AKUSTICKÉ LISTY

České akustické společnosti
www.czakustika.cz

ročník 25, číslo 3–4 prosinec 2019

Obsah

ČESKÁ AKUSTICKÁ SPOLEČNOST

Fundamental Frequency Tracks of Question-word Questions in Natural and Synthetic Speech

Kontury základní hlasivkové frekvence v doplňovacích otázkách v přirozené a syntetické řeči

Jan Volín a Pavel Šturm

Univerzita Karlova, Filozofická fakulta, Fonetický ústav – nám. Jana Palacha 2, 116 38 Praha 1

The relationship between the prosodic feature of speech melody (intonation) and fundamental frequency (F0) of voice is briefly introduced together with an overview of the communicative functions of the phenomenon. The core of the study builts upon a sample of 448 sentences spoken by 28 Czech speakers and the same sentences produced by the text-to-speech synthesis system of ARTIC, which is based on concatenation of variable-size units without a special prosodic module. The differences in global statistic descriptors between human and synthetic production of fundamental frequency tracks were sought together with information about cummulative slope index (CSI) and the pattern provided by k-means cluster analysis of the sample. Relatively clear differences between statements and question-word questions emerged.

1. Fundamental Frequency in Speech

1.1. General exploitation of F0 in speech

The speech signal can be briefly characterized as a relatively quickly changing complex sound. As such it generally serves communicative purposes – various configurations of the rapid change of spectral properties can be perceived as short units (known as words or morphemes) and these units are assigned a content by convention: the users of a given language agree with the referential meanings ascribed to the individual configurations (i.e., words, morphemes).

An important part of the coding system is the presence or absence of the fundamental frequency (F0) in the complex. Thus, a broad band of aperiodic higher frequencies without F0 is recognized as [s], while with the F0 present it will be [z]. Similarly, aperiodic noise with less power and a lower centroid without F0 will function as [f], while with F0 as [v].

However, F0 is used for other communicative purposes as well. This is allowed by its prevalence in the speech signal: all the vowels (like [e], [a] or [o]) and sonorant consonants (like [m], [j] or [l]) occur with F0 in modal speech. Given that languages of the world are always spoken in syllables and a syllable must have a vowel or a sonorant consonant in its nucleus, F0 is present in spoken utterances most of the time. It can, therefore, serve as a carrier of prosodic information, especially of speech melodies. Those can signal differences between questions and statements or between finished and unfinished utterances. Apart from these obvious functions, speech melodies contribute substantially to the creation of word groupings that are essential for effective decoding of the utterance meanings. A chain of words has to be grouped into the so-called phrases, without which the decoding process becomes demanding, tiresome and sometimes even erroneous.

F0 in speech is also the carrier of very important affective meanings. On the level of emotions, it may signal, e.g., anger, boredom, happiness or surprise. On the level of immediate stances, which are more common in everyday situations, it allows for signaling politeness, involvement, reservation, willingness to co-operate, irritation, doubts, decisiveness, sarcasm and many other states that are often more important than the actual words of an utterance (see, e.g., [1]).

Last but not least, the melodic patterns resulting from F0 courses can reveal important sociophonetic or idiosyncratic facts. This means that the geographic origin or the socio-economic status of the speaker can be estimated together with the information concerning the age, health condition, tiredness or intoxication. All of these, in combination with individual unique features, can be used for forensic purposes [2], [3].

1.2. Correlation of F0 and speech melody

In many research and technological application tasks, it is important to remember that F0 values extracted from speech cannot be simply equalled with pitch, and, more importantly, F0 tracks do not reflect the speech melody directly. Human perceptual mechanisms are unable to 'read' the F0 values objectively in hertz (Hz) in real time. Rather, the melodic information seems to be retrieved only in middle parts of the syllabic nuclei and the individual values are then interpolated [4], [5]. Moreover, the melody as such is perceived relatively or relationally, so even if the mel scale is an established psychoacoustic instrument, the speech melodies alone seem to be better reflected by semi-

Přijato 24. listopadu 2018, akceptováno 25. června 2019.

tone measures [6]. (Also, *cf.* alternative ERB measures in [7].)

The account given here is only sketchy since a precise model of speech melody perception has not been built yet. For instance, speakers somehow normalize values depending on the vowel quality and vowel [i] requires higher objective F0 to be perceived with the same pitch as the vowel [a] (see, e.g., [8]). Also, longer vowels allow for perception of pitch movements within the syllable nucleus (like rises and falls in phrase-final positions), but how long exactly a nucleus must be to allow for this and whether the effect is influenced by global articulation rate still remains to be investigated.

To summarize, currently accepted practice in estimating phrasal speech melody from F0 tracks rests in extracting the mean values from the second third of every syllabic nucleus and conversion of the values into semitones (ST). Our methodology in the present study reflects that.

1.3. QW-questions in Czech

A well-known long-term tradition of dichotomic categorization of sentences into statements and questions is undisputed, and it is especially important in languages where the only differentiating factor is the melody (intonation) in speech and a question mark in writing. However, it should be remembered that languages often possess more than just one type of questions. Two major types are yes-no questions (e.g., *Are you hungry?* or *Do you know him?*) and wh-questions (e.g., *What time is it?* or *Where is your car parked?*). More generally, the wh-questions should be called *question-word questions* (QW-questions) since spelling in various languages does not indicate question words with letters 'wh' (in Czech it is usually letter 'k': *kde* (*where*), *kdy* (*when*), *kam* (*where to*), *kudy* (*which way*), *kolik* (*how much*), etc.), but the English terminology is quite widely used across various language descriptions. The yes-no questions are also called *polar questions*.

The typical melody used for polar questions in the Czech language differs from that of QW-question. While polar questions are signalled by rising intonation, QW-questions display the opposite. Their typical intonation is falling. However, if a typical statement melody is falling, is there any difference between the contours on QW questions and on statements? The accounts in the Czech most quoted sources suggest that there is no difference ([9], [10], [11]), although [11] admits variants. Generally, melodies in standard statements and QW-questions are supposed to be the same.

