SPECTRAL SLOPE PARAMETERS AND DETECTION OF WORD STRESS

J. Volín, J. Zimmermann

Institute of Phonetics, Charles University in Prague

Phonetic research in the past decades has exerted a great effort to describe the principles of relative prominence in speech. The importance of this issue is felt in many areas of both pure and applied phonetics spilling over to the fields of speech technology, where prominence can help to disambiguate words in recognition tasks or make synthesized speech more natural, convincing and unequivocal, the field of didactics, where it can, for instance, help students of foreign languages to acquire better pronunciation, or the field of forensic science, where it can help to establish affective state of the speaker or his/her individual habits in accenting parts of linguistic messages in order to check the identity. These are just a few examples from a large pool of possibilities.

Despite the ingenious studies of Fry (1955, 1958) the questions of the relevant factors of perceptual prominence have not been satisfactorily answered. D. Fry manipulated durations of synthetic segments, their relative fundamental frequency (*F*0), the overall amplitude, and the type of the vowel in the syllabic peak to test perceptually the impact on listeners. His experiment indicated the decisive role of F0 in marking word stress in English. However, these pioneering results were later re-evaluated mainly because the test items all had targets in the nuclear position, where *F*0 must play the key role by definition. Moreover, it was discovered that in the case of such complex sounds as those found in speech the overall sound pressure level (SPL) or other crude correlates of intensity are not directly related to the perceived loudness. The new impetus to the research came among others from the studies by Sluijter et al. (1996, 1997), who attracted attention to the spectral balance of sonorous nuclei of Dutch syllables. They demonstrated that the earlier proposals to use spectral properties to determine vocal effort were correct. Listeners apparently estimate the intended prominence on syllables by the vocal effort assessment, information of which they draw from the spectral properties of the sonorous parts of the speech signal. These suggestions were later corroborated by several scientists for other languages (e.g., Campbell & Mokhtari, 2003 for Japanese; Prieto & Ortega-Llebaria, 2006 for Spanish; Tamburini, 2006 or Plag et al., 2011 for English, etc). It was shown that more prominent syllables exhibit greater spectral balance, i.e., smaller spectral tilt, than less prominent syllables. In other words, the gradient (slope) of the line connecting spectral peaks is steeper in syllables that are perceived as less prominent. It was, nonetheless, also shown that the specifics of the situation are different in each language or even an accent of a language (e.g., Gordeeva, 2006 for Scottish English). Since the above mentioned languages are typologically different from Czech, we decided to ascertain whether the spectral tilt plays any significant role in prominence patterns produced by Czech speakers.

One of the major problems is that there is no single established method of quantifying the spectral slope. One of the early attempts to provide an index of spectral balance is that of Britta Hammamberg's and her colleagues who used the difference between the energy peaks (maxima) in 0 – 2 kHz and 2 – 5 kHz frequency bands (Hammamberg et al., 1980: 448). Various modifications were later suggested to this method. The so-called α measure is based on the ratio between the sound energy above and below 1000 Hz (e.g., Sundberg & Nordenberg, 2006). In their overview, Hanson et al. (2001) discuss several further measures, each with its advantages and disadvantages in the context of various tasks. It has to be remembered, however, that all of the above-mentioned measures were used either for disordered speech or general LTAS characteristics of utterances. Banse and Scherer (1996) tested several spectral characteristics including Hammamberg's index to determine affective states (fourteen different emotions, attitudes or affective stances). Yet their measurements were also taken for a sentence as a unit, albeit a short one. Our focus, on the other hand, is the attribute of a single syllabic nucleus, which usually spans over about 70 to 90 milliseconds in average speaking rates. Also, the syllabic nucleus is normally sonorous, which in terms of energy means quite substantial disparity to consonantal parts of syllables. It follows that we looked for adjustments to the existing approaches. For our experiment we built a program tool using the MATLAB platform including its Signal Processing Toolbox. The cycle of signal processing was planned as displayed in Figure 1.

