Measuring acoustic correlates of voicing in stops and fricatives

Zoltán G. Kiss

1 Introduction

The aim of this paper is to enumerate the various acoustic correlates of voicing in stops and fricatives and how these correlates can be measured in the phonetic analysis software, Praat (Boersma & Weenink 2012). In addition, the paper aims to asses the validity and reliability of these correlates and measurement methods in comparison with visual inspection of waveforms and spectrograms.

2 Data for the acoustic measurements

The data in this paper make use of recordings of standard Hungarian from one female subject (aged in her early 20s). This recording was part of an independent research into sonorant voicing (Bárkányi & G. Kiss 2012). This paper focuses on the voiceless alveolar stop /t/ and the voiceless alveolar fricative /s/ in three positions: (i) intervocalic, (ii) before /p/, and (iii) before /b/. The test words were *net* 'net' and *szesz* 'alcohol'. These words were embedded in carrier sentences, which the subject had to read out from a monitor at a normal, casual speech tempo. The sentences were recorded in a sound proof cabin onto a laptop computer through an M-Audio MobilePre USB preamplifier, using a Sony ECM-MS907 microphone. As the aim of the paper is to enumerate and illustrate the various voicing measurement methods and not to carry out thorough acoustic and statistical analyses, only one token was chosen for each segment in each position, thus there were altogether six tokens. The test sentences are shown in (1)

and (2), the underlined words constitute the focus of the acoustic measurements of this paper.[1]

(1) a. A netet tartják az évezred találmányának.
'The internet is considered to be the discovery of the millenium.'
b. Egy netprobléma lépett fel.
'An internet problem occurred.' (referred to as *netpro* in the text)
c. A netbeállításokon múlik az egész.
'All depends on the internet settings.' (referred to as *netbe* in the text)

(2) a. A szeszes italok körében jól ismert.
'It is well-known among alcoholic drinks.'
b. A szesz pirosra színezte a főzetet.
'The alcohol coloured the concoction red.' (referred to as *szeszpi* in the text)
c. Sajnos a szesz belefolyt a szemébe.
'Unfortunately, the alcohol got into his eyes.' (referred to as *szeszbe* in the text)

3 Segmentation and manual measurements of material

Segmentation of the test words were carried out manually by visual inspection in Praat with the help of 5-ms-long Gaussian window broadband spectrograms (bandwidth = 260 Hz) and the waveforms of the recordings in the following way. In the case of /t/, the boundary between the preceding vowel and the stop was placed where the formants cease completely. A separate section was marked for the release noise, where release noise is defined as a sudden transient aperiodic burst noise in both the spectrogram and the waveform. The boundary between the release and the following vowel was placed where the burst noise ceases, the formants appear and the periodic wave begins (see figure 1). The boundary between /t/ and the following stops was marked where there was any visual sign of the release of /t/, and simply with the help of listening to the recording (see figure 2). In all the test words used in this paper there was always a short burst noise between the two stops, which made the segmentation relatively straightforward.

[1] The six test sentences, together with their TextGrid files, can be found at tinyurl.com/buogokw

In the case of /s/, the boundary between the preceding vowel and the fricative was placed where the first noisy marks (aperiodicity) appear in the waveform and where the formants cease in the spectrogram. The end of the fricative and the beginning of the following vowel was marked in the position where the waveform does not show aperiodicity anymore, and the periodic wave begins, plus where the formants of the vowel first appear (see figure 3). The boundary between /s/ and the following stops was placed where the waveform becomes almost completely flat, free of any aperiodicity (see figure 4).

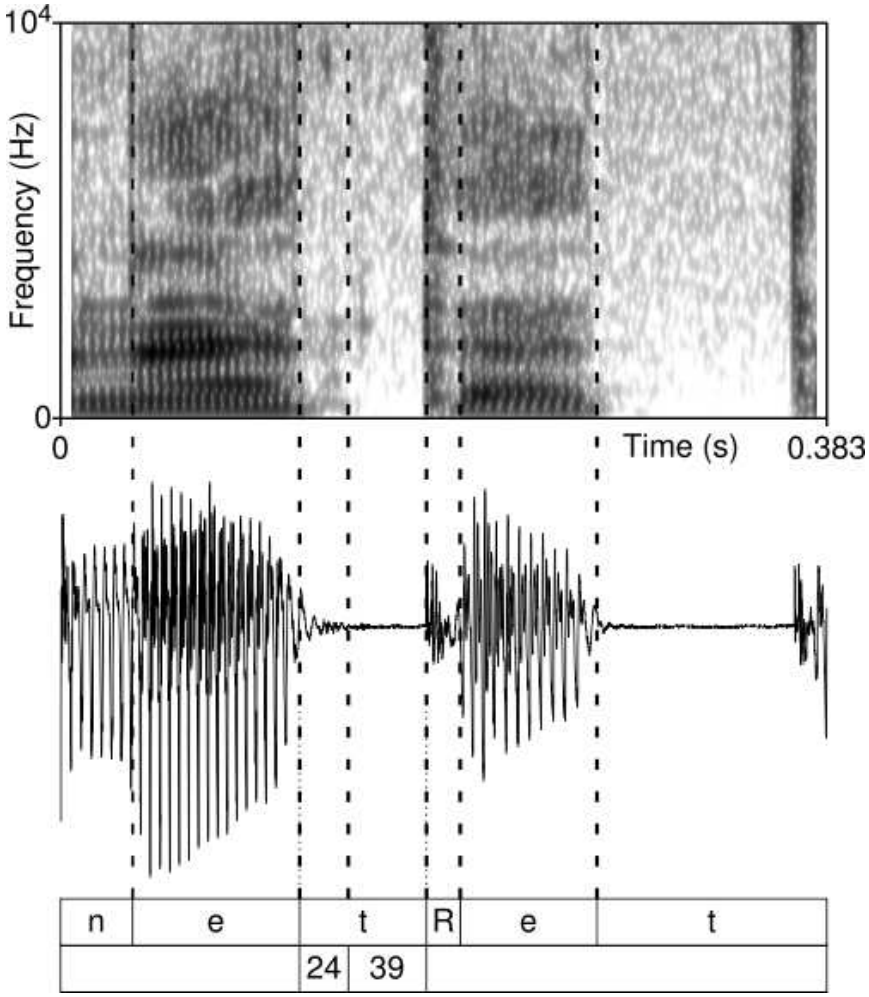
Phonetic voicing can be visually inspected in waveforms by checking for periodic waves, and in spectrograms by looking for energy at low frequencies around 100–300 Hz, ie looking for the presence of the fundamental frequency f_0, the most important acoustic correlate of vocal fold vibration. For the purposes of the current investigation, a separate tier was designated for calculating the percentage of phonetic voicing during the closure portion of /t/ and the constriction phase of /s/. The "voiced" portion was marked wherever signs of periodic waves could be seen in the waveform, this constituted the primary visual cue for vocal fold vibration. As a secondary cue, the low-frequency energy of the spectrogram was also used to mark the boundary of the voiced domain. In the tokens used in this paper, these voiced sections always begin from the boundary between the preceding vowel and the following /t/ or /s/. The length of the voiced and the voiceless domains were measured (in Praat: Query, Get selection length), and the percentage of the *voiceless* section to the whole domain was calculated. If the whole of the segment was voiced, no separate tier was created to measure voicing.