Our observations of current intonation, however, suggest otherwise, and no newer accounts of Czech questioning melodies exist to the best of our knowledge. Therefore, an experiment was prepared to test the hypothesis about the uniformity of Czech statements and QW-questions (the null hypothesis) or their differentiation in speech (the alternative hypothesis in this study). Moreover, naturally spoken utterances will be compared with synthetic sentences (see below, Section 2.3) produced by a system based on the pre-suppositions of [9], [10], [11], i.e., on the idea that melodies of statements and QW-questions can be constructed by the same procedure.

2. Method

2.1. Speech material design

A set of 8 sentence pairs was created where one member of a pair was a QW-question, while the other member was a statement differing in just the initial phone (*kam* × *tam*), syllable (*jak* × *a*) or, in just one pair, the initial two syllables (*odkud* × *tak už*). An example of one of the pairs follows:

QW-question: Kam pojedete v poledne? (*Where will you go at noon?*)	1a
Statement: Tam pojedete v poledne. (*There you will go at noon.*)	1b

The 8 target pairs together with several distractors (other sentences and polar questions) were jumbled so that the contrast between the statements and the questions was not obvious. We did not want the respondents to adjust their natural speech habits to an explicit research hypothesis. Moreover, the individual targets were supplemented by a lead-in phrase. For instance, the example 1a above was preceded by a sentence *Někdy bych jel s váma* (I'd like to go with you some time), while the example 1b was preceded by *To teď nechte být* (Leave this alone now). The individual members of a pair were kept separated from each other by a large number of other sentences. All these measures were taken because we did not want the respondents to consciously portray the possible intonational contrast. It is highly recommended not to reveal the tested hypothesis to the respondents.

2.2. Recording procedure

The list of jumbled items was given to 28 respondents (14 men and 14 women), who were native speakers of Czech without any speech, hearing or sight impairment. The speakers were individually asked to read out the sentences in a most natural manner. They were advised to imagine a friend or family member of theirs and utter each sentence as if they were saying it to them, not reading it from a sheet of paper. Also, they were invited to self-correct themselves if they thought that the rendering of a sentence was not 'good', i.e., not sounding natural.

The sound-treated recording studio of the Institute of Phonetics in Prague was equipped with a condenser microphone AKG C4500 B-BC, which was plugged directly into an external soundcard SB Audigy 2 ZS. The sentences were recorded with 32-kHz sampling rate and 16-bit resolution and saved in an uncompressed format as WAV files.

2.3. Synthetic speech material

In order to synthesize the target speech material we used the ARTIC synthesis system (*Artificial Talker in Czech*, see [12], [13]). Three features of the system should be mentioned. First, it is a text-to-speech (TTS) application, converting input text to output speech. Second, it is based on the principles of concatenation, where acoustic units are selected from an extensive corpus of pre-recorded natural speech, and joined together in a linear sequence. Importantly, the QW-questions are not treated as a special category due to suggestions in [9], [10], [11], which we challenge in the current study.

The corpus in ARTIC was built on more than 10 000 sentences, which were annotated orthographically and phonetically and described in terms of various acoustic parameters. The size of the concatenated units varies depending on how much of the desired input text corresponds to one of the database sentences (e.g., half of a sentence can be used, one word, or just a diphone). Lastly, the system is based on unit selection, where speech units are selected from several alternatives to meet some specific criteria, which can be local (e.g., spectral shape) or global (e.g., prosodic position). Consequently, each synthesized sentence has a certain overall 'cost' related to how well it is assembled from the individual units.

Using the online ARTIC interface, we synthesized the target sentences in a male voice (version 'artic_images/spkr_AJ.rev698.img'). The first offered alternative was always selected in order to provide consistency and automaticity without human interference into the selection process. Presumably, the first alternative should be the best output of the system, but in some cases it is normally worth selecting a lower-rated alternative if the first contains any artefacts. The synthesized sentences were saved as WAV files onto the computer and further processed in the same fashion as the human-spoken sentences.

2.4. F0 tracks extraction and processing

Autocorrelation method built in the software analysis package Praat [14] was used to extract the F0 tracks of the target sentences. The individual contours were built from values extracted every 10 milliseconds. The F0 tracks were inspected and manually corrected where necessary for octave jumps (period doubling and period halving) and non-modal phonation phenomena like creaky phonation.

The individual tracks were interpolated through the voiceless regions (see above Section 1.2.) and several metrics of central tendency and data dispersion were extracted. In the presentation of the results in Section 3 below, the measures will be abbreviated as:

MN – arithmetic mean
MD – median
vSD – standard deviation
VAR – variation range (from maximum to minimum)
PER – 80% percentile range (from 10^{th} to 90^{th} percentile)
IQR – interquartile range (from 25^{th} to 75^{th} percentile)
CSI – cumulative slope index
GRD – gradient of a regression line

Most of the metrics are basic concepts of descriptive statistics, but CSI and GRD perhaps deserve a comment. Cumulative slope index summarizes differences in values between all discrete points of measurements but normalizes the outcome with the duration of sentences in seconds (or in number of syllables as in [15]). The gradient of a regression line is calculated with the least-sum-of-squares method through the F0 contour. It is known that the gradients of phrases and sentences are prevalently negative, even if there is a final rise in F0 [16]. All the differences in the above-listed parameters are tested at the level of $\alpha = 0.05$.

K-means clustering of individual trajectories as one of the exploratory methods was opted for because it has a transparent capacity to group similar data points (trajectory shapes) together and provide "centroid shapes" with unproblematic interpretation. All the F0 values were normalized by the speaker's mean before clustering, i.e., the mean was set to 0 ST. The deviation as such (as a correlate of pitch range) was not normalized since the semitone scale safeguards comparability of male and female speakers [6].

