Experiments

Listening experiments were conducted with four paid volunteers: two male and two female, all native talkers of American English aged 18-36 who passed a hearing test and training session.

Experiments were forced choice, with two options: voiced or unvoiced. For each /CV/ pair (for example, /ba/ and /pa/), a series of 64 utterances were played from a randomized list consisting of all the voiced and unvoiced tokens in the set, each token listed twice. The subject was played one token at a time, and asked to label the sound heard as voiced or unvoiced (in explaining the task , the subject was actually asked to identify the sound as – for example – either /ba/ or /pa/), and, if necessary, to guess.

Experiments were conducted at seven different SNR levels: no noise, 10 dB, 5 dB, 0 dB, -5 dB, -10 dB, and –15 dB.

The percentage of correct responses for each /CV/ pair was calculated for each SNR level, and a threshold SNR level corresponding with 79% correct responses was computed for each /CV/ pair from the best-fit sigmoid.

3. Results

Most of the /CV/ pairs had 100% correct responses in the absence of noise. Those which didn't were close to 100%: /bu-pu/ 99.22%, /du-tu/ 99.22%, /ga-ka/ 97.27%, /gi-ki/ 99.61%, and /gu-ku/ 99.61%. Even tokens without a significant F1 transition cue were correctly identified, indicating that the F1 transition is not a necessary cue for voicing. Seemingly, the VOT duration is a sufficient cue in quiet.

The threshold SNR levels corresponding to 79% correct responses for each /CV/ pair are listed in Table 5. /CV/s in the /a/ vowel context appear to be more noise robust than those in the /i/ and /u/ contexts, having threshold SNR levels ranging from –6.9 to –8.3 dB, 2 to 6 dB lower than thresholds for /i/ and /u/. Furthermore, thresholds for /CV/s in the /i/ and /u/ contexts were lower (more robust) for velars (g-k) than labials and alveolars. Velars with /i/ or /u/ had thresholds ranging from –3.8 to –5.2 dB, while labials and alveolars with /i/ or /u/ had thresholds ranging from –1.6 to –3.0 dB.

	all talkers	male only	female only
bapa	-7.4	-7.2	-7.6
data	-8.3	-8.0	-8.6
gaka	-6.9	-6.9	-7.0
bipi	-1.6	-0.9	-2.3
diti	-3.0	-1.0	-4.6
giki	-5.2	-4.7	-5.7
bupu	-2.4	2.3	-4.8
dutu	-2.0	-0.7	-4.5
guku	-3.8	-2.9	-2.1

Table 5: Threshold SNR levels (in dB) computed using data from all talkers, male talkers only, and female talkers only.

Threshold SNR values were also computed with the results separated by gender of the talker. These results are also listed in Table 5. In most cases, the threshold SNRs obtained were approximately the same for both males and females with respect to the same /CV/ pair. Usually, female tokens had slightly better perceptual results than their male counterparts, with one exception /gu-ku/. For three /CV/ pairs, /di-ti/, /bu-pu/, and /du-tu/, female tokens resulted in significantly lower threshold SNRs than their male counterparts, by over 3 dB.

3.1. Correlation between perceptual thresholds and physical measurements

Correlation analysis was applied to the perceptual results and the acoustic measurements. Correlation was first computed using the threshold SNR data from the perceptual experiments, and the difference between the mean values of a measured acoustic feature for the voiced and unvoiced /CV/s in each pair. The correlation coefficients obtained are listed in Table 6.

F1 freq change	0.86	F3 onset freq	0.00
F1 onset freq	0.86	F3 slope	-0.07
F1 slope	0.79	vot pk amp	-0.09
F1 duration	0.78	F0 freq change	-0.14
F2 duration	0.73	voicebar rms amp	-0.38
burst duration	0.65	voicebar duration	-0.40
F2 freq change	0.48	F2 slope	-0.40
F2 onset freq	0.46	F3 amp change	-0.41
vot rms amp	0.23	F2 amp change	-0.44
burst pk amp	0.11	F1 amp change	-0.46
F3 freq change	0.09	F3 duration	-0.62
burst rms amp	0.04	vot duration	-0.85

Table 6: Correlation coefficients computed from correlating the threshold SNR values with difference between the means of the voiced/voiceless measurements.