Figure 1 shows the spectrogram and waveform of *netet* 'net-ACC'. Visual inspection indicates that /t/ is only slightly voiced, the vocal fold vibration continues from the preceding vowel, but ceases very rapidly. The percentage of the voiceless section is 62%.

In the case of *netprobléma* 'net problem' (figure 2, left), the situation is very similar to *netet*: voicing continues into the closure phase of /t/ but ceases rapidly. The unvoiced section is 75%. In *netbeállításokon* 'net settings-SUPERESS', the closure portion of /t/ is clearly voiced all through (notice, however, that the following /b/ is actually only very slightly voiced), hence the percentage of the unvoiced section is 0% (see figure 2, right).

In the case of the fricative /s/, it is more difficult to visually locate the voiced sections because the intense frication noise of the fricative at high frequencies (around 8000–10 000 Hz) masks the potential periodicity of the low frequencies. For this reason it is instructive to filter out the high

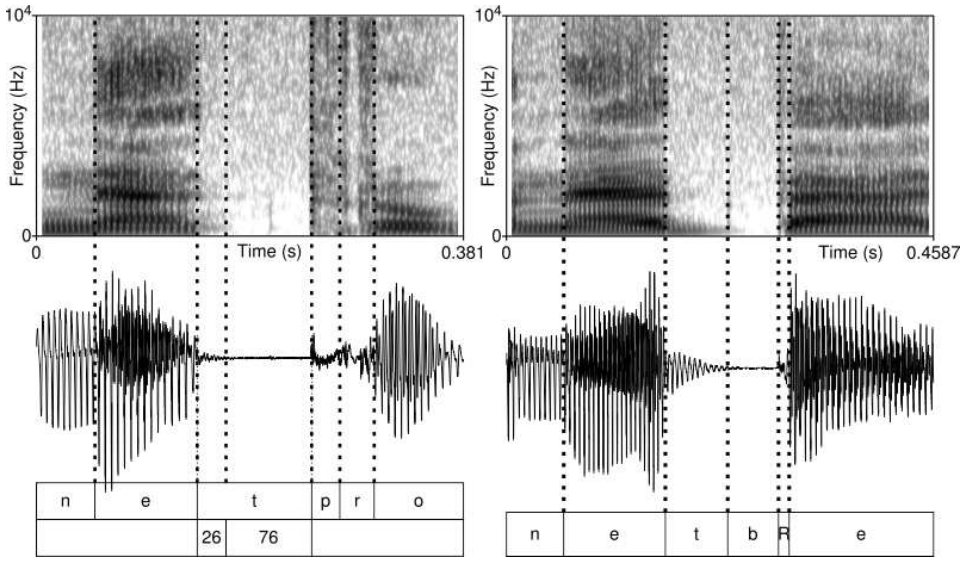
figure 1: Spectrogram and waveform of netet *'net-*ACC*'. The numbers at the bottom indicate the length (in ms) of the voiced+voiceless portion of the closure, 'R' stands for the release noise*



frequencies so that the waveform may only contain the low frequencies. For the purposes of this paper, a low-pass filter was used between 0–500 Hz (with a smoothing of 100 Hz) for the creation of the waveforms to preserve only the low frequencies of the vocal fold vibration (in Praat: Filter Pass Hann band, from frequency = 0 Hz, to frequency = 500 Hz, smoothing = 100 Hz).

Figure 3 shows the spectrogram and filtered waveform of *(sz)eszes* 'alcoholic (drink)'. /s/ is only partially voiced, just like in the case of *netet* and

figure 2: Spectrogram and waveform of netpro(bléma) *'net problem' (left) and* netbe(állításokon) *'net settings-*SUPERESS*' (right). 'R' stands for the release noise. The numbers at the bottom indicate the length (in ms) of the voiced+voiceless portion of the closure, if the closure was voiced throughout, no number is given*



netprobléma, as a result of the continuation of vocal fold vibration from the preceding vowel. The voiced section is 19 ms long, the voiceless portion is 69 ms. The voiceless domain is thus 78% of the whole fricative portion.

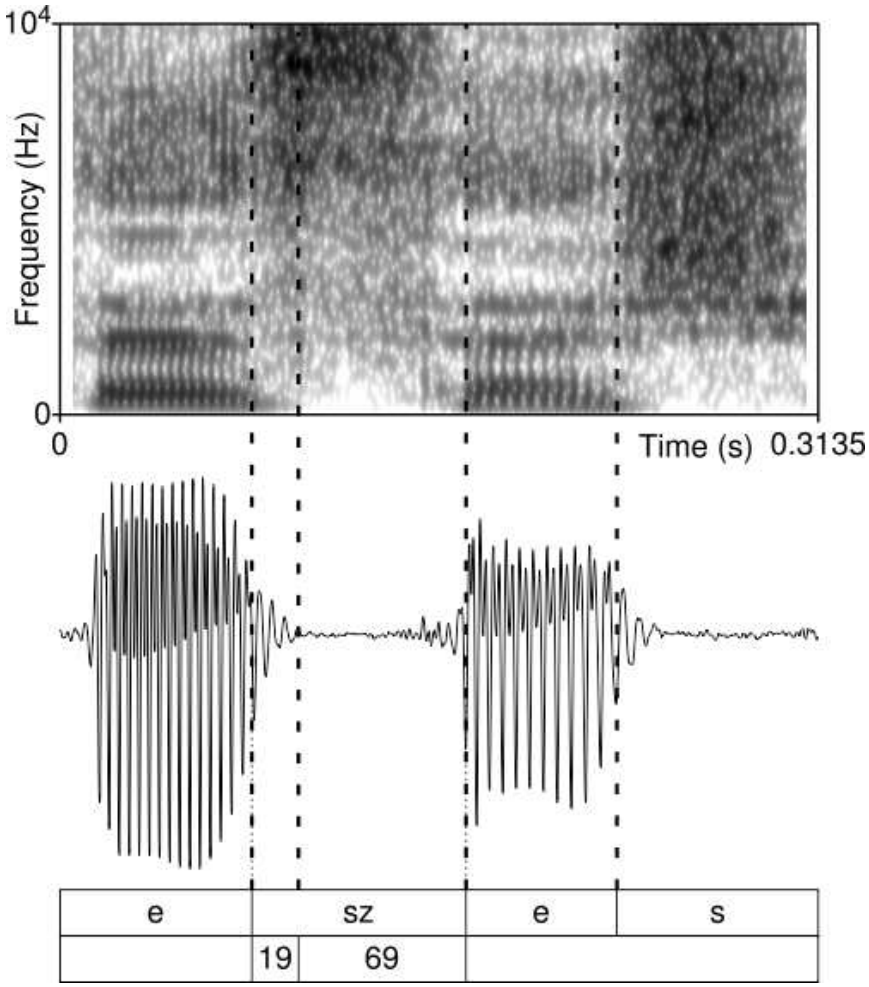
Figure 4 exhibits /s/ before /p/ in *(sz)esz pi(rosra)* 'alcohol red-SUBLAT' (left) and before /b/ in *(sz)esz be(lefolyt)* 'alcohol flowed-ILLAT' (right). In the case of *szeszpi*, the figure shows that there is only a short section of /s/ that is voiced (13 ms). The voiceless section is 38 ms long, which is 75% of the whole fricative domain. /s/ in *szeszbe* is voiced all through the fricative constriction, which is clearly indicated by periodicity of the filtered waveform (in this case, the following /b/ is also fully voiced). Thus the percentage of the voiceless portion is 0%.

