Automated voicing analysis in Praat: statistically equivalent to manual segmentation

Christopher D. Eager

University of Illinois at Urbana-Champaign
eager2@illinois.edu

ABSTRACT

The "fraction of locally unvoiced frames" measure in Praat's Voice Report (VR) is an automated method of obtaining the percentage of a segment which is voiced, but its accuracy has been called into question due to values that change based on scrolling and zooming in Praat's viewing window and don't always match manual voicing segmentation. This study offers statistical support for the accuracy of VR when certain guidelines are followed: (1) use the object window; (2) decrease the time step to increase temporal resolution; and (3) use gender-specific pitch ranges. The closure and frication portions of 277 affricates were analyzed using VR in this way and the results were compared to manual voicing segmentation using paired Wilcoxon tests. The results show that there is no significant difference between VR and manual segmentation, regardless of whether only the closure portion, only the frication portion, or the entire affricate is considered.

Keywords: voice report, automated voice analysis

1. INTRODUCTION

The phonetic implementation of phonological voicing distinctions involves both durational cues and the vibration of the vocal folds themselves. Durational measures which have been studied include the duration of the segment itself [c.f. 11, 15, 19, 21], of the surrounding segments [c.f. 3, 18], and VOT [c.f. 11, 15, 16]). Measures of vocal fold vibration include the percentage or duration of vocal fold vibration within the segment [c.f. 1, 2, 6, 7, 8, 10, 12, 13, 15, 19, 20, 21] and the intensity of the resulting periodic energy [c.f. 10, 13]. The focus of this paper will be the percentage of a segment which is voiced and automated ways of obtaining this measure. There are articulatory means of obtaining this measure such as EGG [20], but the measure can also be obtained from inspection of the acoustic waveform and spectrogram.

While manual segmentation of this measure is common [c.f. 7, 8, 12, 15], it can be time-consuming and is easier for some segments (i.e. stop consonants where there is no added noise to the signal) than for others (i.e. fricative segments where there can be a great amount of aperiodic noise superimposed on the voiced signal). It is thus desirable to have a reliable automated method of extracting this information in these differing environments.

Praat [5] has a function called "Voice Report" (VR) which returns, among other things, the "fraction of locally unvoiced frames" in a segment. If this number is subtracted from 1, the result is the portion of the segment which has vocal fold vibration, ranging from 0 to 1. This and similar automated methods based on the pitch contour have also been used [c.f. 1, 2, 6, 10, 13, 19, 21].

Other methods have been proposed in the literature (i.e. center of gravity [c.f. 2]) and have been reviewed in [10] for three different measures of validity: content-related, criterion-related, and construct-related. Of the voicing measures, [10] found that Praat's Voice Report (VR) and the intensity-based measurements had the closest correlation to trained perceptual judgments of voicing in pre-palatal fricatives taken from data on Argentine Spanish. The study also found that Praat's default pitch settings (75-600 Hz [5]) could lead to chance high-frequency periodicity being interpreted as voicing. This was remedied by lowering the pitch ceiling to 300 Hz. However, the tokens were not manually segmented for voicing duration and so the VR measure's accuracy as a gradient measure could not be tested. [14] called into question whether the "Fraction of locally unvoiced frames" is a reliable measure, since, as the algorithm's creator points out in [4], the values change based on zooming and scrolling in the viewing window. [14] also pointed out that the values obtained from VR varied from manual segmentation. However, the accuracy of VR could not be statistically measured in [14] as the purpose of the article was to demonstrate several voicing measures using only a small sample of six tokens.

2. MOTIVATION FOR THE CURRENT STUDY

To the author's knowledge, an explanation of and way of fixing the scrolling problem and a direct statistical comparison of VR against manual

marking of voicing of segments has not been performed, which could answer questions regarding the reliability of the algorithm.

