2017 ACTA UNIVERSITATIS CAROLINAE PAG. 35–44
PHILOLOGICA 3 / PHONETICA PRAGENSIA

TEMPORAL VARIABILITY OF FUNDAMENTAL FREQUENCY CONTOURS

ROBIN HRUŠKA AND TOMÁŠ BOŘIL

ABSTRACT

Intonation is one of the means of performing a speech style. Thus, observing pitch variation in an utterance may be a clue to identifying speech style. We design a cumulative slope (CS) index based upon the amount of pitch variation in a measured F0 contour and the duration of that contour. The more pitch changes there are and the greater their frequency range is, the greater the CS index is. This is confirmed by an experiment we conduct: the CS index of utterances with expressive intonation is higher than that of utterances with neutral intonation, and for utterances with neutral intonation the CS index is higher than for utterances with monotonous or flat intonation. However, as there is a great variability between speakers, the CS index as defined currently, cannot be used to universally differentiate between the styles. Results obtained using automatic voice activity detection (VAD) are close to those obtained with manual VAD and thus the extraction of CS index can be reliably automatized.

Key words: fundamental frequency, melody of speech, stylization, variability of pitch contours

1. Introduction

Intonation is a prosodic feature intrinsic to any spoken text. In a broad sense, intonation is a complex function of speech melody, dynamics and rhythm. In this paper, we focus solely on the melodic component since it is the most salient one and we take a narrower look at intonation as perceptually relevant movement of voice pitch, regardless of its linguistic motivation.

Observing intonation is equally as important as observing the segmental properties of speech, since in everyday communication it carries valuable linguistic and non-linguistic information. In tone languages, intonation has a lexical function and is used to differentiate between word meanings. In many languages it serves a grammatical purpose and helps to distinguish whether an utterance is a question or a statement. It also helps to build the syntactic structure of spoken text. Intonation carries pragmatic information; speakers use it to express their moods, emotions and attitudes. It is used to deliberately

https://doi.org/10.14712/24646830.2017.31

accentuate certain words or syllables. It also has an indexical function; it gives us non-linguistic cues about the speaker, his/her age, gender or social status. (Volín & Bořil, 2013)

In analysing the intonation of an utterance, we usually look for pitch events – movements and patterns that are relevant for the perception of intonation. It is still hard to determine what is and what is not a pitch event and there are different approaches to this issue. A pitch event may be represented by a turning point (a peak, a valley or an inflection point) in the pitch contour, by a simple rising, falling or level pitch movement or by a compound pitch movement such as a rise-fall or fall-rise. We can either observe individual pitch events and patterns and their alignment in relation to the segmental layer, or we can observe overall characteristics of the intonational contour, such as pitch range or the rate at which pitch movements occur. While the former is suitable for examining the intonational features that have a linguistic function (such as lexical tones, pitch accents or boundary tones), the latter may be a source of more abstract linguistic or even non-linguistic information. For example, flat intonation with only slight changes may be a sign of a speaker's boredom or lack of interest, while intonation curves with dramatic rises and falls may indicate the speaker's excitement. In Czech, expressive utterances are often characterised with a wider intonation range and more dramatic pitch changes. Emotional utterances frequently use rising pitch movements, or upward transposition of the nuclear accent (or melodeme) (Palková, 1997).

2. F0 contour, stylization and modelling

Intonation melody is a psychoacoustic variable that cannot be measured instrumentally, but we can measure its acoustic correlate, the contour of voice fundamental frequency, or F0 contour for short. The raw F0 contour extracted from an utterance (e.g. using the pitch extraction algorithm in *Praat* software (Boersma & Weenink, 2016)) includes macrointonation (perceptually relevant pitch movements), but also microintonation (pitch fluctuations caused by segmental properties of the utterance). To filter out the perceptually irrelevant information, we use pitch contour *stylization*.

Stylization is a process that tries to simulate human perception of intonation by smoothing and simplifying the F0 contour so it can be interpreted as a sequence of basic elements such as pitch rises, falls or peaks and valleys, whose properties can be quantified by a set of parameters. The core principle of stylization is *perceptual equality*, i.e. the requirement that an utterance resynthesized with a stylized F0 contour should be perceptually indistinguishable from the same utterance resynthesized with the original contour (Hermes, 2006).