To sum up, manual/visual inspection of spectrograms and waveforms tell us that /t/ and /s/ are fully voiced before /b/, but only partially voiced intervocalically and before /p/. The results are summarized in table 1.

table 1: Percent of unvoiced portions manually measured for all six tokens

netet	netpro	netbe	szeszes	szeszpi	szeszbe
62	75	0	78	75	0

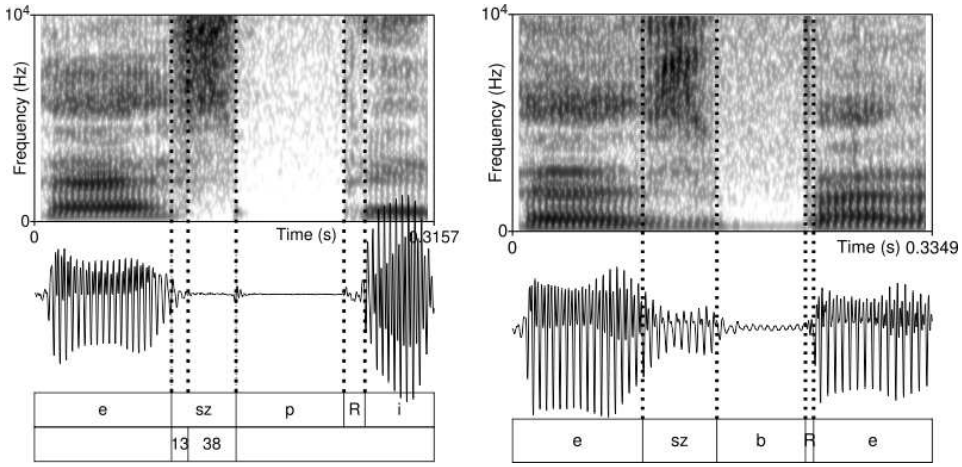
figure 3: Spectrogram and low-pass filtered (0–500 Hz) waveform of (sz)eszes *'alcoholic (drink)'. The numbers at the bottom indicate the length (in ms) of the voiced+voiceless portion of the constriction*



4 Automatic measurement of the acoustic correlates of voicing

In this section I will enumerate the correlates of voicing contrast usually cited in the literature, and show how they can be measured in Praat. I will also compare the results of Praat's measurements with those of the manual/visual measurements in the previous section. Most of these measurements can be carried out automatically with the help of scripts in Praat,

figure 4: Spectrogram and low-pass filtered (0–500 Hz) waveform of (sz)esz pi-(rosra) *'alcohol red-*SUBLAT*' (left) and* (sz)esz be(lefolyt) *'alcohol flowed-*ILLAT*' (right). The numbers at the bottom indicate the length (in ms) of the voiced+voiceless portion of the constriction, if the closure was voiced throughout, no number is given. 'R' stands for the release noise*



and so it is for this reason that we can call them 'automatic' measurements, as opposed to the manual/visual inspection of the spectrograms and waveforms, and reading off data from them in Praat.

4.1 Pulse-based measurements

Perhaps the most well-known method to measure phonetic voicing in Praat is to use its 'Voice Report' (henceforth VR), which measures voicing based on the glottal pulses that it finds in the speech signal. There are many factors that Praat's VR lists but the one that is meant to measure the ratio of unvoiced portions is what is called 'Fraction of locally unvoiced frames'. In this paper, the standard settings were used among the Pitch settings (in the 'advanced' pitch settings, too), except that the 'analysis method' was set to 'cross-correlation', which is recommended by Praat as the optimal method for voice analysis. Also, in the 'Advanced pulses settings', the standard values were set. The measurements were made the following way: all the sound file was read in and displayed in the sound editor window. The sound files contained the focus sounds and their immediate environments as well as additional portions as shown by figures 1–4. With the whole sound file displayed, the /t/ and /s/ domains were selected, and then the VR was called. The results are shown in table 2.

table 2: Results of Praat's VR for "Fraction of locally unvoiced frames" (%) for all six tokens. The results of the manual measurements are also repeated for comparison

	netet	netpro	netbe	szeszes	szeszpi	szeszbe
VR:	74	87	10	82	80	0
manual:	62	75	0	78	75	0

Compared with the manual measurements, the VR results are slightly different. In the case of *netet,* VR judges /t/ to be more voiceless than the manual measurement (VR: 74% vs manual: 62%); similarly, /t/ in *netpro* is reported to be 87% voiceless by VR, whereas it was 75% voiceless when measured manually. The results for *szeszes* and *szeszpi* are very similar as well in both methods, indicating that the focus segments are mostly voiceless. In the 'voicing environments' (before /b/) the results were similar again, with the same results for *szeszbe*: both methods found /s/ to be fully voiced.

There are two problems, however, with the VR method of Praat. One is mentioned in, for example, Gradoville (2011). Using the default settings, Praat's VR function may be "fooled" if the speech signal contains a periodic sound that is not a result of glottal vibration. This can often happen in the case of fricatives. "Phantom" pulses like these can be eliminated if the maximum pitch is set to around 250 Hz.

The second problem concerning Praat's VR is, however, more serious, and greatly affects the validity of the results VR reports and the reliability of the voicing measurement method of Praat. The VR is sensitive to the length of the sound signal that Praat reads in and displays in the sound editor. VR often reports different results if a section is selected for measurement while, say, the whole sound file is read in as opposed to when exactly the *same* section is selected while a different length of the file is read in by Praat. For example, in the case of *netet,* the VR value for unvoiced frames is 74% when the whole sound file was read in and displayed (383 ms long). But, if we cut out or display only the V-/t/-V portion of the signal (218 ms), VR reports that the fraction of unvoiced frames for the *same* /t/ closure section is now 68%! If we further decrease the length of the analysis section displayed (so that only half of the flanking vowels are visible, total length = 157 ms), the value moves up to 74%. If only the /t/ section is visible, the value reported by VR is 65%, which is the closest to the value that was measured manually. Depending on the length of the analysis domain, the values by VR vary slightly or greatly. The values varied in this fashion for all six tokens discussed in this paper.

Praat's manual mentions this potential issue:

> Most of Praat's voice analysis methods start from the glottal pulses that are visible in the SoundEditor window as blue vertical lines through the waveform. [...] If your sound is long, you may have to zoom in in order to see the separate pulses. You may notice that for some sounds, the time location of the pulses can vary when you zoom or scroll. This is because only the visible part of the sound is used for the analysis. The measurement results will also vary slightly when you zoom or scroll.

In my experience, the variation can sometimes be large and not 'slight', and thus seriously question the validity of the results. This issue is linked to the underlying algorithm that is based on frames, and the number of those frames, as discussed by Boersma (1993 : 104), on which Praat's voicing measurements are based:

> Because our method is a short-term analysis method, the analysis is performed for a number of small segments (frames) that are taken from the signal in steps given by the TimeStep parameter (default is 0.01 seconds). For every frame, we look for at most MaximumNumberOfCandidatesPerFrame (default is 4) lag-height pairs that are good candidates for the periodicity of this frame. This number includes the unvoiced candidate, which is always present.

Based on the limited data of this paper, the values remain close to those measured using the whole sound file when the visible section contained *exactly* the domain of /t/ and /s/ (see table 3). It was especially true for /t/; for /s/, the values did not change much. It thus seems advisable to cut these analysis domains and use *only these* for the VR measurements. In order to gain the most valid and reliable results, it is, however, recommended that one resorts to the manual/visual measurement method of finding the ratio of voiceless–voiced portions in a given section of the sound signal, but this, unfortunately, may render the processing of the data set time-consuming.