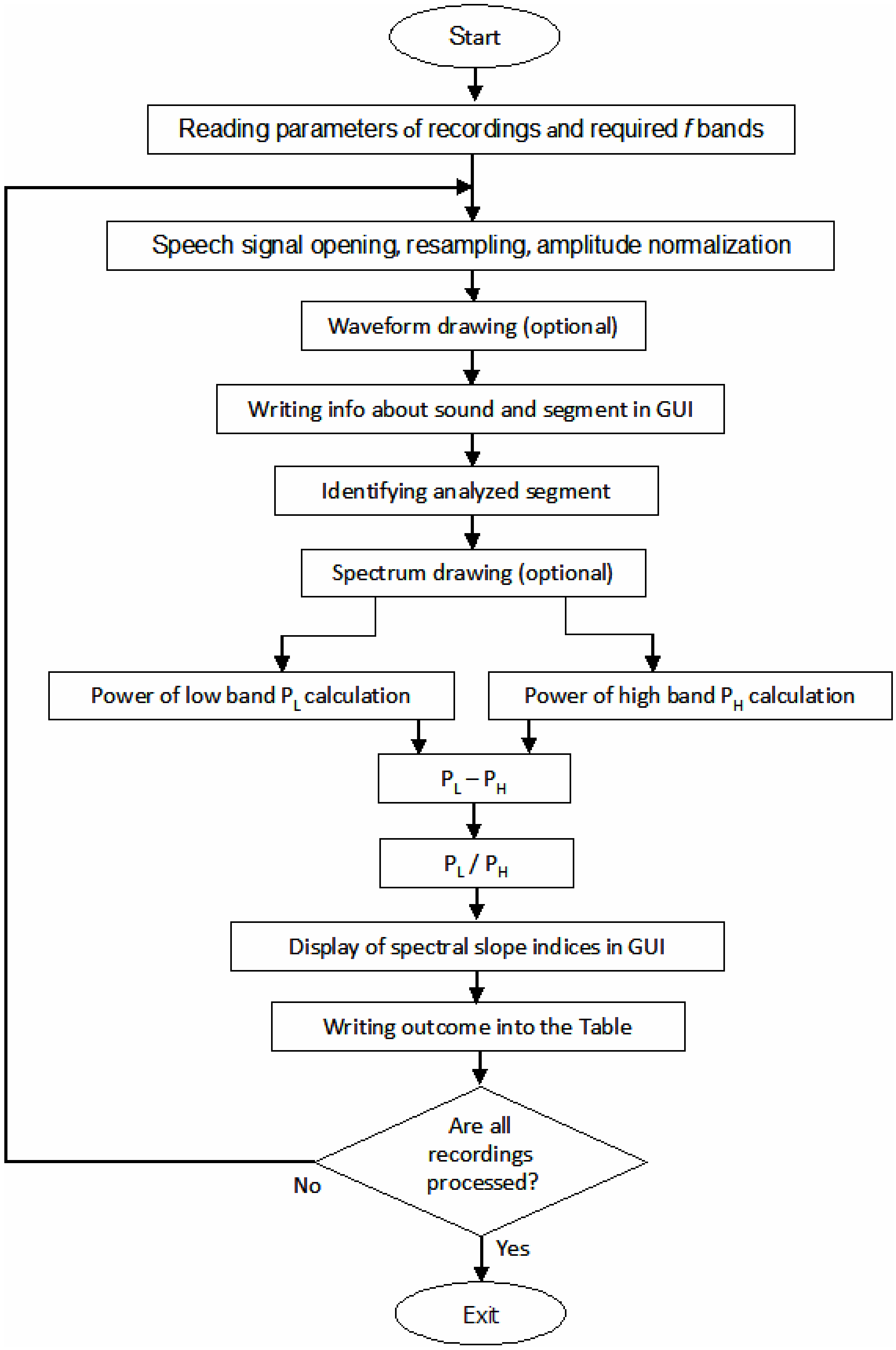
Start
Reading parameters of recordings and required *f* bands
Speech signal opening, resampling, amplitude normalization
Waveform drawing (optional)
Writing info about sound and segment in GUI
Identifying analyzed segment
Spectrum drawing (optional)
Power of low band P_L calculation
Power of high band P_H calculation
$P_L - P_H$
P_L / P_H
Display of spectral slope indices in GUI
Writing outcome into the Table
Are all recordings processed?
No
Yes
Exit