Many stylization algorithms have been proposed so far, ranging from simple, purely statistical ones to those that incorporate advanced perceptual and psychoacoustic phenomena. MoMel (Hirst & Espesser, 1993) is one of the former. It uses quadratic regression to find target points in the contour (in most cases, the points identify with local maxima, minima and inflection points) and then interpolates between them using quadratic splines. The Fujisaki model (Fujisaki & Ohno, 1997) presents a different approach in that it treats the F0 contour as a superposition of "phrase components" and "accent (or tone) components". The phrase component layer is modelled by low-pass filtering

weighted impulses ("phrase commands") and the accent components are modelled by filtering square pulses ("accent commands") with variable amplitude and duration. The *tonal perception model* (Mertens & d'Alessandro, 1995) assumes that we do not perceive intonation as a continuous contour, but rather as a sequence of syllabic tones. The algorithm detects syllabic nuclei, approximates their pitch with static (level) or dynamic (rise/fall) tones and linearly interpolates the contour outside the nuclei.

In this paper we adopt the tonal perception model for several reasons. Compared to the formerly mentioned stylizations, this one is not just based on statistical smoothing, rather it builds on experimental findings in speech pitch perception. It uses *glissando threshold* to determine whether a pitch change in a syllable nucleus is salient enough to be perceived as a dynamic tone and *differential glissando threshold* to determine whether the change of pitch movement rate is salient enough to be perceived as a compound tone. Another advantage is that the algorithm has been implemented in the intonation analysis software *Prosogram* (Mertens, 2004). It is freely available for download as a set of Praat scripts with a graphic interface.

3. Method and research question

The rate and intensity of pitch events may be essential in deciding whether we perceive speech as monotonous or variable. We suppose that speech with frequently occurring and prominent pitch changes will be generally perceived as expressive and variable. Speech with few melodic events that use a small pitch range will be regarded as rather monotonous. In theory, the simplest monotonous intonation is stylized by one slightly declining straight line. The more variable intonation is, the more pitch events appear in the contour and the more it deviates from a straight line. In effect, more turning points and line segments are needed for a reliable stylization.

Suppose an F0 contour that has been stylized as pitch points interpolated with linear segments. Hruška (2016) defines a *cumulative slope index* (in semitones per second) so that

$$CumSlope = \frac{1}{T_{tot}} \sum_{n=2}^{N} |f(n) - f(n-1)|$$

where T_{tot} is the total duration of voice activity (in seconds), N is the number of discrete points in the pitch contour and $f(n)$ is the frequency in semitones of the n-th point.

Frequency in semitones (ST) is

$$f_{ST} = \frac{12\ log(f_{Hz}/100)}{log(2)}$$

where f_{HZ} is frequency in hertz and 100 is a reference value[1] in hertz for 0 ST.

The CS index is merely the sum of absolute frequency differences between subsequent pitch points divided by the duration of the measured speech segment. Consequently, variable and expressive utterances should show higher CS index values than monotonous utterances.

[1] As the cumulative slope index reflects solely frequency differences, the reference value has no impact and any other value would lead to exactly equal results.

4. Experiments

To test how the proposed index captures/reflects the variability of F0 contours, we conducted two experiments. The data set for Experiment 1 (presented in Hruška (2016)) was taken from the Mini-Dialogue part of the Prague Phonetic Corpus. These scripted dialogues in Czech were performed by students of philological programmes at Charles University. For this study, we randomly chose 5 males and 15 females from the corpus. Each of 10 dialogue lines (see Table 1) was read and acted in neutral style by every speaker, 7 utterances were removed (faulty, omitted or inserted words by the speaker) leading to 193 utterances in total. These recordings were manually segmented, allowing for a precise measurement of duration of voice activity. The purpose of the experiment was to test whether the CS index would be stable across subjects and utterances in the case of similar reading style and to explore its variability.

Experiment 2 was designed to study the impact of neutral / flat / expressive style on the CS index and to test whether a fully automatic voice activity detection (VAD) algorithm could be used to substitute the manual segmentation process without significantly altering the results. Hence, two voice activity detection (VAD) methods were compared. First, fully automatic VAD algorithm was used (Sohn et al., 1999) and second, voice activity segments were marked manually. In both methods, for a pause to be labelled as a voiced-inactive segment, the minimum of 100ms was chosen. Text of three mini-dialogues from the Prague Phonetic Corpus were chosen (15 turns, see Table 2) but completely new recordings were taken with 3 males and 2 females ranging from 23 to 32 years of age (no speaker of Experiment 1 took part in Experiment 2). To test the capability of the CS index to distinguish among different dynamics of F0 contours, the speakers were instructed to perform the dialogues in three styles: a) neutral acting, b) monotonous (flat intonation), and c) with expressively dynamic intonation.