4.2 Harmonicity

The harmonics-to-noise ratio (HNR) can be used to compare the relation of periodicity and noise in a sound signal, and so it can be a measure for both the degree of voicing and "noisiness", ie how periodic (as opposed to aperiodic) a sound is (see, eg Hamann & Sennema 2005, Bárkányi & Kiss

table 3: Results of Praat's VR for "Fraction of locally unvoiced frames" (%) for all six tokens, when the whole sound file is visible (VR_0) vs when only the analysis section is visible (VR_1). The results of the manual measurements are also repeated for comparison

	netet	netpro	netbe	szeszes	szeszpi	szeszbe
VR_0:	74	87	10	82	80	0
VR_1:	65	77	9	82	81	0
manual:	62	75	0	78	75	0

2009, 2010, Gordeeva & Scobbie 2010, Gradoville 2011). An HNR of 0 dB means that there is equal energy in the periodic and noisy part, while an HNR approximating to 20 dB indicates that almost 100% of the energy of the signal is in the periodic part, hence the sound is a (sonorant) voiced sound (cf the Praat manual, for the technical details, see Boersma 1993). Based on its definition, HNR can only be used reliably as a measure of voicing in the case of fricatives (and other sounds that contain turbulence) and voiced sounds, but not in the acoustic analysis of stops that contain neither periodicity nor noise (ie unreleased voiceless stops).

Praat's VR also lists the HNR, but just like the "unvoiced frames" measure, this value is sensitive to the length of the file read in by Praat. Also, it only lists the HNR of what it judges to be "voiced parts." Thus, the method to be followed is to extract the focus domain as a separate sound object (in our case: the constriction portion of /s/), and create what is called a "harmonicity object," and measure HNR on that object only, using the cross-correlational method (in Praat: Periodicity, To Harmonicity (cc), the standard settings: Time step = 0.01 s, Minimum pitch = 75 Hz, Silence threshold = 0.1, Period per window = 1.0). Calling the Info window, various HNR values are given, in this paper I use the HNR median value. The HNR medians for /s/ are summed up in table 4.

table 4: HNR medians of /s/ in dB (top row). The results of the manual measurements of the unvoiced sections (in %) are also repeated for comparison (bottom row)

szeszes	szeszpi	szeszbe
0.12	−1.11	8.54
78	75	0

The HNR median values in table 4 seem to correspond to the manual voicing measurements: the higher HNR median is indicative of a periodic and somewhat noisy sound (a voiced fricative).

4.3 Centre of gravity

The centre of gravity (CoG), or spectral mean/centroid, corresponds to the average of frequencies over the entire frequency domain weighted by the amplitude (the power spectrum). CoG is then interpreted as the frequency that divides the spectrum into two halves such that the amount of energy in the higher frequencies (the "top" half) is equal to the amount of energy in the lower frequencies (the "bottom" half). If, for example, most energy can be found at higher frequencies, the CoG will have a relatively large value. The most frequent use of CoG is to quantify the *place-of-articulation* differences between fricatives and released stops (Jassem 1979, Forrest et al. 1988, Jongman et al. 2000, Gordon et al. 2002, Ladefoged 2003, Johnson 2003, Machač & Skarnitzl 2005, Boersma & Hamann 2006, 2008).

CoG can also be used to quantify the *manner* of articulation of fricatives, namely whether the spectrum of a fricative contains energy at higher frequencies (relatively high CoG) or at lower frequencies (relatively low CoG). In the former case, the fricative can be considered noisy; in the latter case, the fricative can be characterized with formant structure and the presence of voicing (both skew the energy distribution of the spectrum towards the lower frequencies). This interpretation of CoG was made use of in the differentation between the various labiodental fricatives: voiceless and noisy fricative [f], voiced and noisy fricative [v], and voiced ("narrow") approximant [ʋ] in German and Dutch (Hamann & Sennema 2005), as well as Hungarian and Slovak (Kiss & Bárkányi 2006, Kiss 2007, Bárkányi & Kiss 2009).

Assuming that the place of articulation of the sound to investigate does not change considerably, CoG is expected to be pulled towards the low frequencies in a sound whose production involves excitation by vocal fold vibration (and vowel-like formant structure), in other words, it seems logical that CoG can be used to quantify the differences between voiced and voiceless stops whose place of articulation is (more or less) the same (cf eg Gradoville 2011).

The measurement methods for CoG reported in the literature vary widely, and hence the values reported also vary a lot.[2] In this paper, I assume the standard setting of Praat for CoG, namely, that the spectral mean is computed by weighing the frequencies in the spectrum by their power densities. This is the method used in Forrest et al. (1988), Jongman et al. (2000), Żygis & Hamann (2003), Padgett & Żygis (2003) as well as

[2] Very often, the exact, detailed methods of measuring CoG are unfortunately missing in the descriptions, as in, eg Gradoville (2011).

Boersma & Hamann (2008).[3] In Gradoville (2011), this method is referred to as "COG2" ("power weighing COG"). For the CoG measurements in this paper, the stop closure and the fricative constriction were cut out based on the segment boundaries in Praat's TextGrid (see figures 1–4). I added 20+20 ms to both sides of the analysis domain because of the width of the analysis window we are using for the spectrum analysis (20 ms, see below)—this way the full domain of the sound was preserved for analysis. The recordings were resampled at 22 050 Hz, and low-pass filtered between 0 and 11 000 Hz. An FFT spectral object was created by placing the cursor in the middle of the closure for /t/ and the constriction of /s/, using the following spectrogram settings: window length = 0.02 s (this means that the physical length of the analysis window was 40 ms long) and window shape was Gaussian (all other spectrogram settings were left standard in Praat).[4] To get the CoG, we query the central of gravity, with power = 2.0.

Another spectral moment, the spectral standard deviation (StD), or spectral dispersion, can be used in conjunction with CoG to measure whether the energy is concentrated mainly in a small band around the centre of gravity or spread out over a wide range of frequencies. For this paper, I measured StD on the same spectral object as the CoG, and simply queried Praat to measure it, with power = 2.0. Figure 5 shows the CoG and StD values for our six test tokens.

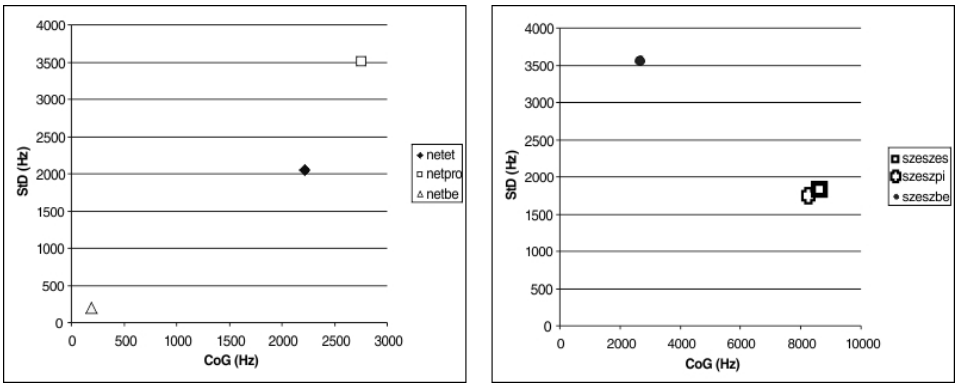
Figure 5 indicates that the voiced tokens *netbe, szeszbe* are separated from the unvoiced tokens. In the case of *netbe,* the CoG is 193 Hz, with a StD of 167 Hz, which means that the energy is concentrated around the low frequencies. In the case of the other two /t/-tokens, the CoG is higher, and the centroid is wider-spread, indicative of an unvoiced, somewhat noisy sound.[5] The CoG values for /t/ thus correspond well with the manual measurements of voicing: /t/ in *netbe* is voiced, while in *netet, netpi...* it is unvoiced. Similar conclusions can be drawn from the values of /s/: its CoG is drawn towards lower values in *szeszbe* (this was manually measured as voiced), with a relatively high dispersion: this sound looks to be

[3] According to Boersma & Hamann (2006, footnote 7), Gordon et al. (2002) and Ladefoged (2003) "apparently used the incorrect method [of measuring CoG...] which weighs the frequencies by their intensity values in dB and is therefore sensitive to arbitrary recording settings."