2.1 An explanation of the zooming/scrolling problem

The changing values of the VR when scrolling/zooming in Praat's viewing window and their deviation from manually segmented voicing percentages can be accounted for by the default time step used in the viewing window, the length of the analysis window, and the duration of the segment being analyzed. If left at default settings, the cross-correlated pitch algorithm in the viewing window in Praat uses a time step of 0.25/PitchFloor seconds and an analysis window equal to one longest pitch period [5]. While this makes sense for the viewing window (a decreased timestep would cause lag when scrolling or zooming), a shorter time step may be beneficial in certain circumstances.

Consider a 20ms sound analyzed in the viewing window. With the pitch floor at the default 75 Hz, the time step will be 3.33ms and the analysis window will be 13.33ms long. For the 20ms segment, the maximum number of frames to analyze will be 6, and most likely will be 5 since the chance is high that the first frame of analysis within the segment will not be at the very beginning of the segment (the position of the frames being determined by the start point of the viewing window, the length of the analysis window, and the time step).

Additionally, the number of frames determines the resolution of the measurement: for *n* frames there will be *n+1* possible values. So for this example, with 6 frames there are 7 possible values: 0%, 16.7%, 33.3%, 50%, 66.7%, 83.3%, and 100%. If the viewing window is shifted such that only 5 frames of analysis fall within the segment, there are 6 possible values: 0%, 20%, 40%, 60%, 80% and 100%. If you analyze and re-analyze the same sound after scrolling left and right and zooming in and out, you will get a set of 11 different possible values from the combination of these sets (0%, 16.7%, 20%, 33.3%, 40%, 50%, 60%, 66.7%, 80%, 83.3%, 100%) which are not equally spaced. If the manual segmentation of voicing (a truly continuous variable) shows that exactly 27.2% of the segment is voiced, it is impossible for the algorithm to return this value, and it will return inconsistent values surrounding this number based on where the center points of the frames fall due to scrolling and zooming. This situation becomes worse as segment duration decreases and better as segment duration increases.

This situation can be remedied, however, by using the object window instead of the viewing window. From the object window, the "To Pitch (cc)…" command allows the user to control the time step. Reducing the time step and running the "To Pitch (cc)…" command on the entire sound file requires more computing time, but may yield more accurate results. With a time step of 0.001 s, the 20ms sound described earlier would have 20 frames of analysis and 21 possible values, greatly increasing its resolution and potential accuracy when compared to manual segmentation. [22] show that gender-specific pitch ranges of 70-250 for males and 100-300 for females yield results statistically equivalent to speaker-specific pitch ranges. By using these pitch ranges, the problem of chance high-frequency periodicity described in [10] is avoided due to lowered pitch ceilings and the analysis window length determined by the pitch floor is set such that it will yield better results for both genders.

2.2 Hypothesis

When temporal resolution is increased by decreasing the time step of the cross-correlated pitch object to 0.001 seconds, gender-specific pitch ranges of 70-250 (males) and 100-300 (females) are used, and a whole sound file is processed from the object window, Praat's Voice Report will return voicing percentages which are statistically equivalent to manual segmentation.

3. METHODOLOGY

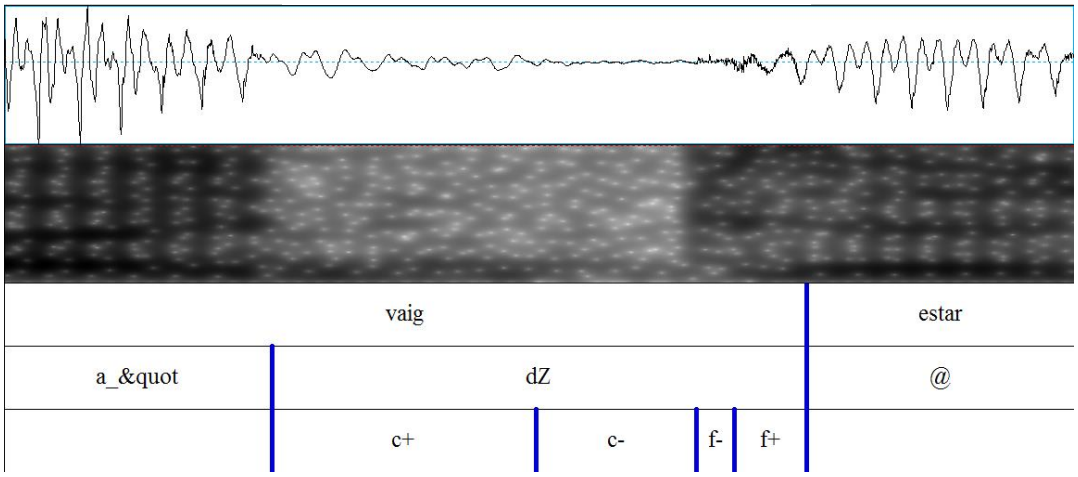
To answer this question, recordings of unscripted Central Catlan in the Glissando corpus [9] were used. According to [9], "the Sony Vegas program running on a PC with a RME Hammerfall HDSP 9652 soundcard, and a Yamaha 02R96 mixer with ADAT MY16AT cards, were used for recordings, at a sampling frequency of 48 KHz" using a AKG C 414 B-ULS directional microphone.