Figure 1: Diagram of the processing cycle of the designed tool

The tool is controlled through GUI for easier qualitative insight into the search for optimal parameters. Before finalizing our tool for the core experiments, we tested four methods of signal power calculations. The first method was based on filtering the source sound leaving two bands (high and low) and calculating the power in each of them in the time domain with the formula

$$P = \frac{\sum |x(n)|^2}{N},$$

where $x(n)$ are individual samples of the signal and N is the number of the samples. The second method was based on determining the power in the frequency domain. This method comprised three consequent steps. First, the specification of the spectrum calculation was set – in our case FFT with the Hamming window of the length derived from the signal properties. Next, the power density of the target segment was determined. Finally, the mean power was calculated. These steps did not require formulae as they were accomplish through MATLAB commands from the inventory in the Signal Processing Toolbox. In the third method, the power in both frequency bands was also determined in the frequency domain. The difference from the previous method was in that the signal was not filtered but the frequency intervals were entered directly in the function for power calculation. The fourth method was based on calculations of relative acoustic energy with the RMS function (i.e., root mean square):

$$RMS = \sqrt{\frac{1}{n}\sum_{n}\left(|x(k)|^2\right)}$$

which likens the speech signal to a uniform signal of constant amplitude with equal power. In other words, the sequence of samples in the time frame is replaced with just one parameter, *RMS*, that directly correlates with the energy of the signal. The relationship between the *RMS* of a frequency band and the power *P* of that band can be expressed by the formula

$RMS = 10*\log10(P/2)$.

We used this formula to transform the results of all four methods in order to compare them mutually. The comparison is displayed in Table 1 for the bands of 800-1200 Hz and 2300-3000 Hz of a vocalic element in one of our recordings. It is obvious that the four methods lead to non-identical results. On the other hand, the outcomes do not differ in any disconcerting manner. For our further experimenting the second method was chosen for its relative simplicity and straightforwardness.

Table 1: COMPARISON OF THE OUTCOMES OF FOUR METHODS OF CALCULATION

Method	**Power [dB/Hz]**		**RMS [dB]**	
1.	**P_L**	0.0081		-23.925
	P_H	1.83e-6		-60.386
2.	**P_L**	0.0094		-23.279
	P_H	1.28e-6		-61.938
3.	**P_L**	0.0067		-24.750
	P_H	6.97e-7		-64.578
4.			**B1**	-27.500
			B2	-62.872

Three male speakers were asked to pronounce 45 three-syllable sequences (pseudowords) with five Czech short vowel [i], [e], [a], [o], and [u], and three options for the stress placement: initial, medial, final. Consonants in syllabic onsets were [h], [t], and [m]. The recording procedure resulted in the set of 405 syllables. A subset of 135 syllables was further modified by equalizing the overall SPL of the vowels in the three-syllable word to –6 dB. The whole set 540 syllables was subjected to a series of analyses with the aim to (a) verify if there is any spectral slope difference between stressed and unstressed syllables produced by Czech speakers, and if yes, then (b) find the spectral slope measure that differentiates well between stressed and unstressed syllables. Unlike all the previous methods, we decided to exclude the band of the second vocalic formant (*F*2). This is because we hypothesized that the band of *F*2 is used primarily for marking the identity of the vowel and should be perhaps slightly less charged with the prosodic functions in the utterance. Similarly, the *F*0 spectral band was excluded, as it is too robust compared with other contributors to the spectral shape. Seven different settings were tested and their effectiveness was estimated from the *F* parameter of three way analysis of variance (ANOVA) with energy ratio as the dependent variable (energy difference proved much less sensitive), and STRESS, SPEAKER, and VOWEL as independent variables or factors. The outcomes are displayed in Table 2.