Both Experiment 1 and 2 were recorded in a sound-treated studio. Experiment 1 recordings have 32 kHz sample rate with 16bit PCM, Experiment 2 recordings are sampled with 48 kHz and 16bit PCM. The difference in sample rate should not have any impact on the following analyses as the only concern is about VAD and F0 extraction.

F0 contours were extracted and stylized in Prosogram v2.13 (Mertens, 2004) for Praat (Boersma & Weenink, 2016) with the following settings: *Task – Calculate intermediate files, Segmentation method – Automatic: acoustic syllables*. Then, stylized contours were processed using rPraat (Bořil & Skarnitzl, 2016). Frequency values were converted to semitones and the CS index was calculated for each utterance among all voice-active segments.

Table 1. Experiment 1, 10 utterances.

ID	Text
H1a_5	To bych se nebál.
H2d_1	Máte nějakou představu, jak to bude probíhat?
H3a_1	Až začnou o tom transportu, nastražíš uši.
H3b_5	Na internetu to není?
H3c_1	Trochu se bojím, že je tím rozčílíte.
H4b_1	Hodilo by se trochu víc informací.
H5c_1	Takže všechno to teď závisí na vás.
H5d_5	A ze vzduchu nic vidět není?
H6c_5	A právě s tím nikdo nepočítá.
H6d_1	Jen abyste měli dost peněz, až to přijde.

Table 2. Experiment 2, 15 utterances.

ID	Text
H1b_1	Začni opatrně. A o další spolupráci bych se nezmiňoval.
H1b_2	To by mě ani nenapadlo.
H1b_3	Řekneš jim, co si myslíš?
H1b_4	Nevím, snad nebudu muset.
H1b_5	Dobře, nezapomeň, hlavní heslo: opatrnost.
H1c_1	Nejlepší bude, když zatroubíte a počkáte, až vylezou.
H1c_2	Hmm, no a potom?
H1c_3	Řeknete jim, co si myslíte.
H1c_4	A vy na nás počkáte?
H1c_5	Jasně, ani se nehneme z místa.
H2a_1	Takže se posadíš a budeš se usmívat. Žádnou paniku.
H2a_2	Jasně. A co mám dělat, až přijde ten jejich šéf?
H2a_3	Zeptáš se, co bude dál.
H2a_4	To je všechno? Nemám mu říct, že už to víme?
H2a_5	Ne, nic nevíš. Ani slovo.

5. Results

Results of Experiment 1 are shown in Fig. 1. Fig. 1a shows the distribution of cumulative slope index for each dialogue turn across all speakers. Most values are spread in the interval from 5 to 15 ST/s with no sign of any anomalous (outlier) trend. A few turns seem to be slightly higher (H3a1, H3c_1, and H4b_1) with more than 75% of values larger than 10, which may be due to apparent pragmatic element in their meaning which might have led to the speakes's more dynamic acting. Fig. 1b presents the distribution of CS index of all dialogue turns grouped by speaker. Again, the variability disallows to observe any unique trend, although speaker F9 reaches consistently mildly higher values

and speaker F12 seems to remain more often in lower values. These results are not surprising as the conditions of the experiment naturally lead to a neutral acting style which can explain the similar behaviour of intonation dynamics of utterances.

Experiment 2 is presented in Fig. 2 and 3. Figure 2 compares automatic and manual VAD and although manual VAD is certainly more precise (and significantly more time-consuming), the results are very similar. As expected, expressive style leads to higher values of CS index (see Fig. 2a) and monotonous (flat) style produces generally lower values. A detailed look at the dialogue turns (Fig. 2b) shows the same overall trend, although due to evident variability across dialogue turns one cannot determine a threshold that would reliably separate the three styles.