3.1 Segmentation

Affricates serve as an ideal candidate to test the VR algorithm under several different circumstances, as they have a stop portion (with little to no aperiodic energy) and a frication portion (with much more aperiodic energy) which can be analyzed separately and together. Furthermore, only intervocalic segments were used, as the borders between the preceding and following segments can be more accurately determined in this environment, and Catalan affricates enter into a more complex

allophony with fricatives in other environments [1]. In all, 277 intervocalic prepalatal affricates (193 /ʤ/ and 84 /ʧ/) were manually segmented in Praat.

Figure 1: Example of manual segmentation



The segmentation criteria were as follows: (1) the beginning of affricate closure was marked when there was a large drop in intensity and the second formant in the preceding vowel ceased; (2) the border between closure and frication was marked based on the beginning of a burst (if present) and the beginning of uninterrupted aperiodic energy in the waveform if no burst was present; (3) the boundary between the frication portion and the following vowel was marked at the cessation of uninterrupted aperiodic energy; (4) if there was periodicity in the waveform and energy below 300 Hz in the spectrogram during either the closure or frication periods, these portions were further segmented; and (5) each segment was marked with "c" or "f" for closure and frication respectively and with a "+" or a "-" for voiced or voiceless respectively. An example of this segmentation is given in Figure 1.

3.2 Praat script and voice report settings

A Praat script written by the author created a single cross-correlated pitch object for each channel of each conversation (which were several minutes long). The pitch floor was 70 for males, 100 for females. The pitch ceiling was 250 for males, 300 for females. The time step was set to 0.001 seconds. The other settings (maximum number of candidates, silence threshold, voicing threshold, octave cost, octave-jump cost, and voiced/unvoiced cost) were left at the default values recommended in the Praat manual [5]. The script then created a PointProcess object for each Pitch object and obtained the "Fraction of locally unvoiced frames" from the VR for the closure portion, frication portion, and entire affricate for each token and subtracted from 1 to obtain a percent voiced measure. Additionally, the script computed the percentage of the closure, frication and affricate of each token which was manually marked as voiced.

3.3 Statistical analysis

The three VR measures (closure, frication, whole affricate) were compared with the corresponding manually segmented measures via paired Wilcoxon tests in R [17], and the probability density functions of the difference between manual segmentation and VR were plotted for each of the three measures.

4. RESULTS AND DISCUSSION

Table 1 gives the major distributional characteristics of VR, manual segmentation, their difference, and significance tests. Figure 2 shows probability density functions for the difference between Manual and VR for each segment type.