3. Results

3.1. Descriptors of the F0 Tracks

Generally, arithmetic mean and median are measures of central tendencies, but in terms of F0 they describe the average level. Table 1 captures the trends in our data: for both men and women the level for questions is higher than for statements. In synthetic speech, however, this trend is not present. The significance of the differences was confirmed only for the mean but not median by ANOVA for paired measures (Q vs. S of the same speaker): $F(2, 214) = 647.6; p < 0.001$ and Tukey HSD post-hoc test revealed that the significance was achieved through the male and female speakers, not the synthetic sentences.

	Women		Men		Synthesis	
	Q	S	Q	S	Q	S
Mean	233.4	214.2	127.9	119.6	121.8	122.3
Median	222.9	216.5	123.8	122.1	119.1	121.1

Table 1: Mean measures of F0 central tendencies in Hertz for questions (Q) and statements (S) produced by female, male and synthetic speakers

Table 2 is organized analogically, but it is focused on dispersion or, rather, variation within the F0 contours. Clearly, all metrics are higher for questions than for statements when men or women talk, but slightly opposite (insignificantly) in synthetic speech. The signifi-

cance of the differences was ascertained with ANOVA for paired measures (Q vs. S of the same speaker) as follows: $F(2,214) = 14.44$; $p < 0.001$ for SD, $F(2,214) = 5.37$; $p < 0.05$ for VAR, $F(2,214) = 12.54$; $p < 0.001$ for PER, $F(2,214) = 12.60$; $p < 0.001$ for IQR, and $F(2,214) = 7.41$; $p < 0.01$ for CSI. Tukey HSD post-hoc test revealed that the significance was achieved due to the male and female speakers, but not the synthetic sentences.

	Women		Men		Synthesis	
	Q	S	Q	S	Q	S
SD	44.6	27.3	23.2	15.3	21.4	22.0
VAR	10.5	8.5	10.4	8.3	11.4	21.1
PER	8.1	5.9	8.2	6.0	7.9	8.2
IQR	4.6	3.4	4.6	3.0	4.8	4.6
CSI	18.4	15.8	18.5	15.1	19.1	19.8

Table 2: Mean measures of F0 variation for questions (Q) and statements (S) produced by female, male and synthetic speakers. For abbreviations see above Sect.2.4. SD is in Hertz, VAR, PER and IQR in semitones, CSI is ST/sec

Finally, the differences between questions and statements in the gradient of linear regression lines were tested. Table 3 shows that the slopes are steeper for questions than statements in both male and female speakers, but not in synthesized sentences. The ANOVA test did not find the differences significant ($F(2,214) = 1.71$; $p = 0.193$), but when the synthetic sentences were discarded, the effect became highly significant: $F(1,208) = 4.78$; $p < 0.001$.

	Women		Men		Synthesis	
	Q	S	Q	S	Q	S
GRD	−6.13	−4.90	−5.91	−4.37	−5.11	−5.45

Table 3: Mean gradient in ST/sec. for questions (Q) and statements (S) produced by female, male and synthetic speakers

Obviously, the descriptors of F0 tracks differentiate between Czech QW-questions and statements despite the claims in some sources that the melodies are the same. The difference was, however, found only for human speakers. The synthesis does not display it for a simple reason: its design was informed exactly by those sources that denied the difference.

3.2. K-Means clustering

Given the magnitude of the sample, k-means clustering was limited to 2-cluster, 3-cluster and 4-cluster solutions only. Clustering to five or more groups might already lead to artefacts. Moreover, our major concern was testing the hypothesis of the identical melody for both the statements and QW-questions (see above) and **not** to establish a number of potential contour types used in current Czech language.

The null hypothesis would then predict equal number of questions (Q) and statements (S) in each cluster delineated by the clustering procedure. Conversely, if the Q to S ratio in the established clusters differs from 1, then the null hypothesis is not supported, and the alternative hypothesis becomes more probable. Specifically in our case, this would mean different behaviour of F0 in QW-questions and statements.

The results of the cluster analysis will be presented in three sections according to the assigned number of clusters.

3.3. Two-set clustering

Figure 1 displays the average outcome of clustering into two sets. Each pair was clustered separately and the resulting set with more actual questions in it was labelled Q-type, while the set with more statements in it was labelled S-type. The null hypothesis would predict equal distribution of actual questions and statements in Q-type and S-type clusters. This apparently is not the case.

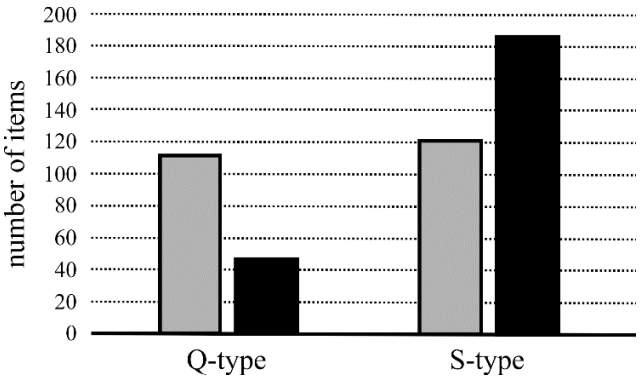


Figure 1: Distributions of questions (grey columns) and statements (black columns) into two types of clusters (see text)

The statistical significance of the difference between balanced distribution and the observed outcome was calculated with a chi-square test. The result returned very high significance: $\chi^2(1) = 39.43$; $p < 0.001$.

In terms of the individual pairs, the most clear-cut result occurred for the pair *Jak brzo viděl výsledek?* vs. *A brzo viděl výsledek* (*How soon did he see the result* vs. *And soon he saw the result*). The Q-type cluster contained 22 questions and 1 statement, while the S-type cluster contained 27 statements and 7 questions. Both synthetic sentences ended up in the S-type cluster, although the synthetic question had a greater distance from the cluster centroid. Figure 2 shows the centroid trajectories. It is quite evident that the Q-type has a prominent peak on the word *brzo* (s2 and s3), which rises more than 4 ST above the mean (i.e., 0 ST) and from the third syllable (s3) falls steadily until the end. The S-type starts 2 ST below the mean, rises moderately, stays around the mean for s3, s4 and s5, and after that falls in similar fashion as the Q-type.