[4] This is the method employed in Jongman et al. (2000), except that here I only use one analysis window, and not several windows. On the method of using several analysis windows for CoG measurements, see, eg Kiss (2007).

[5] This may well be because of the background noise during the silent closure of the stop.

figure 5: Centre of gravity and standard deviation ("dispersion") for /t/ (left) and /s/ (right) in the three positions



voiced as well as noisy.[6] The other two tokens are very similar with respect to CoG and StD: the values are indicative of a sound whose energy is centred at very high frequencies, and this centroid is relatively wide-spread. We conclude that the CoG/StD values correspond well with the manual measurements of voicing in the case of /s/, too.

4.4 Duration

It has been long observed that there is a correlation between the voicing properties of obstruents and the duration of preceding stressed vowels (or vowel + sonorant sequences), and the duration of closure or constriction of the obstruent (see, among others, House & Fairbanks 1953, Chen 1970, Lehiste 1970, Kluender et al. 1988). More closely, voiceless obstruents as opposed to voiced obstruents are relatively long, and vowels (or vowel + sonorant sequences) before them are relatively short. This has been referred to in the English literature as Pre-Fortis Clipping (Wells 1982, Harris 1994). On the other hand, voiced obstruents are relatively short, while vowel or vowels + sonorants before them are relative long. This is often called Pre-Lenis Lengthening, especially in the American literature (Chomsky & Halle 1968). Jongman (1989) and Kreitman (2008) found that different fricatives seem to have different intrinsic duration. According to the results of Jongman (1989) for instance, English /f/ is longer than any

[6] If we low-pass filter the sound with a cutoff at 3000 Hz for example (thus we get rid of the effect of the high-intensity and high-frequency noise), the CoG and StD of /s/ in *szeszebe* move down radically to 236 Hz and 238 Hz respectively, showing that without the influence of the noise at high frequencies, the sound is voiced.

other fricative. He also found that fricative length varied according to the quality of the neighbouring vowel (he examined CV sequences).

Since speakers typically talk at different rates, the absolute durations of the segments are highly variable, and this is of major concern in acoustic experiments, too. It has been found, however, for English and German for instance (Port & Dalby 1982, Port & Leary 2005) that the ratio of vowel duration to stop closure or fricative constriction remains rather constant in words with the same voicing feature. More closely, the V-to-C duration ratio is generally larger for voiced obstruents than for voiceless obstruents. This ratio is relatively invariant across changes in speaking rate, syllable stress, and segmental context. The durational effects have been given both articulatory and perception-based accounts. Already Chomsky and Halle argue, for instance, that 'the very common lengthening of vowels before voiced obstruents can be explained on the grounds that it requires time to shift from the glottis configuration appropriate for vowels to that appropriate for obstruents' (Chomsky & Halle 1968 : 301; see also Belasco 1953). Stevens et al. (1992) argue, however, that voiced fricatives have shorter frication intervals because they are produced with a smaller glottal abduction gesture, which satisfies the aerodynamic requirements for turbulent noise generation for a relatively short interval in comparison to the large abduction gesture that accompanies voiceless fricatives.

On the other hand, many perception-driven accounts derive the inverse patterning of voiced-voiceless obstruent length and preceding vowel duration as a form of mutual auditory enhancement for the voicing contrast. The idea is that increased vowel duration makes the duration of a following obstruent appear shorter, and conversely that a decrease in vowel duration increases the perceived duration of a following obstruent, and that vowel duration and obstruent duration are therefore integrated into a single percept (Port & Dalby 1982, Port & Leary 2005, Massaro & Cohen 1983, Kluender et al. 1988). This hypothesis has been largely supported by experimental evidence. Thus, listeners pay attention especially to the relative duration of a vowel and the constriction duration of a following obstruent (Javkin 1976, Parker et al. 1986, Kingston & Diehl 1994).

Since this paper focuses on two phonologically voiceless segments /t/ and /s/, and not their contrast with, say, /d/ and /z/ respectively, the durational correlate of voicing will not be further investigated here. It should be noted, however, that based on the discussion above, the expectation is that the vowel-to-consonant durational ratio should not vary significantly because the underlying segments are all voiceless, unless voicing assim-

ilation in *netbe* and *szeszbe* fully neutralizes the contrast of /t/–/d/ and /s/–/z/.[7]

4.5 Intensity

Gradoville proposes that intensity can be used as another correlate of phonetic voicing in the case of fricatives. He warns that a crucial aspect of intensity-based approaches is that "they must be normalized for the recording level and the volume of speech." He suggests two ways for the normalization of intensity: (i) the consonant-to-vowel intensity ratio on both sides of the fricative (when available): the fricative-to-left-vowel intensity ratio (FTL) and the fricative-to-right-vowel intensity ratio (FTR), and (ii) the ratio of low frequency intensity to that of the entire sound (LFT). The logic behind this second normalization approach is that "voicing is only ever going to happen at low frequencies and the frication from the [fricative] is only ever going to happen at higher frequencies. Such a ratio tells us how much of the intensity for the fricative is accounted for by low frequency intensity, which in general will be voicing if the recording is relatively free of extraneous noise" (2011 : 63).

In this paper I measure FTL and the LFT for /s/, following Gradoville's (2011) method: the intensity values were taken in the middle 50 ms of each vowel and consonant. The ratios were taken by subtracting the intensity of the left vowel from that of the fricative. This yielded FTL. LFT was measured by applying a pass Hann band filter on the extracted sound object with the following parameters: from frequency = 0 Hz, to frequency = 900 Hz, smoothing = 100 Hz. The mean intensity was then taken from the unfiltered sound and the filtered sound. The former was subtracted from the latter to yield the LFT. According to Gradoville, "values closer to zero are predicted to be more voiced, whereas values farther from zero are predicted to be more voiceless." Table 5 shows the values for FTL and LFT for /s/.[8]

[7] Results in Bárkányi & G. Kiss (2012) indicate that in intervocalic position, there is no significant difference between /t/–/d/ and /s/–/z/ with respect to the duration of the preceding vowel; however, the duration of the consonants and the vowel-to-consonant duration ratios remained significantly different (the vowels were longer before the underlyingly voiced segments). For the /tp/–/dp/, /sp/–/zp/ and /tb/–/db/, /sb/–/zb/ contrasts, none of the three durational correlates were significantly different, except the duration of the preceding vowel in the case of the /sp/–/zp/ contrast.