A closer look at individual speakers with manual VAD (see Fig. 3) implies that some speakers performed the three styles of recordings very diversely. On the other hand, female F2 shows very similar range of values for both the neutral and the expressive style. After listening to a few of her utterances, the reason is apparent. To achieve the expressive style, she used other means than intonation such as variation of whisper and shout, vocal timbre or dynamic rhythm changes. As the CS index is meant to measure variability of the F0 contour, this result matches the expectations.

We used R (R Core Team, 2017) and lme4 (Bates, Maechler & Bolker, 2015) to perform a linear mixed effects analysis of the relationship between CS index and style. As fixed effects, we entered style and VAD method into the model (the interaction term was tested using likelihood ratio test but was insignificant with $p = 0.580 >= 0.05$). As random effects, we had intercepts for subjects and utterances, as well as by-subject and by-utterance random slopes for the effect of both style and VAD method. P-values were obtained by likelihood ratio tests of the full model with the effect in question against the model without the effect in question.

Style affected CS index ($p = 0.008 < 0.05$), lowering it by about 5.9 ST/s ± 1.2 standard errors (flat style) and rising it by about 4.7 ST/s ± 1.6 standard errors (expressive style). VAD method also affected CS index ($p = 0.012 < 0.05$), manual VAD rised it by about 1.4 ± 0.5 standard errors in comparison to the automatic VAD. Although statistically significant, this value is low in absolute terms in comparison to the effect of style, and thus the practical impact of the VAD method to the overall use of CS index value to distinguish among styles is not notable.

Detailed look at model coefficients showed that the manual VAD method impact is very consistent across utterances (min. 1.16, max. 1.68 ST/s increase) and also across subjects (min. 0.95, max. 1.72 ST/s). Also, the effect of styles is very consistent across utterances. Although there are differences of the effect of styles across subjects (in three male subjects the differences among styles are more distinct than in two females), the sample size is too small to make any generalization in this sense. Even though each subject had different range of style dynamics, the direction of CS index difference among styles was consistent across all five subjects.

Figure 1. Experiment 1, cumulative slope index of all utterances (a) by utterance, (b) by speaker.

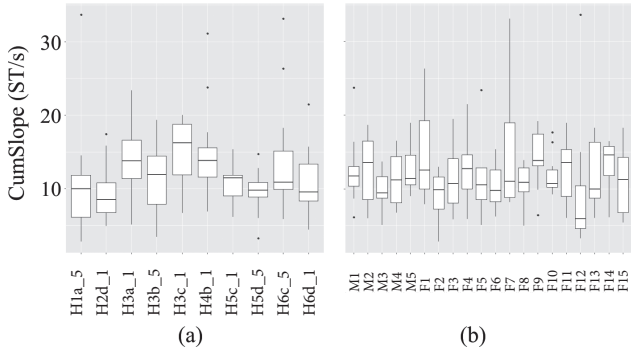


Figure 2. Experiment 2, cumulative slope index for the three intonation styles: neutral, flat, and expressive. Automatic and manual methods of voice activity detection are compared. (a) depicts summary plot of all utterances by all speakers, (b) displays identical data sorted by utterances, i.e. each colour group consists of 15 boxes, each corresponding to one utterance (see Table 2) performed by all speakers.

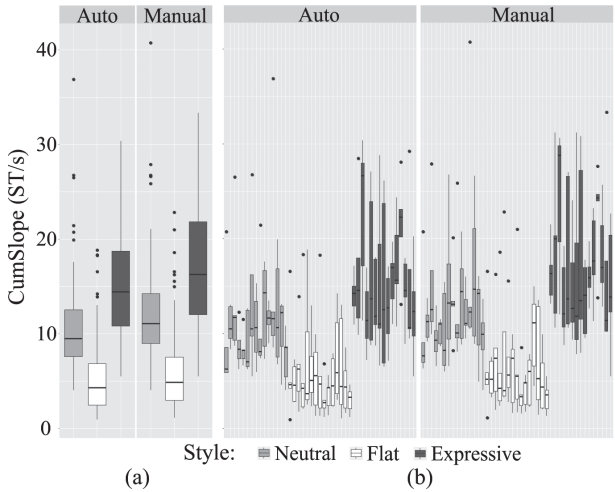
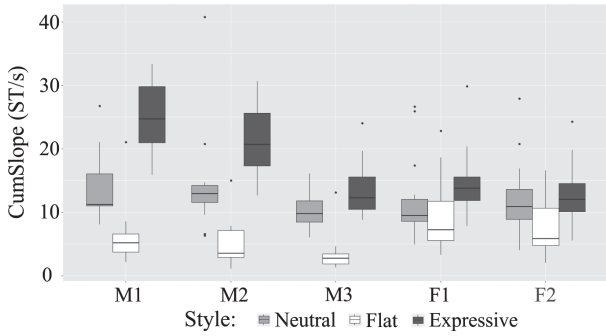


Figure 3. Experiment 2, cumulative slope index of all utterances by 3 male (M1, M2, M3) and 2 female (F1, F2) speakers in three style scenarios, manual voice activity detection only.