Table 2: COMPARISON OF SEVEN DIFFERENT SETTINGS FOR SPECTRAL BANDS

	LB from	*LB to*	*HP from*	*HP to*	*F*
Analysis 1	350	1100	1300	4000	30.0
Analysis 2	400	1100	2300	3000	28.1
Analysis 3	350	1100	2300	4000	44.6
Analysis 4	300	1100	2300	4000	34.1
Analysis 5	350	1000	2300	4000	44.1
Analysis 6	350	1100	2450	4000	44.0
Analysis 7	350	1100	2300	5500	47.1

The last setting was the most successful one. Figure 2 depicts some of the ANOVA results, namely the main effect of the STRESS factor and the interaction between the VOWEL and STRESS. Stressed syllabic nuclei consistently manifested smaller spectral slope (i.e., greater spectral balance) than the unstressed ones (S vs. U in the figure). It is obvious that different Czech vowels produce different spectral slopes but all of them preserve the effect of prominence. It also emerged that each of the speakers produced different results – the effect was greater than the mean displayed in Figure 2 for one of them and smaller for another. Clearly, biological predispositions and personal habits imprint on the phenomenon. This fact supports the idea of forensic applicability suggested above in the first paragraph of this report. On the other hand, we tested separately the factors of syllable position and the consonantal onset (together with the potential influence of the normalized SPLs in the word). None of these secondary factors appeared to have any significant effect.

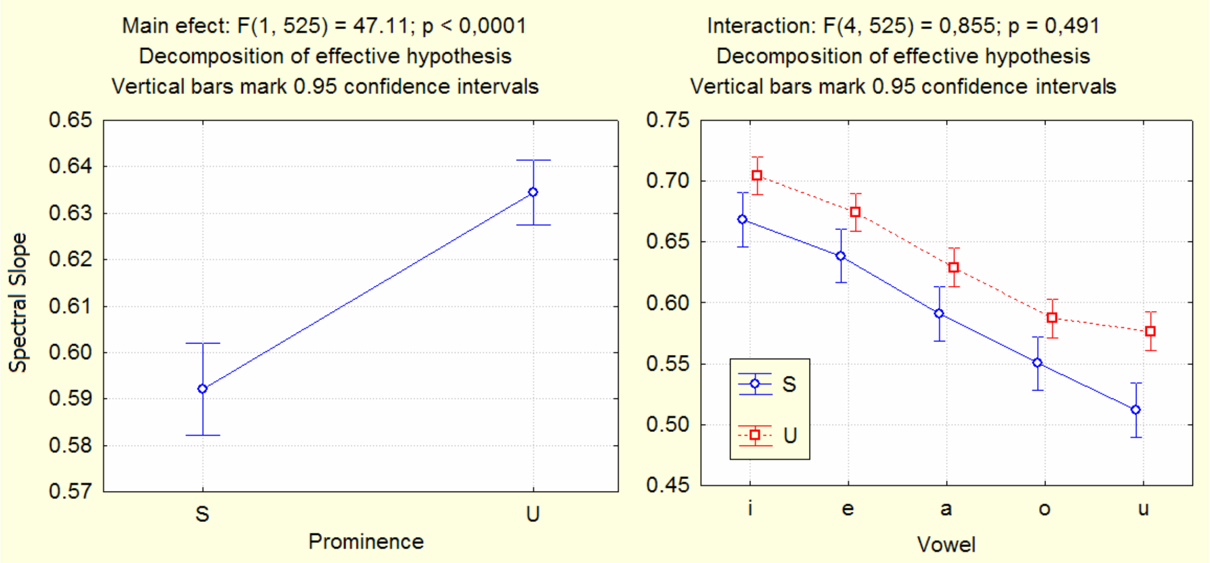

Figure 2: The main effect of STRESS on spectral slope and the interaction of vowel types and stress. Stressed vowels are marked S, the unstressed ones U.