[8] Gradoville (2011) does not detail the intensity settings he used in Praat; here I use the standard settings.

table 5: Fricative-to-left-vowel intensity ratio (FTL) and low frequency-to-total intensity (LFT) ratio, both in dB. The results of the manual measurements of the unvoiced sections (in %) are also repeated for comparison (bottom row)

	szeszes	szeszpi	szeszbe
FTL	12.52	14.51	11.46
LFT	−23.43	−16.59	−10.02
manual	78	75	0

This very preliminary result indicates, as hypothesized by Gradoville (2011), that the voiced token has a value closest to 0 dB. Needless to say, for a thorough testing of this voicing parameter, more tokens and statistical analyses are needed.

4.6 Zero-crossing rate

Zero-crossing rate (ZCR) is a measure of the number of times in a given time interval that the amplitude of the speech signals passes through a value of zero (the time-axis), divided by the number of frames. ZCR is frequently used in automatic speech recognition systems to separate the voiced/unvoiced portion of the speech signal (see, among others, Ito & Donaldson 1971, Rowden 1992 : 45–46, Heffernan 2007, Bachu et al. 2008). Phonetic/laboratory phonological application of ZCR can be found in Bombien (2006) for the acoustic analysis of voiceless sonorants in Icelandic and in Gordeeva & Scobbie (2010) for the analysis of preaspiration in voiceless fricatives in Scottish English. According to Gordeeva & Scobbie (2010), in the case of fricatives, the ZCR tends to be higher in voiceless fricatives than in voiced fricatives, thus lower values of ZCR reflect more modal excitation (voicing).

ZCR is usually counted in 10-ms-long frames per second divided by the number of frames (as in Gordeeva & Scobbie 2010). For example, if we measure a sound which is 10 ms long (ie we have only one 10-ms-long interval), and in this 10-ms-long frame the speech signal crosses the time-axis 78 times, the ZCR will be 7800 crossings per second ("crps") divided by 1, thus 7800 crps.

In Praat, the number of zero-crossings can be calculated by creating a PointProcess object from the Sound object (select the Sound object, then Points, ToPointProcess (zeros)). The channel number is 1 if the recording is mono. For the purposes of this paper, both "Include raisers" and "Include fallers" were selected. Querying "Get number of points" is supposed to

give the total number of zero-crossings. This number should be multiplied by 100 to get the per-second value, and divided by the number of 10-ms frames to measure the ZCR.

Figure 6 shows the waveform of the intervocalic fricative in *szeszes* (on the left). The sound file was resampled at 22050 Hz and low-pass filtered between 0 and 11000 Hz, just like in the case of the CoG measurements. The length of the sound is 88 ms, thus there are 8.8 10-ms-long analysis frames. Praat reports that there are 1065 zero-crossings in this domain, that is, 106500 zero crossings per second. The ZCR is 106500 divided by 8.8, amounting to 12102 crossings per second. The ZCR of /s/ in *szeszbe* (on the right of figure 6) is much lower: 3305 crossings per second.

figure 6: Waveforms of the intervocalic fricative in szeszes *(left) and* szeszbe *(right). Their ZCR is 12 170 and 3305 crossings per second, respectively*

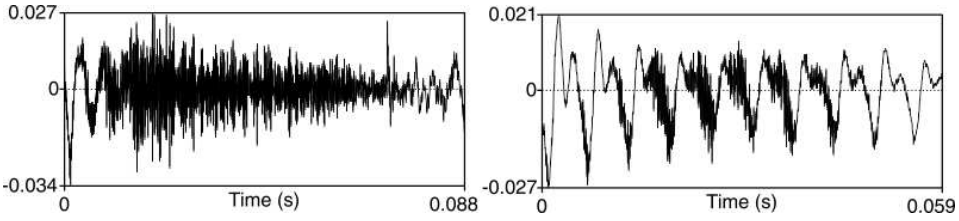


Table 6 displays the ZCR for all six tokens, measured in the closure portion of /t/ and in the constriction portion of /s/.

table 6: Zero-crossing Rate (ZCR) for all six tokens. The results of the manual measurements are also repeated for comparison

	netet	netpro	netbe	szeszes	szeszpi	szeszbe
ZCR:	4031	5235	537	12 102	13 647	3305
manual:	62	75	0	78	75	0

The results go hand-in-hand with the results of the manual measurements, and back the expectation that voiceless fricatives and voiceless occlusions of stops have a higher ZCR than voiced fricatives and voiced stops.

4.7 Low frequency spectral features

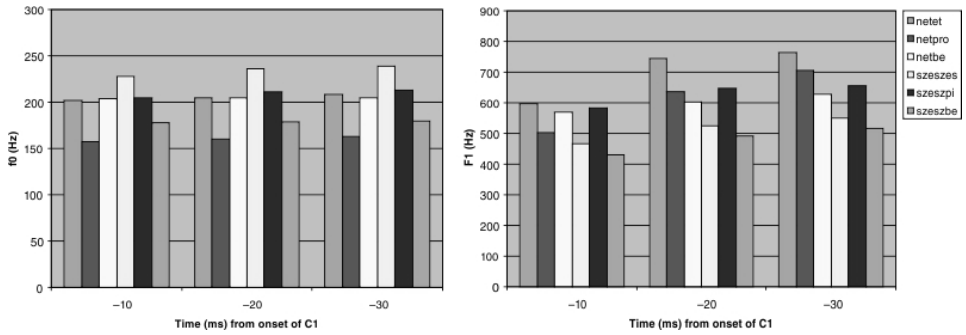
Low frequency spectral features are often cited in the literature to act as acoustic correlates of the underlying voiceless-voiced contrast of stops and fricatives (see Jansen 2004, and the references therein). According to Jansen,

the fundamental frequency f_0 and first formant F1 of a vowel following a voiceless stop and fricative start somewhat higher than the f_0 and F1 of a vowel following a voiced stop and fricative. The low frequency spectral effect is "much stronger following than preceding stops, and it decays over time, so that f_0/F1 differences are normally maximal at the time of voicing onset" (2004 : 52). As Kingston & Diehl (1994) show, the presence of low f_0/F1 values do not imply the presence of voicing, only the reverse seems to hold: voiced stops and fricatives are usually accompanied by f_0/F1 lowering. Jansen (2004 : 141) also warns that the articulatory underpinnings of the low frequency spectral effects are unclear, and so the validity and interpretation of f_0/F1 measurements with respect to voicing are questionable.[9]

The f_0/F1 values in the six tokens of this paper were measured the following way. The original recordings were resampled at 11 025 Hz, and low-pass filtered (pass Hann band between 0–5500 Hz). Measurements were taken at three positions in the pre-consonantal vowel /E/: 10 ms, 20 ms, and 30 ms preceding the onset of the closure of the stop and the constriction of the fricative (based on the boundaries of the TextGrids, see figures 1–4). The values of f_0 and F1 were calculated in Praat with the standard settings.

Figure 7 shows the results of the f_0/F1 measurements.

figure 7: f_0 *(left) and F1 (right) of the preceding vowel in the six tokens, measured at* -10*,* -20 *and* -30 *ms from the onset of the consonant*



[9] In contrast, spectral tilt measurements (such as the difference between the first and second harmonics (H1–H2), or between the first harmonics and first formant (H1–F1)) seem to be reliable correlates of various phonation types, as well as the presence vs lack of aspiration; see, eg Hanson & Chuang (1999), Ladefoged (2003), Gordeeva & Scobbie (2010).