6. Discussion and conclusions

In Experiment 1, the speakers did not focus specifically on their acting style. Moreover, they were students of similar age, education and social circumstances. Thus, the cumulative slope values are obviously variable, but no trend can be observed. In Experiment 2, speakers were instructed to stylize their speech with expressive, neutral and flat intonation and the values differ between the individual styles quite consistently across the speakers except for F3, who did not use intonation to achieve the expressive style, but rather used other means, such as variation of articulation rate or vocal timbre. In these cases, the CS index values approach those of neutral style utterances, which is an expected result. However, similarly to Laan (1996), who compared the prosodic properties of read and spontaneous speech, we found that F0 variation alone could not reliably discriminate between the three speech styles.

When we compare the results obtained through automatic voice detection to those obtained through manual voice detection we see the values are fairly similar. Even though the automatic VAD is not as precise, it is consistent in the errors it produces and significantly less time-consuming since it saves manual work. The errors mostly occurred at voiceless plosives or when there were audible breaths or mouth clicks in the recording. They led to longer voice activity having been detected and thus slightly lower CS index values. Nevertheless, this behaviour is consistent across all the styles and speakers and it does not skew the relative differences between them. Thus, CS index can be measured and computed in a fully automatic manner with no human interference required.

The CS index is a simple sum of absolute values of pitch changes divided by the duration of the measured segment. It does not provide information about the actual slope or steepness of the F0 contour, rather it shows how much the contour varies in a given speech segment (in our case a dialogue turn). There is no absolute scale that could be used as a reference for the measured values, since the index seems to depend on the speaker and on the utterance. As it is defined now, it can only be used to compare speech segments with/relative to each other, e.g. two utterances by the same speaker. Furthermore, the CS index does not reflect the shape of the measured contour. A consistently rising intonation may produce the same value as a consistently falling one and no difference is made between a pitch movement at the beginning and at the end of an utterance. However, the CS index was designed to be simple and suitable for fully automatic processing of large quantities of data.

In future research, it would be useful to test a few variants of the CS index, such as instead of dividing the sum of pitch differences by the total duration, dividing it by the number of syllables in the measured segment, obtaining the index value in semitones per syllable. This method would still allow for fully automatic implementation since syllabic nucleus detection is already involved in the Prosogram stylization we use. Another experiment could measure the CS index using a sliding window to see how it changes over time and to test whether it can be used in real-time signal processing, e.g. for monitoring actual mood of a client during a phone call with an operator.

ACKNOWLEDGEMENTS

This research was supported by the Czech Science Foundation project No. GA16-19975S "Age-related changes in acoustic characteristics of adult speech".

REFERENCES

Bates, D., Maechler, M., Bolker, B. & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1–48.

Boersma, P. & Weenink, D. (2016). *Praat: doing phonetics by computer* [Computer program]. Version 6.0.14, retrieved 11 February 2016 from http://www.praat.org/.

Bořil, T. & Skarnitzl, R. (2016). Tools rPraat and mPraat. In: P. Sojka, A. Horák, I. Kopeček & K. Pala (Eds.), *Text, Speech, and Dialogue*, 367–374. Berlin: Springer International Publishing.

Fujisaki, H. & Ohno, S. (1997). Comparison and assessment of models in the study of fundamental frequency contours of speech. In: *INT – 1997*, 131–134.

Hermes, D. J. (2006). Stylization of pitch contours. In: S. Sudhoff et al. (Eds.), *Methods in empirical prosody research*, 29–61. Berlin: Walter de Gruyter.

Hirst, D. & Espesser, R. (1993). Automatic modelling of fundamental frequency using a quadratic spline function. In: *Travaux de l'Institut de Phonétique d'Aix*, Vol. 15, 75–85.