Table 1: Distribution of voicing measures and paired Wilcoxon tests

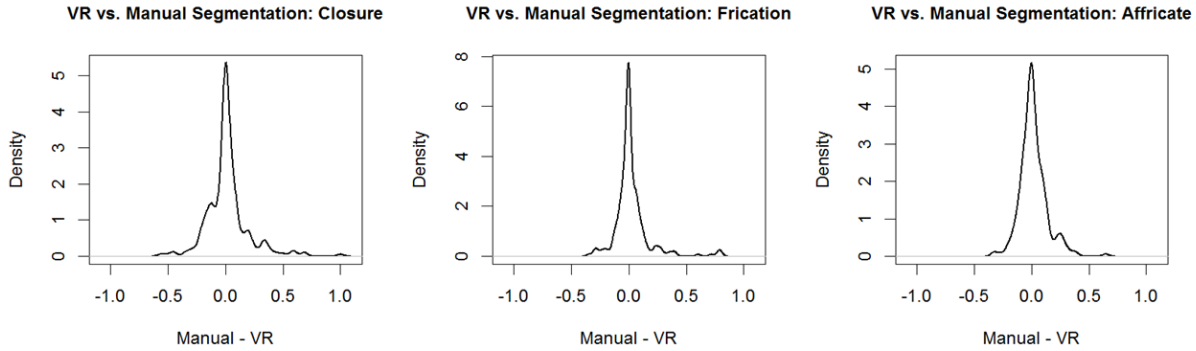
Percent Voiced		Closure	Frication	Affricate
VR	mean	.62	.28	.48
	s.d.	.30	.34	.30
	median	.63	.13	.38
Manual	mean	.64	.31	.49
	s.d.	.32	.37	.32
	median	.61	.14	.40
Manual – VR	mean	.02	.02	.02
	s.d.	.18	.15	.12
	median	0	0	0
Paired Wilcoxon	V	13037	12787	16110
	p	.42	.25	.13

As can be seen in Table 1, under all three segmentations (closure, frication, whole affricate), the mean difference was .02 and the median difference was zero. The reason for the mean being slightly higher than the median is evident in the probability density functions in Figure 2: there are a few strong outliers on the positive side, while there are not strong outliers on the negative side. The median values of zero in combination with the large peaks at zero in each of the probability density functions show that the majority of tokens are not different under VR and manual segmentation, and

following [22] in the setting of gender-specific pitch settings at 70-250 (males) and 100-300 (females) yields comparable results to manual segmentation of voicing by increasing temporal resolution, avoiding the chance high-frequency periodicity described in [10], and guaranteeing appropriate analysis window lengths.

This study contributes to further research by offering guidelines for the use of VR and confirming that these guidelines result in automated measurements which are statistically equivalent to manual segmentation.

Figure 2: Probability density functions for Manual/VR difference.
Negative values indicate overestimated voicing, positive values underestimated



further shows that VR does not have an inherent bias on either the positive or negative side. The paired Wiclxon tests (also given in Table 1) show that the differences are not significant (closure p=.42; frication p=.25; affricate p=.13). This confirms the hypothesis that the VR algorithm is accurate when the guidelines described in this paper are followed.

The number of tokens where voicing is underestimated and the number where voicing is overestimated are about equal, but in this dataset the tokens where voicing was underestimated had a greater absolute difference from the manual segmentation than the tokens where voicing was overestimated. It is possible that alterations to the other settings in the "To Pitch (cc)…" function (specifically the voicing threshold) could reduce the deviance of these outliers. Regardless, the results of this study are strong evidence that, in spite of the occasional outlier, a properly used VR is statistically equivalent to manual voicing segmentation.

5. CONCLUSION

In this study it has been shown that the changing VR values observed when scrolling/zooming in Praat's viewing window [14] can mostly be attributed to the low temporal resolution caused by the high default time step in the viewing window. It was shown that creating Pitch and PointProcess objects from the object window with a time step of 0.001 seconds and

It should be noted that though the problem of temporal resolution is lessened by decreasing the time step, resolution still varies with segment length, and this cannot be fixed.

While VR reliably answers questions related to the amount of a segment that is voiced above the voicing threshold, it does not give information about the gradient strength of voicing. Further research should focus on ways of bringing these various measures together to give a more complete view of voicing phenomena.

6. REFERENCES

[1] Hualde, JI., Eager, CD., Nadeu, M. Catalan prepaltals: Effects of nonphonetic factors on phonetic variation? JIPA. To appear.