Based on this limited data, the low spectral features seem to be consistent with the view that voicing in a consonant is accompanied by a relatively low f_0/F1 only in the case of the /s/: the lowest value among the /s/-tokens can be found in *szeszbe*, in which /s/ was found to be phonetically voiced by the manual measurement. The values for /t/, especially the f_0 values, do not support the correlation between voicing and low spectral features. The values in the token *netbe* (in which /t/ is phonetically voiced) do not seem to be different from *netet* (where /t/ is mostly voiceless). Also, /t/ in *netpro*, where it was found to be largely voiceless by the manual measurement, has actually the lowest f_0 values. Lastly, the values do not seem to be changing much over time as the measurements move farther away from the consonant onset. We need to stress again, however, that for a thorough investigation, the measurements should be carried out on more tokens, from more speakers, accompanied by vigorous statistical analyses, especially considering the fact that f_0/F1 differences in vowels in the vicinity of voiceless vs voiced consonants rarely seem to exceed 30 Hz (Jansen 2004 : 52), therefore small differences may turn out to be significant.

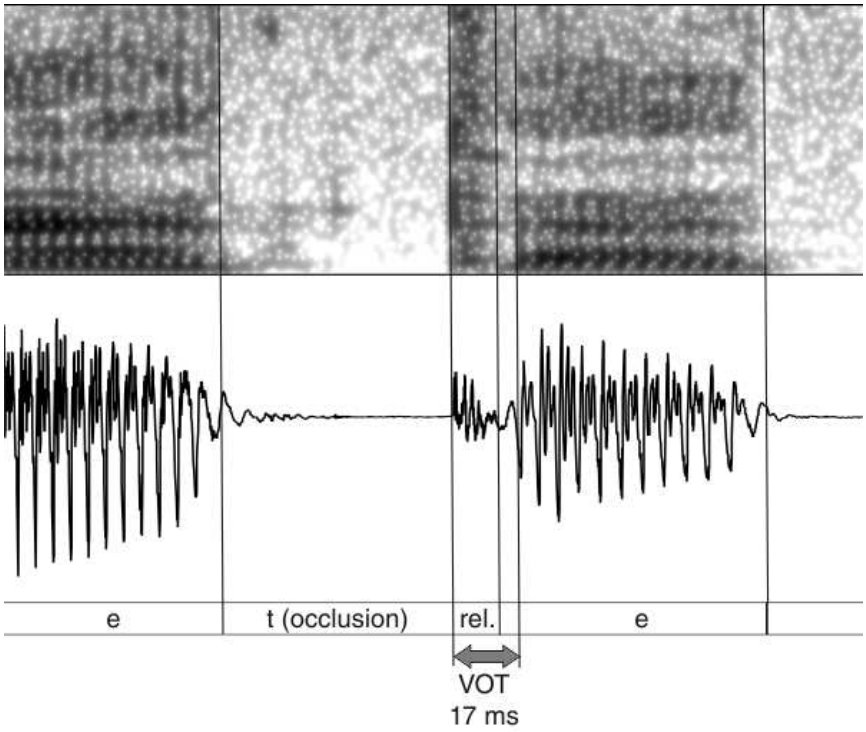
4.8 Voice Onset Time

Voice onset time (VOT), the interval between the release of a stop and the beginning of vocal fold vibration of a following vowel or sonorant consonant, is regarded as one of the most important phonetic correlates of the laryngeal contrast of stops in utterance-/word-initial presonorant (especially prevocalic) position (Lisker & Abramson 1964, Keating 1984, Westbury & Keating 1986, Ladefoged & Maddieson 1996, Jessen 1998, Cho & Ladefoged 1999, Jansen 2004). VOT values allow the analyst to distinguish between three main groups of stops: (i) voiceless aspirated (long lag positive VOT), (ii) voiceless unaspirated (zero or short lag positive VOT), and (iii) voiced unaspirated (negative VOT/prevoiced). "Voicing languages," like Hungarian, contrast negative VOT voiced stops to zero/short lag VOT voiceless stops. According to Jansen (2004 : 46), voicing languages have very similar VOT-targets in medial prevocalic position as in word-initial position, and so it can be used to characterize the difference between voiced vs voiceless stops in this context, too. In preconsonant position, VOT of course cannot be used as a correlate of voicing contrast in the lack of a following vowel or sonorant consonant.

VOT can most reliably be measured manually on the basis of broadband spectrograms and corresponding waveforms: the domain of VOT begins from the appearance of the release burst noise and lasts until the first occurrence of the periodic wave of the following vowel or sonorant

consonant. The analyst also needs to check the presence of vocal fold vibration during the occlusion phase to find possible negative VOT. The only intervocalic stop token among our test words is *netet*, the VOT of the intervocalic /t/ in this word is positive and 17 ms long (see figure 8). According to Keating (1984), the cut-off point between short vs long lag VOT is 35 ms, and so the /t/ in *netet* can be classified as a voiceless stop with short lag VOT.

figure 8: VOT of netet



5 Conclusion

This paper has presented a short overview of the most important acoustic correlates of voiceless and voiced stops and fricatives, and how these correlates can be measured in Praat, using some illustrative examples from Hungarian. The paper also touched upon the assessment of the validity and reliability of these correlates and measurement methods in comparison with manual/visual inspection of waveforms and spectrograms. It has been suggested that Praat's automatic pulse-based measurements (Voice Report) and its harmonics-to-noise ratio measurements can poten-

tially give varied results depending on the sound length read in by the software, and so in this case, the manual/visual measurements may provide more reliable results. The other measurements (especially CoG, duration, intensity, and zero-crossing rate) seem to correlate well with the results of the manual/visual method. It must be stressed that a thorough assessment of the various methods would require more tokens and rigorous statistical analysis, and so the findings of this paper must be treated only as preliminary results.

REFERENCES

Bachu, R. G., S. Kopparthi, B. Adapa and B. D. Barkana. 2008. Separation of voiced and unvoiced using zero crossing rate and energy of the speech signal. Ms, Electrical Engineering Department, School of Engineering, University of Bridgeport.

Bárkányi, Zsuzsanna and Zoltán G. Kiss. 2012. On the border of phonetics and phonology: Sonorant voicing in Hungarian and Slovak. Paper presented at the 20th Manchester Phonology Meeting (mfm20), 24-26 May 2012.

Bárkányi, Zsuzsanna and Zoltán Kiss. 2009. Hungarian /v/: Is it voiced? In: Marcel den Dikken and Robert M. Vago (eds.), *Approaches to Hungarian 11: Papers from the New York Conference.* Amsterdam & Philadelphia: John Benjamins. 1–28.

Bárkányi, Zsuzsanna and Zoltán Kiss. 2010. A phonetic approach to the phonology of *v*: A case study from Hungarian and Slovak. In: Fuchs et al. 2010 : 103–142.

Belasco, Simon. 1953. The influence of articulation of consonants on vowel duration. *Journal of the Acoustical Society of America* 25 : 1015–1016.

Boersma, Paul. 1993. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *Proceedings of the Institute of Phonetic Sciences (University of Amsterdam)* 17 : 97–110.

Boersma, Paul and Silke Hamann. 2006. Sibilant inventories in bidirectional phonology and phonetics. Paper presented at the Third Old World Conference in Phonology (OCP3), 17–19 January 2006, Budapest.

Boersma, Paul and Silke Hamann. 2008. The evolution of auditory dispersion in bidirectional constraint grammars. *Phonology* 25 : 217–270.

Boersma, Paul and David Weenink. 2012. Praat: Doing phonetics by computer. (Version 5.3.23) [Computer program].