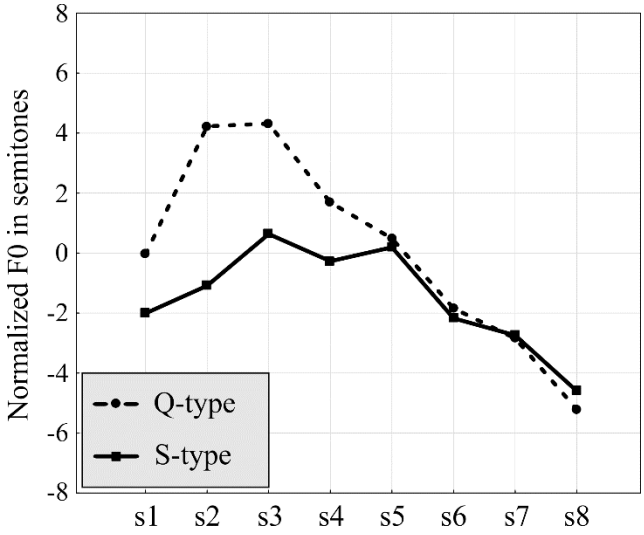


Figure 2: Centroid contours returned by k-means clustering method for one of the sentence pairs. The symbols s1, s2, ..., s8 on the axis x stand for individual syllables of the sentences. (For Q-type and S-type see text.)

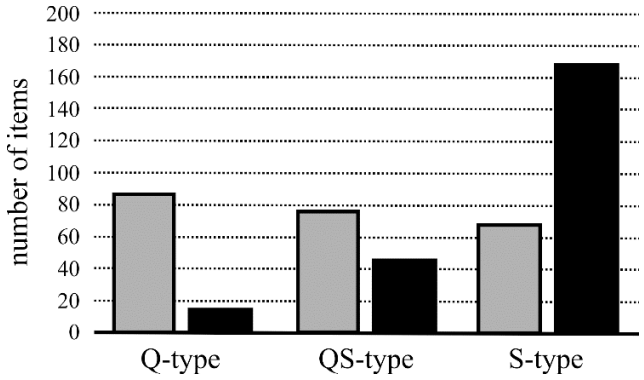


Figure 3: Distributions of questions (grey columns) and statements (black columns) into three types of clusters (see text)

3.4. Three-set clustering

Figure 3 portrays the mean outcome of clustering into three sets. Similarly to the previous, the clusters with prevalence of questions in them were labelled Q-type, whereas the clusters with prevalence of statements were labelled S-type. The remaining clusters were marked as QS-type.

Contrary to the null hypothesis the ratio of questions and statements in individual clusters varies considerably. In the Q-type clusters the number of questions is more than 5 times higher than the number of statements. The S-type clusters contained more than twice as many statements as questions. The QS-type is obviously more balanced. The chi-square test confirmed the difference from equalized distribution as highly significant: $\chi^2(1) = 99.20$; $p < 0.001$.

In terms of the individual pairs, the most straightforward result occurred for the pair *Kdy nemáme žádnou záruku?* vs. *My nemáme žádnou záruku* (*When do we have no guarantee?* vs. *We have no guarantee*).

The Q-type cluster contained 17 questions and 1 statement, while the S-type cluster contained 26 statements and 8 questions. The QS-type comprised 4 questions and 2 statements. Figure 4 shows the centroid trajectories. It is quite evident that the Q-type and S-type are very similar to the types in Fig. 2: questions start with a high peak on the second and third syllables (s2, s3) with a steady fall afterwards, while statements have much flatter contour mostly around and below the speakers' mean pitch (0 ST). The QS-type represents 6 items (i.e., questions) with a high end, which do not sound particularly typical.

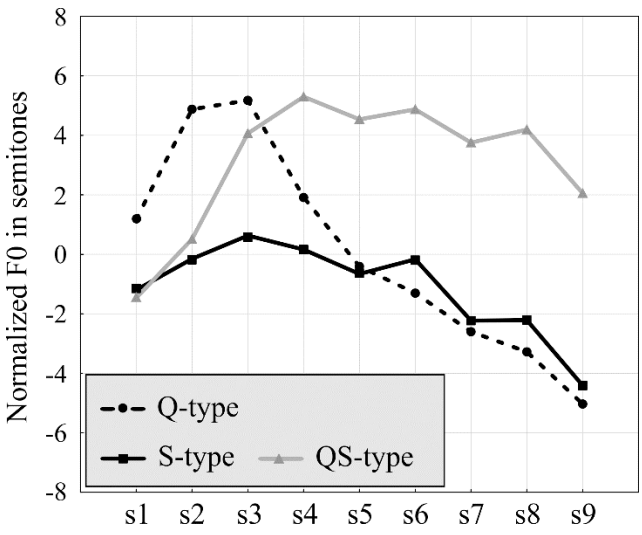


Figure 4: Centroid contours returned by k-means clustering method for one of the sentence pairs. The symbols s1, s2, ..., s9 on the axis x stand for individual syllables of the sentences. (For Q-type, QS-type and S-type see text.)

3.5. Four-set clustering

Figure 5 shows the mean outcome of clustering into four sets. Analogically to the two-set and three-set clusterings, the clusters with prevalence of questions in them were labelled Q-type, the clusters with prevalence of statements were labelled S-type, and the remaining two clusters were marked as QS-type and SQ-type depending on the ratio of questions and statements in them.

It can be observed that the ratio of questions to statement in individual types differs from each other. The Q-type contains more than 6 times as many questions as statements, and the number of statements in the S-type is more than 5.63 times higher than the number of questions. The chi-square test confirmed high significance of the observed differences: $\chi^2(1) = 251.80$; $p < 0.001$.