The F1 transition measurements show the highest correlation, F1 frequency change and F1 onset frequency both having a correlation coefficient of 0.86. F2 transition duration and burst duration also showed relatively high correlation coefficients of 0.73 and 0.65, respectively. VOT duration, on the other hand, shows poor correlation, with a correlation coefficient of -0.85 (a negative correlation coefficient indicates that a larger distance between the means correlated with a worse threshold SNR, the opposite of what is expected).

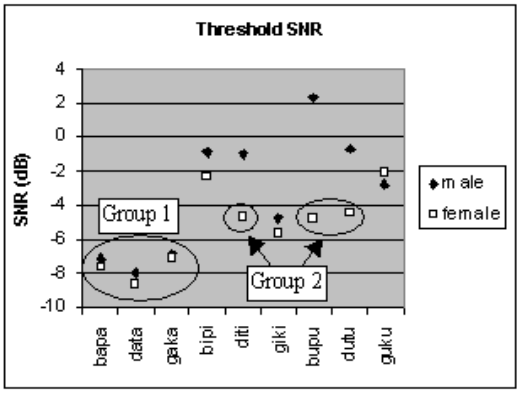

Figure 1: Threshold SNR levels.

Figure 1 shows the threshold SNR values corresponding to 79% correct obtained from the perceptual experiments, separated by the gender of the talkers. Clearly, the points labeled Group 1 – all the /Ca/ points – performed much better than the /Ci/ and /Cu/. This seems to correspond with the F1 transition cue. The F1 transition measurements generally resulted in high correlation coefficients. Large differences between voiced and voiceless F1 onset frequency and F1 frequency change were present in /CV/s in the /a/

context, but not as significant in /CV/s in the /i/ or /u/ context.

The points labeled Group 2 in Figure 1 are /di-ti/, /bu-pu/, and /du-tu/ with female talkers. These three thresholds were significantly lower than their male counterparts. A possible explanation lies again with the F1 transition.

	male	female
bipi	93.8%	68.8%
diti	100.0%	81.3%
giki	75.0%	68.8%
bupu	87.5%	93.8%
dutu	68.8%	87.5%
guku	68.8%	75.0%

Table 7: Percent correct voicing classification using F1 duration, separated by talker gender, for /Ci/s and /Cu/s.

Table 7 lists the percent correct classification obtained using the F1 transition duration, with the data separated by gender of the talker, for the /i/ and /u/ /CV/s. For the female talkers, high percentages were obtained for /di-ti/, /bu-pu/, and /du-tu/ (Group 2). Some of the male /i/ and /u/ /CV/s also obtained high percentage correct classification using the F1 transition duration, but didn't have lower threshold SNRs. This may be due to their shorter F1 transitions.

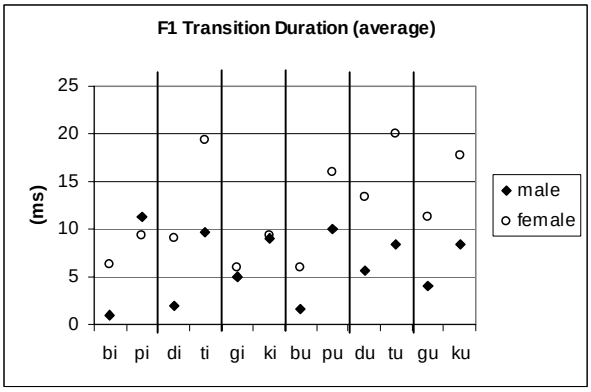

Figure 2: Average male and female F1 transition durations for /i/ and /u/ /CV/s.