Hruška, R. (2016). *Properties of fundamental frequency contours in segmental contexts*. Unpublished bachelor thesis. Prague: Institute of Phonetics, Faculty of Arts, Charles University.

Laan, G. P. M. (1996). *The Contribution of Intonation, Segmental Durations, and Spectral Features to the Perception of a Spontaneous and a Read Speaking Style*. Amsterdam: Institute of Phonetic Sciences, University of Amsterdam.

Mertens, P. & d'Alessandro, C. (1995). Pitch contour stylization using a tonal perception model. In: *Proceedings of 13th International Congress of Phonetic Sciences*, Vol. 4, 228–231.

Mertens, P. (2004). The Prosogram: Semi-Automatic Transcription of Prosody based on a Tonal Perception Model. In: B. Bel & I. Marlien (Eds.), *Proceedings of Speech Prosody 2004*, Nara (Japan): ISCA.

Palková, Z. (1997). Modelling intonation in Czech: Neutral vs. marked TTS F0-patterns. In: *Intonation: Theory, Models, and Applications*. Athens, Greece.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Sohn, J., Kim, N. S. & Sung, W. (1999). A statistical model-based voice activity detection. *IEEE Signal Processing Lett.*, 6 (1), 1–3.

Volín, J. & Bořil, T. (2013). General and speaker-specific properties of F0 contours in short utterances. *AUC Philologia 1/2014, Phonetica Pragensia XIII*, 9–20.

RESUMÉ

Intonace zásadně ovlivňuje vyznění i celkový význam řeči a při různých mluvních stylech může mít rozličné průběhy. Z toho důvodu je zajímavé sledovat variabilitu časového vývoje intonační výšky jednotlivých promluv. Navrhli jsme ukazatel kumulativní strmosti, tzv. cumulative slope (CS) index, který jedním souhrnným číslem vyhodnocuje míru proměnlivosti kontury základní frekvence řečového signálu (F0) vzhledem k celkovému trvání analyzovaného úseku.

Aby tento ukazatel skutečně pracoval s průběhem vnímané výšky intonace, rozhodli jsme se jej použít na tzv. stylizované kontury F0, které byly vypočteny algoritmem modelujícím percepční dopad vývoje základní frekvence řeči. Předpokládáme, že intonačně dynamičtější promluvy obdrží vyšší hodnotu CS indexu, zatímco monotónnější řeč bude ohodnocena číslem nižším.

V rámci dvou provedených experimentů jsme zkoumali chování hodnot CS indexu. Nejdříve jsme analyzovali nahrávky předem připravených krátkých dialogů, u kterých jsme očekávali podobnou míru dynamičnosti intonace, jelikož se mluvčí adaptovali na obdobný mluvní styl, příhodný pro tento typ úlohy. Obdržené výsledky CS indexu skutečně odhalily, že vzhledem k přirozené variabilitě hodnot ukazatele nebyly výrazné rozdíly mezi jednotlivými mluvčími. Ve druhém experimentu byli mluvčí explicitně instruováni k tomu, aby dialogy namlouvali postupně několika styly, a sice neutrální intonací, expresivní intonací a monotónní, resp. plochou intonací. V této úloze dosahoval CS index v jednotlivých stylech skutečně výrazně rozdílných hodnot. Pokud však byla expresivnost promluvy dosažena jinou cestou než výraznými změnami výšky intonace, tedy například dramatickým střídáním mluvního tempa a přechodem mezi šepotem a hlasitou řečí, CS index nabýval hodnot v rozsahu typickém pro neutrální styl. Takové chování považujeme za správné, protože tento ukazatel byl navržen pro vyhodnocování jedné konkrétní složky intonace a pro měření tempa řeči nebo stylu fonace mohou být použity jiné ukazatele.

Výstupy obou experimentů napovídají, že CS index je do jisté míry nezávislý na mluvčím a odráží míru dynamičnosti průběhu výšky v rámci intonace, která souvisí i s mluvním stylem. Vzhledem ke srovnatelným výsledkům v případě použití automatické detekce řečové aktivity s detekcí manuální je možné proces výpočtu CS indexu spolehlivě plně automatizovat.

Robin Hruška, Tomáš Bořil
Institute of Phonetics
Faculty of Arts, Charles University
robinhruska@email.cz