Results for individual pairs produced the clearest picture for the pair *Proč vám to nevyšlo?* vs. *Moc vám to*

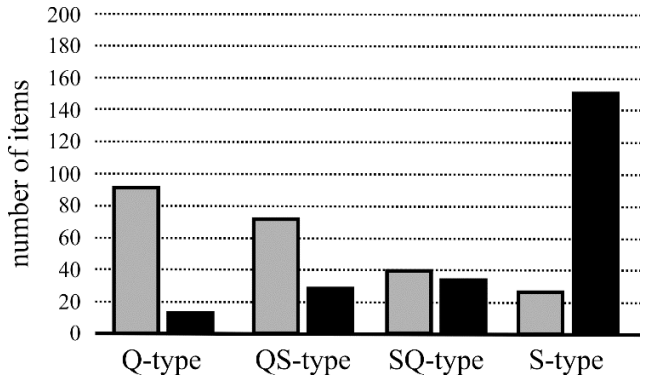


Figure 5: Distributions of questions (grey columns) and statements (black columns) in four types of clusters (see text)

nevyšlo (*Why didn't it work?* vs. *It didn't work much.*). The centroid trajectories are presented in Figure 6. Again, the Q-type and S-type are very similar to the types in Fig. 2 and Fig. 4.

This is encouraging, since although we observe different questions of different contexts and different lengths, the pattern is consistent: the questions have an initial high peak triggered by the question word, while the statements (S-types) are relatively flat with the final syllable in about the same position as that of the questions, that is about 4 to 5 ST below the mean value. These outcomes are pertinent to the dilemma mentioned above in Section 1.3.

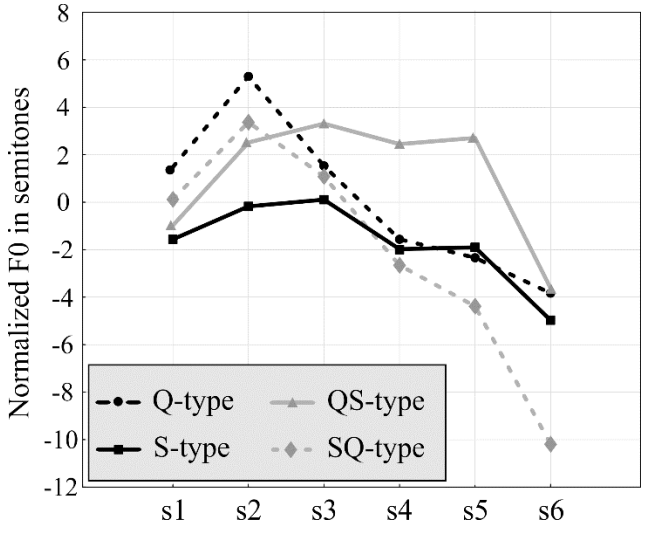


Figure 6: Centroid contours returned by k-means clustering method for one of the sentence pairs. The symbols s1, s2, ..., s6 on the axis x stand for individual syllables of the sentences. (For Q-type, QS-type, SQ-type and S-type see text.)

4. Conclusion

The descriptors in Section 3.1, which are global in that they represent each sentence by one value only, showed that both men and women produce the difference between QW-questions and statements in a similar fashion. The statistical significance of all the differences would be actually even higher than the one that is reported if the synthetic sentences were excluded. (In this study we only checked that for the variable GRD because the other significances were high anyway.)

The clustering approach indicated that the original claim of equal melodies for both statements and question-word questions in Czech is unfounded. Moreover, the increasing χ^2 criterion in three- and four-set clustering suggests that the language offers more than just two types of melodies, even if some seem to be more prototypical (or common).

We should also remember that k-means clustering relies on arithmetic measures of central tendencies, which may obscure some uncommon, yet linguistically valid cases. To ascertain those would be the task for the future analysis of individual items in the follow-up research.

Furthermore, it should be kept in mind that we only examined speech production at this point. That is just one side of the communicative process. Perceptual testing will have to follow to investigate how adequate the individual contours sound to listeners.

As to synthetic items, their questions and statements occurred in the same cluster in 22 out of 24 clustering steps, even if with different distances from the centroid. (This means that they did not have exactly identical melodies, but they did not differ substantially enough to be classified separately.) They fell into different clusters in two steps only, but in these two steps, the result was contrary to the desired outcome: the synthetic statement fell in the Q-type cluster and the synthetic question fell into the S-type cluster. This, of course, is no surprise, since the design of the examined version of ARTIC system relied on older descriptions of Czech intonation and did not incorporate any difference between statements and question-word questions.

Acknowledgment

This research was supported by the Czech Science Foundation project No. 16-04420S "*Kombinované využití fonetických a korpusově založených postupů při odstraňování rušivých jevů v řečové syntéze*".

References

[1] Scherer, K.: Vocal communication of emotion: A review of research paradigms, *Speech Communication* **40**, p. 227–256, 2003.

[2] Skarnitzl, R., Hývlová, D.: Statistický popis hodnot základní frekvence. In: R. Skarnitzl (ed.) *Fonetická identifikace mluvčího*, Praha: FF UK, p. 49–64, 2014.

[3] Volín, J., Bořil, T.: Základní frekvence v konturách a průbězích. In: R. Skarnitzl (ed.) *Fonetická identifikace mluvčího*, Praha: FF UK, p. 65–76, 2014.

[4] House, D.: *Tonal perception in speech*, Lund University Press, Lund, 1990.

[5] Mertens, P.: The Prosogram: semi-automatic transcription of prosody based on a tonal perception model, *Proc. of Speech Prosody 2004*, Japan, Nara, 2004.

[6] Nolan, F.: Intonational equivalence: an experimental evaluation of pitch scales, *Proc. of 15th ICPhS 2003*, p. 771–774, Barcelona, 2003.

[7] Hermes, D., van Gestel, J.: The frequency scale of speech intonation. *Journal of the Acoustical Society of America* **90**, p. 97–102, 1991.