[2] Bárkányi, Z., Kiss, Z.G. 2009. Hungarian v. In: Dikken, M. and Vago, R.M. (eds) *Approaches to Hungarian Vol 11: Papers from the 2007 New York Conference*. John Benjamins Publishing. 1-28.

[3] Baum, S. R., & Blumstein, S. E. 1987. Preliminary observations on the use of duration as a cue to syllable-initial fricative consonant voicing in English. *The Journal of the Acoustical Society of America*, 82(3), 1073–1077.

[4] Boersma, P. 1993. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In: *Proceedings of the institute of phonetic sciences* Vol. 17. Amsterdam. 97-110.

[5] Boersma, P., Weenink, D. 2014. Praat: doing phonetics by computer (Version 5.3.77). Amsterdam. Retrieved from http://www.fon.hum.uva.nl/praat/

[6] Davidson, L. 2014. The implementation of voicing in obstruents in American English connected speech. *The Journal of the Acoustical Society of America*, 135(4), 2291–2291.

[7] Delforge, A. M. (2009). *The Rise and Fall of Unstressed Vowel Reduction in the Spanish of Cusco, Peru: A Sociophonetic Study.* Dissertation at University of California, Davis. ERIC.

[8] Docherty, G. J. (1992). *The timing of voicing in British English obstruents*. Foris: New York.

[9] Garrido, J. M., Escudero, D., et al. 2013. Glissando: a corpus for multidisciplinary prosodic studies in Spanish and Catalan. *Language Resources and Evaluation*, 47(4), 945–971.

[10] Gradoville, M. S. 2011. Validity in measurements of fricative voicing: Evidence from Argentine Spanish. In: Alvord, S.M. (ed.) *Selected proceedings of the 5th Conference on Laboratory Approaches to Romance Phonology*. Somerville: Cascadilla Proceedings Project. 59–74.

[11] Green, K. P., Zampini, M. L., & Clarke, C. 1998. The role of preceding closure interval and voice onset time in the perception of voicing: A comparison of English versus Spanish-English bilinguals. *The Journal of the Acoustical Society of America*, 104(3), 1835–1835.

[12] Haggard, M. (1978). The devoicing of voiced fricatives. *Journal of Phonetics* 6. 95-102.

[13] Hualde, J. I., Simonet, M., & Nadeu, M. 2011. Consonant lenition and phonological recategorization. *Laboratory Phonology*, 2(2), 301–329.

[14] Kiss, Z. G. 2013. Measuring acoustic correlates of voicing in stops and fricatives. In Szigetvári, P. (ed.), *VL1xx-Papers in linguistics presented to László Varga on his 70th Birthday* (pp. 289–312). Budapest: Tinta Publishing House.

[15] Lavoie, L. M. 2001. *Consonant strength: phonological patterns and phonetic manifestations*. New York: Garland Pub.

[16] Lisker, L., & Abramson, A. S. 1964. A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, 20(3), 384–422.

[17] R Core Team. 2014. R: A language and environment for statistical computing (Version 3.1.2). Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

[18] Raphael, L. J. 1972. Preceding Vowel Duration as a Cue to the Perception of the Voicing Characteristic of Word-Final Consonants in American English. *The Journal of the Acoustical Society of America*, 51(4B), 1296–1303.

[19] Rivas, M. V. (2006). Does the perception of fricatives correspond to their production? The case of Italian vs. Dutch. MA Thesis University of Amsterdam.

[20] Smith, C. L. 1997. The devoicing of /z/ in American English: effects of local and prosodic context. *Journal of Phonetics*, 25(4), 471–500.

[21] Torreira, F., & Ernestus, M. 2011. Realization of voiceless stops and vowels in conversational French and Spanish. *Laboratory Phonology*, 2(2), 331–353.

[22] Vogel, A. P., Maruff, P., Snyder, P. J., & Mundt, J. C. 2009. Standardization of pitch-range settings in voice acoustic analysis. *Behavior Research Methods*, 41(2), 318–324.