Bombien, Lasse. 2006. Voicing alternations in Icelandic—A photoglottographic and acoustic investigation. *Arbeitsberichte des Instituts für Phonetic der Universatät Kiel* 37 : 63–82.

Chen, Matthew. 1970. Vowel length variation as a function of the voicing of the consonant environment. *Phonetica* 22 : 129–159.

Cho, Taehong and Peter Ladefoged. 1999. Variations and universals in VOT: Evidence from 18 endangered languages. *Journal of Phonetics* 27 : 207–222.

Chomsky, Noam and Morris Halle. 1968. *The sound pattern of English.* New York: Harper & Row.

Forrest, Karen, Gary Weismer, Paul Milenkovic, and Ronald N. Dougall. 1988. Statistical analysis of word-initial voiceless obstruents: Preliminary data. *Journal of the Acoustical Society of America* 84 : 115–123.

Fuchs, Susanne, Martine Toda and Marzena Żygis (eds.). 2010. *Turbulent Sounds. An Interdisciplinary guide.* Berlin & New York: De Gruyter Mouton.

Gordeeva, Olga B. and James M. Scobbie. 2010. Preaspiration as a correlate of word-final voice in Scottish English fricatives. In: Fuchs et al. 2010 : 167–207.

Gordon, Matthew, Paul Barthmaier, and Kathy Sands. 2002. A cross-linguistic acoustic study of voiceless fricatives. *Journal of the International Phonetic Association* 32 : 141–174.

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

Hamann, Silke and Anke Sennema. 2005. Acoustic differences between German and Dutch labiodentals. *ZAS Papers in Linguistics* 42 : 33–41.

Hanson, Helen M. and Erika S. Chuang. 1999. Glottal characteristics of male speakers: Acoustic correlates and comparison with female data. *Journal of the Acoustical Society of America* 106 : 1064–1077.

Harris, John. 1994. *English Sound Structure.* Oxford & Cambridge, MA: Blackwell.

Heffernan, Kevin Michael. 2007. Phonetic distinctiveness as a sociolinguistic variable. Doctoral dissertation, University of Toronto.

House, A. and G. Fairbanks. 1953. The influence of consonantal environment upon the secondary acoustical characteristics of vowels. *Journal of the Acoustical Society of America* 25 : 105–113.

Ito, M. and R. Donaldson. 1971. Zero-crossing measurements for analysis and recognition of sounds. *IEEE Transactions on Audio and Electroacoustics* 19 : 235–242.

Jansen, Wouter. 2004. Laryngeal contrast and phonetic voicing: A laboratory phonology approach to English, Hungarian, and Dutch. Doctoral dissertation, Rijksuniversiteit Groningen.

Jassem, Wiktor. 1979. Classification of fricative spectra using statistical discriminant functions. In: Björn Lindblom and Sven Öhman (eds.), *Frontiers of Speech Communication Research.* New York: Academic Press. 189–206.

Javkin, H. 1976. The perceptual basis of vowel duration differences associated with the voiced/voiceless distinction. *Report of the Phonology Laboratory, UC Berkeley* 1 : 78–92.

Jessen, Michael. 1998. *Phonetics and Phonology of Tense and Lax Obstruents in German.* Amsterdam & Philadelphia: John Benjamins.

Johnson, Keith. 2003. *Acoustic and Auditory Phonetics (Second Edition).* Malden, MA & Oxford: Blackwell.

Jongman, Allard. 1989. Duration of frication noise required for identification of English fricatives. *Journal of the Acoustical Society of America* 85 : 1718–1725.

Jongman, Allard, Ratree Wayland, and Serena Wong. 2000. Acoustic characteristics of English fricatives. *Journal of the Acoustical Society of America* 108 : 1252–1263.

Keating, Patricia A. 1984. Phonetic and phonological representation of stop consonant voicing. *Language* 60 : 286–319.

Kingston, John and Randy L. Diehl. 1994. Phonetic knowledge. *Language* 70 : 419–454.

Kiss, Zoltán. 2007. The phonetics-phonology interface: Allophony, assimilation and phonotactics. Doctoral dissertation, Eötvös Loránd University (ELTE), Budapest.

Kiss, Zoltán and Zsuzsanna Bárkányi. 2006. A phonetically-based approach to the phonology of /v/ in Hungarian. *Acta Linguistica Hungarica* 53 : 175–226.

Kluender, Keith R., Randy L. Diehl and Beverly A. Wright. 1988. Vowel length differences before voiced and voiceless consonants: An auditory explanation. *Journal of Phonetics* 16 : 153–169.

Kreitman, Rina. 2008. The phonetics and phonology of onset lusters: The case of Modern Hebrew. Doctoral dissertation, Cornell University.

Ladefoged, Peter. 2003. *Phonetic Data Analysis: An Introduction to Fieldwork and Instrumental Techniques.* Malden, MA & Oxford: Blackwell.

Ladefoged, Peter and Ian Maddieson. 1996. *The Sounds of the World's Languages.* Cambridge MA & Oxford: Blackwell.

Lehiste, Ilse. 1970. *Suprasegmentals.* Cambridge, MA: The MIT Press.

Lisker, Leigh and Arthur Abramson. 1964. A cross-language study of voicing in initial stops: Acoustical measurements. *Word* 20 : 384–422.

Machač, Pavel and Pavel Skarnitzl. 2005. Spectral moments of Czech plosives. Paper presented at the Conference on Turbulences, 13–14 October 2005, Berlin.

Massaro, D. and M. Cohen. 1983. Consonant/vowel ratio: An improbable cue in speech perception. *Perception and Psychophysics* 33 : 502–505.

Padgett, Jaye and Marzena Żygis. 2003. The evolution of sibilants in Polish and Russian. *ZAS Working Papers in Linguistics* 32 : 155–174.

Parker, E., Randy L. Diehl and Keith R. Kluender. 1986. Trading relations in speech and non-speech. *Perception and Psychophysics* 39 : 129–142.

Port, Robert F. and Jonathan Dalby. 1982. C/V ratio as a cue for voicing in English. *Perception and Psychophysics* 2 : 141–152.

Port, Robert F. and Adam P. Leary. 2005. Against formal phonology. *Language* 81 : 927–964.

Rowden, Chris. 1992. Analysis. In: Chris Rowden (ed.), *Speech Processing (The Essex Series in Telecommunication and Information Systems).* Maidenhead: McGraw-Hill. 35–72.

Stevens, Kenneth N., Sheila Blumstein, Laura Glicksman, Martha Burton and Kathleen Kurowski. 1992. Acoustic and perceptual characteristics of voicing in fricatives and fricative clusters. *Journal of the Acoustical Society of America* 91 : 2979–3000.

Wells, John Christopher. 1982. *Accents of English 1–3.* Cambridge: Cambridge University Press.

Westbury, John R. and Patricia A. Keating. 1986. On the naturalness of stop consonant voicing. *Journal of Linguistics* 22 : 145–166.

Żygis, Marzena and Silke Hamann. 2003. Perceptual and acoustic cues of Polish coronal fricatives. In: Maria-Josep Solé, Daniel Recasens and Joachim Romero (eds.), *Proceedings of the 15th International Congress of Phonetic Sciences (Barcelona, 3–9 August 2003).* Barcelona: Causal Productions. 395–398.

Zoltán G. Kiss
gkiss.zoltan@gmail.com
Eötvös Loránd University,
Budapest