[8] Beckman, M. E.: *Stress and non-stress accent*, Foris Publications, Dordrecht, 1986.

[9] Daneš, F.: *Intonace a věta ve spisovné češtině*, Nakladatelství ČSAV, Praha, 1957.

[10] Petr, J. a kol.: *Mluvnice češtiny – Vol. I*, Academia, Praha, 1986.

[11] Palková, Z.: *Fonetika a fonologie češtiny*, Karolinum, Praha, 1994.

[12] Matoušek, J., Tihelka, D., Romportl, J.: Current state of Czech text–to–speech system ARTIC, *Proc. of the 9th International Conference TSD 2006, Lecture Notes in Artificial Intelligence*, Vol. 4188. Springer Berlin/Heidelberg, 2006, p. 439–446.

[13] Tihelka, D., Kala, J., Matoušek, J.: Enhancements of Viterbi Search for Fast Unit Selection Synthesis, *Proc. of 11th Interspeech 2010*, p. 174–177, Makuhari, 2010.

[14] Boersma, P., Weenink, D.: *Praat: doing phonetics by computer* (Version 5.4.08). Downloaded from `http://www.praat.org`.

[15] Volín, J., Tykalová, T., Bořil, T.: Stability of prosodic characteristics across age and gender groups, *Proceedings of Interspeech 2017*, p. 3902–3906, Stockholm, 2017.

[16] Volín, J.: *Downtrends in standard British English intonation*. Hector, Frankfurt am Main, 2008.

Měření tempa řeči u dětí

Measuring Speech Rate of Children

Jan Vimr

České vysoké učení technické v Praze – Fakulta elektrotechnická, Technická 2, 160 00 Praha 6
vimrjan@fel.cvut.cz

The issue of automatic measuring of speech rate by detecting syllable nuclei in utterances is discussed in this paper. Automatic measurements are necessary for analysing large databases of utterances where manual measurement would take significant time. A small database of 60 utterances by children in age group from 5 to 16 years was used to compare number of syllables counted by human with selected methods for automatic detection of syllable nuclei, namely Praat script, Recognizer VUT, Modified Recognizer VUT and our own detector. They are compared on the basis of mean difference, standard deviation and Pearsons correlation coefficient. The conclusion is that the most accurate of the tested methods for syllable nuclei detection is the Modified Recognizer VUT.

1. Úvod

Tempo řeči nebo také mluvní tempo (anglicky speech rate nebo speaking rate) je často zkoumaný parametr při diagnostice poruch řeči nebo při analýze věkové závislosti řeči. Jeho odhad lze také využít pro lepší nastavení řečových rozpoznávačů, které mohou mít problémy s velmi rychlými nebo naopak velmi pomalými promluvami. Tempo řeči bývá obvykle udáváno jako počet řečových jednotek za jednotku času, nejčastěji v slabikách nebo fónech za sekundu. Na rozdíl od artikulačního tempa (anglicky articulation rate) se určuje v celé promluvě, ne jen v plynulých úsecích. Tím pádem jsou započítány i delší pauzy, hezitační zvuky, přeřeknutí, zakoktání apod.

Měření tempa řeči manuálně je časově velmi náročné, a proto byla navržena řada metod, které mají za úkol určit tempo řeči automaticky. Algoritmy, které hledají lokální maxima ve vyhlazeném průběhu krátkodobé energie nižších kmitočtových pásem, byly navrženy například v publikacích [1, 2]. Tyto postupy vycházejí z faktu, že energie řečového signálu je typicky vyšší ve znělých úsecích, tedy zejména v samohláskách. Vzhledem k tomu, že samohlásky tvoří jádra většiny slabik, bude počet nalezených lokálních maxim v průběhu energie přibližně odpovídat počtu slabik v promluvě. Algoritmy lze rozšířit o zkoumání dalších parametrů řečového signálu jako například krátkodobá amplituda, počet průchodů nulou, krátkodobá autokorelace, jak bylo ukázáno v [3]. Volně dostupný algoritmus, který kombinuje hledání lokálních maxim v krátkodobé intenzitě signálu a přítomnost základní hlasivkové frekvence, napsaný v programu Praat [4], byl publikovaný v [5].

Pokročilejší metody pro automatické měření tempa řeči zahrnují například natrénování modelů Gaussovských směsí (GMM) pro určování, do které kategorie tempa řeči – pomalá, střední, rychlá – promluva patří, což bylo publikováno v [6]. Další možností je využití neuronových sítí, jak bylo ukázáno v [7], kde jsou pomocí hluboké neuronové sítě (DNN) rozdělovány jednotlivé segmenty řeči do čtyř kategorií – pomalá, střední nebo rychlá řeč a ticho. Dalším příkladem pokročilejší metody je algoritmus, který hledá hranice jednotlivých fonémů na základě změn v mel-frekvenčním kepstru (MFCC), který je použit v [8].

Cílem tohoto článku je porovnat vybrané volně dostupné metody pro automatický odhad tempa řeči na databázi dětských promluv. První z nich je již zmiňovaný Praat skript [5]. Dále byl využit fonémový rozpoznávač navržený na VUT v Brně, publikovaný v [9], jehož výstupem jsou jednotlivé fonémy promluvy, které byly dále analyzovány dvěma způsoby v prostředí MATLAB [10], což je dále popsáno v kapitole metody. Nakonec byl na základě postupů z publikací [1–3, 5] navržen vlastní detektor slabičných jader, který kombinuje výkon signálu a počet průchodů nulou.