Figure 2 shows the average male and average female F1 transition duration for the /i/ and /u/ /CV/ pairs. For /di-ti/, /bu-pu/, /du-tu/, and /gu-ku/, the F1 transition duration is much longer for tokens generated by the females (6-20 ms) then the males (1-10 ms), for both the voiced and the voiceless /CV/. A longer F1 transition makes it more easily detectable, particularly in noise [2]. This is, however, only useful if a voicing cue is present in the F1 transition. Better perceptual results were obtained only if the F1 transition contained distinct differences between the voiced and voiceless /CV/s, and if the F1 transition duration was sufficiently long.

For the remaining /CV/s (not labeled Group 1 or Group 2 in Figure 1), the velars had lower threshold SNRs than the alveolars, which were lower than the labials (although the female /bi-pi/ was lower than the male /di-ti/). Furthermore, /Ci/s had lower thresholds than the corresponding /Cu/s.

One possible explanation for the better performance of the velars (in the absence of a F1 transition cue) is the unusual characteristic of velars of having multiple bursts. Velars thus have longer burst durations, and also tend to have louder bursts. In contrast, labials have the shortest and softest bursts.

4. Summary and Conclusions

The duration of the VOT proved to be the single best acoustic feature for classification for voicing in syllable-initial plosives. All tokens could be correctly classified based only on their VOT duration. The F1 transition was an excellent classifier for voicing only in the /a/ context. All /Ca/ tokens could be correctly classified based solely on their F1 onset frequency or F1 frequency change.

Results from listening experiments seem to indicate that the F1 transition was a more dominant cue for the perception of voicing in noise. /CV/s in the /a/ context had significantly better perceptual results than /i/ and /u/. Further, better perceptual results were obtained only if the F1 transition contained distinct differences between the voiced and voiceless /CV/s, and if the F1 transition duration was relatively long. In the absence of a strong F1 transition cue, velars were more robust than alveolars and labials, possibly because they have longer and louder bursts.

5. Acknowledgments

Work supported in part by an NSF fellowship and NIH grant.

6. References

[1] Haggard, M., Ambler, S., and Callow, M. "Pitch as a Voicing Cue," *JASA*, 47:613-317, 1970.

[2] Hant, J. "A Computational Model to Predict Human Perception of Speech in Noise", Ph.D. dissertation, UCLA Electrical Engineering Department, 2000.

[3] Kewley-Port, D. "Time varying features as correlates of place of articulation in stop consonants," *JASA*, 73:322-335, 1983.

[4] Klatt, D. H., "Voice Onset Time, Frication, and Aspiration in Word-Initial Consonant Clusters", *J. Speech and Hearing Res.*, 18:686-706, 1975.

[5] Liberman, A. M., Delattre, P. C., and Cooper, F. S., "Some Cues for the Distinction Between Voiced and Voiceless Stops in Initial Position," *Language and Speech*, 1:153-167, 1958.

[6] Lisker, L., "Is it VOT or a first-formant transition detector?", *JASA*, 57:1547-1551, 1975.

[7] Lisker, L. and Abramson, A. S. "The voicing dimension: some experiments in comparative phonetics," in *Proc. 6th ICPhS*, Prague 1967 (Academia Publ. House of Czechoslovak Acad. of Sci., Prague, 1970), pp. 563-567.

[8] Ohde, R. N. "Fundamental frequency as an acoustic correlate of stop consonant voicing," *JASA*, 75:224-230.

[9] Repp, B. "Relative Amplitude of Aspiration Noise as a Voicing Cue for Syllable-Initial Stop Consonants," *Language and Speech*, 22:173-189, 1979.

[10] Stevens, K. N. and Klatt, D. H., "Role of formant transitions in the voiced-voiceless distinction for stops," *JASA*, 55:653-659, 1974.