To summarize then, the results of the experiment reveal several important facts. First, our Czech speakers produced stressed syllables with significant differences in spectral tilt. Second, the best way to quantify the spectral properties was to use ratios rather than differences and sums of energy rather than maxima. Moreover, it seems beneficial to skip the band of the second vocalic formant (*F*2) and the area of the fundamental frequency (*F*0) fluctuations: out of several candidates the best results were achieved with the bands of 350-1100 Hz vs. 2300-5500 Hz. Third, the position of the syllable in the word does not play a significant role and neither does the onset consonant. The modified items with artificially equalized SPL in vowels still retained some of the spectral properties sufficient enough to differentiate between stressed and unstressed syllables. On the other hand, each of the three speakers participating in the experiment behaved differently and similarly each of the five Czech short vowels produced distinct result. Clearly, different models have to be built for individual speakers as well as for individual vowels. That is our task for the future research, in which we also intend to use longer samples of more natural speech production. For this purpose, an option was added to our MATLAB tool to switch of the waveform and spectrogram drawings, as these would slow down the processing of larger batches of data. At the same time, however, we keep the option to freeze the operation on the batch of recordings at any time to scrutinize individual cases of syllabic nuclei.

Acknowledgement: This research was supported by the MSM0021620825 grant of the Czech Ministry of Education.

References:

Banse, R. a Scherer, K.R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology* ***70/3***, pp. 614-636.

Campbell, N., Mokhtari, P. (2003). Voice Quality: the 4th prosodic dimension, In*: Proceedings of the 15th ICPhS*, pp. 2417–2420. Barcelona: IPA & UAB.

Fry, D.B.(1955).Duration intensity as physical correlates of linguistic stress. *Journal of the Acoustical Society of America* ***32***, pp. 765–769.

Fry, D. B. (1958). Experiments in the perception of stress. *Language and Speech* ***1***, pp. 126–152.

Gordeeva, O. (2006). Interaction between the Scottish English system of prominence and vowel length. In: *Speech Prosody 2006*, Dresden: TUD.

Hammarberg, B., Fritzell, B., Gauffin, J., Sundberg, J., Wedin, L. (1980). Perceptual and acoustic correlates of abnormal voice qualities", *Acta Otolaryngologica* ***90***, 441-451,

Hanson, H.M., Stevens, K.N., Kuo, H-K.J, Chen, M.Y, Slifka, J. (2001). Towards models of phonation. *Journal of Phonetics* ***29***, pp. 451–480.

Plag, I., Kunter, G., Schramm, M. (2011). Acoustic correlates of primary and secondary stress in North American English. *Journal of Phonetics* ***39***, pp. 362–374.

Prieto, P., Ortega-Llebaria, M. (2006). Stress and accent in Catalan Spanish: Patterns of duration, vowel quality, overall intensity and spectral balance. In: *Speech Prosody 2006*, Dresden: TUD.

Sluijter, A., van Heuven, V. (1996). Spectral balance as an acoustic correlate of linguistic stress. *Journal of the Acoustical Society of America* ***100/4***, pp. 2471–2485.

Sluijter, A., van Heuven, V., Pacilly, J. (1997). Spectral balance as a cue in the perception of linguistic stress. *Journal of the Acoustical Society of America* ***101/1***, pp. 503–513.

Sundberg, J., Nordenberg, M. (2006). Effects of vocal loudness variation on spectrum balance as reflected by the alpha measure of long-term-average spectra of speech. *Journal of the Acoustical Society of America* ***120/1***, pp. 453–457.

Tamburini, F. (2006). Reliable prominence identification in English spontaneous speech. In: *Speech Prosody 2006*, Dresden: TUD.