2. Metody

2.1. Databáze

Pro testování byly použity nahrané promluvy od 60 dětí (31 chlapců a 29 děvčat) ve věku od 5 do 16 let. Vybrány byly z větší databáze tak, aby byly děti v uvedeném věkovém rozmezí zastoupeny pokud možno rovnoměrně. Obsahem nahrávek je popis obrázku, který zobrazuje soubor činností malého chlapce před cestou do školy. U promluv byl ručně spočítán počet slabik, což bylo použito jako reference pro porovnání přesnosti zkoumaných automatických metod. Použité promluvy jsou dlouhé několik desítek sekund a obsahují mezi 22 a 126 slabikami (medián 51). Promluvy neobsahují pouze plynulou řeč, ale jejich součástí jsou i řečové pauzy, přeřeknutí, hezitační zvuky apod.

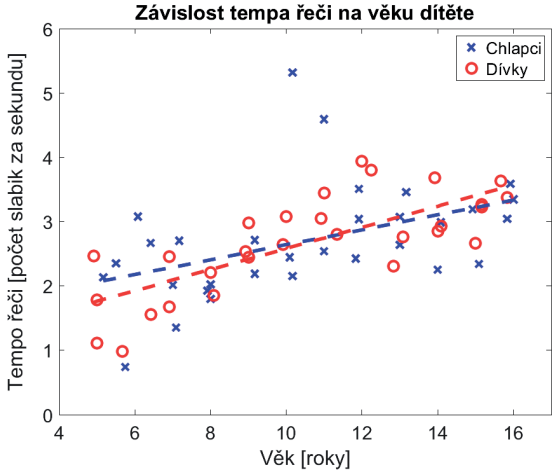
2.2. Manuální

Jako reference pro všechny zkoumané metody na automatické detekování slabičných jader bylo použito manuální počítání slabik v jednotlivých promluvách. To vyžadovalo několikanásobný poslech všech promluv s častým zastavováním. Hlavní nevýhodou této metody je velká časová

Přijato 25. listopadu 2019, akceptováno 23. prosince 2019.

náročnost, která roste s velikostí databáze. Navíc je měření ovlivněno lidským faktorem, kdy posluchač rozumí i zkomoleným slovům apod.

Touto metodou byly spočítány všechny slabiky, a to včetně těch, které byly součástí přeřeknutí, zakoktání apod. Následně bylo vypočteno průměrné tempo řeči v každé promluvě ze zjištěného počtu slabik a známé délky promluvy. Udáváno pak bylo jako počet slabik za sekundu. Na obrázku 1 jsou vyneseny vypočtené hodnoty, které jsou proloženy lineárními modely, zvlášť pro chlapce a dívky. Na základě dvouvýběrového t-testu nebyla prokázána závislost na pohlaví dítěte: $t(59) = 0{,}12$; $p = 0{,}91$. Hodnoty tempa řeči vykazují závislost na věku dítěte. Pearsonův korelační koeficient tempa řeči a věku bez ohledu na pohlaví dítěte vychází: $r = 0{,}59$; $p < 0{,}001$.



Obrázek 1: Závislost tempa řeči na věku

2.3. Praat skript [5]

První zkoumanou metodou pro automatické měření tempa řeči je skript pro detekci slabičných jader napsaný v programu Praat [4], publikovaný v [5]. Jelikož jsou slabičná jádra nejčastěji tvořena samohláskami, jsou v signálu hledány znělé úseky. Ty se vyznačují zpravidla vyšší intenzitou a přítomností základní hlasivkové frekvence. Skript proto zkoumá průběh intenzity v signálu rozděleném na segmenty o délce 64 ms s krokem 16 ms. V průběhu intenzity jsou nalezena lokální maxima, která přesahují zvolený práh, určený jako medián ze všech intenzit, a zároveň jim předchází pokles minimálně o 2 dB. Z nich jsou vybrána pouze ta lokální maxima, která se nacházejí ve znělých úsecích signálu. To je vyhodnoceno na základě přítomnosti základní hlasivkové frekvence, hledané pomocí autokorelace jednotlivých segmentů o délce 100 ms s krokem 20 ms. Zbylá maxima jsou považována za slabičná jádra. Výstupem skriptu je textový soubor, kde jsou zaznamenány časy nalezených slabičných jader. Ze známé délky promluvy pak lze spočítat průměrné tempo řeči v promluvě.

2.4. Rozpoznávač VUT [9]

Druhá testovaná metoda využívá automatickou segmentaci signálu na jednotlivé fonémy pomocí fonémového rozpoznávače založeného na dlouhém časovém kontextu. Ten využívá hybridní systém, který kombinuje skryté Markovovy modely (HMM) a umělé neuronové sítě (ANN). Výstupem z rozpoznávače pro každou promluvu je tabulka udávající jednotlivé fonémy a jejich časové rozmezí v dané promluvě. Určení tempa řeči probíhá pomocí skriptu v prostředí MATLAB [10], kam je načten výstup rozpoznávače. Z něj je následně určen počet rozpoznaných samohlásek v promluvě, ten přibližně odpovídá počtu slabičných jader.

2.5. Modifikovaný rozpoznávač VUT [9]

Další metoda vychází ze stejného fonémového rozpoznávače, nicméně interpretace jeho výstupu je rozdílná a vede k výsledkům, které jsou o něco lepší. Zavedením několika jednoduchých pravidel, viz tabulka 1, je zde korigován celkový počet nalezených slabik v promluvě. Vychází se zde z předpokladu, že výstup rozpoznávače nebude zcela bezchybný, a tedy celkový počet nalezených samohlásek nemusí přesně odpovídat počtu slabik. V rozpoznávači někdy kvůli nepřesné segmentaci dojde k rozdělení samohlásky na dvě, což je třeba korigovat. Podobně je někdy dvojhláska /au/ nebo /ou/ chybně rozdělena na dvě části, ale je třeba je počítat jako jedno slabičné jádro.

Dalším problémem je fakt, že rozpoznávač neumí rozlišit souhlásky /r/, /l/, /m/ od jejich slabikotvorných variant /r̩/, /l̩/, /m̩/. Ty mohou v českém jazyce tvořit jádro slabiky a proto je s nimi nutné počítat. To je jeden z důvodů, proč bylo zavedeno pravidlo, že pokud jsou výstupem rozpoznávače tři po sobě jdoucí souhlásky, jsou započítány jako slabika. Dalším důvodem je to, že může jít o chybu, kdy rozpoznávač neodhalil samohlásku. Toto pravidlo by mohlo způsobit problém u slov, kde se vyskytují tři po sobě jdoucí souhlásky, které ale slabiku netvoří. Taková slova se v použitých promluvách téměř nevyskytovala, ale při dalším použití této metody je to potřeba vzít v úvahu.

/VV/ ⟶ /V/	Dvě po sobě jdoucí stejné samohlásky jsou brány jako jedna samohláska
/au/ ⟶ /a͡u/	Dvojice samohlásek „au“ je brána jako jedna dvojhláska
/ou/ ⟶ /o͡u/	Dvojice samohlásek „ou“ je brána jako jedna dvojhláska
/CCC/ ⟶ /syl/	Tři po sobě jdoucí souhlásky jsou brány jako slabika

Tabulka 1: Pravidla modifikovaného rozpoznávače

Pravidla jsou implementována ve skriptu v prostředí MATLAB [10], do kterého je načten výstup fonémového rozpoznávače. Je zjištěn celkový počet slabičných jader

jako počet nalezených samohlásek, korigovaný podle výše uvedených pravidel.

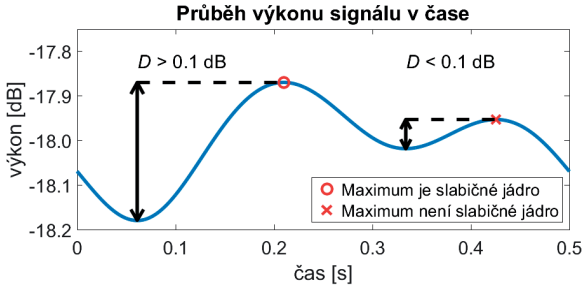
2.6. Vlastní detektor slabičných jader

Na základě postupů uvedených v [1–3, 5] byl navržen vlastní detektor slabičných jader, naprogramovaný v prostředí MATLAB [10]. Vychází z předpokladu, že výkon řečového signálu je výrazně vyšší u samohlásek, tedy pomocí průběhu krátkodobého výkonu lze nalézt jednotlivá slabičná jádra, která jsou nejčastěji tvořena samohláskami. Dalším sledovaným parametrem je počet průchodů nulou, který je nejvyšší v neznělých částech promluvy, zejména u sykavek a v řečových pauzách. Oba parametry jsou často využívány v číslicovém zpracování signálů a jsou popsány například v publikaci [11].

Řečový signál byl nejprve segmentován na úseky o délce 10 ms s krokem 5 ms. V každém segmentu byl spočítán výkon P_i a počet průchodů nulou Z_i. Průběhy P a Z byly následně vyhlazeny mediánovým filtrem o délce 10 segmentů. Poté byla hledána lokální maxima v průběhu P větší než zvolený práh P_{th} s tím, že v úvahu nebyly brány úseky, kde hodnota Z byla větší než práh Z_{th}. Tedy segment musel splňovat podmínku:

$$(P > P_{\mathrm{th}}) \,\&\, (Z > Z_{\mathrm{th}}). \tag{1}$$

U lokálních maxim v průběhu P, která splňovala předchozí podmínku, byl dále sledován parametr D, který určoval rozdíl výkonu mezi maximem a minimem v úseku od předchozího maxima, viz obrázek 2.



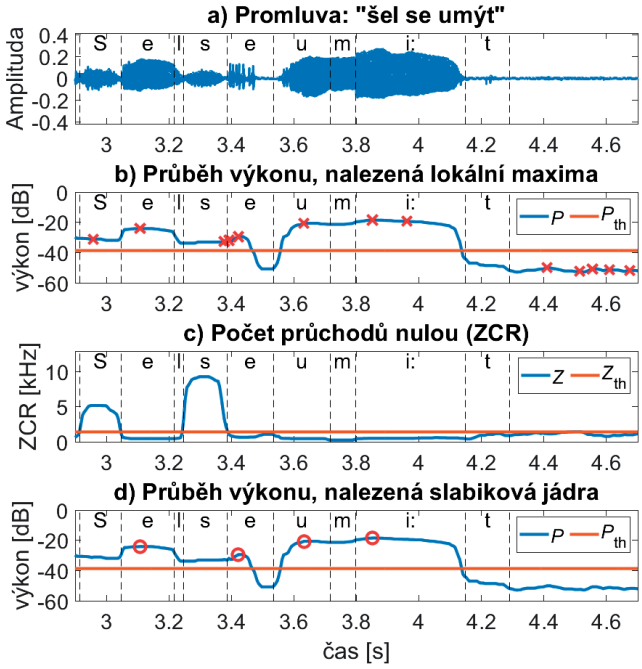
Obrázek 2: Parametr D

Aby bylo maximum považováno za slabičné jádro, musela být splněna podmínka:

$$D > 0{,}1\,(\mathrm{dB}). \tag{2}$$

Zavedení tohoto parametru výrazně přispívá ke zpřesnění algoritmu. Hraniční hodnota byla určena na základě série experimentů, aby bylo dosaženo co největší přesnosti algoritmu.

Na obrázku 3 je naznačen princip vlastního detektoru v několika krocích. Na obrázku 3 a) je zobrazen krátký úsek promluvy s ručně označenými hranicemi fonémů. Na obrázku 3 b) je vynesen vyhlazený výkon signálu s nalezenými lokálními maximy. Na obrázku 3 c) je průběh vyhlazeného počtu průchodů nulou v jednotlivých segmentech



Obrázek 3: Princip vlastního detektoru

a na obrázku 3 d) jsou nalezená slabičná jádra, tedy maxima z obrázku 3 b), která přesahují hodnotu P_{th}, předchází jim pokles minimálně o 0,1 dB a hodnoty Z jsou v daném úseku menší než Z_{th}.

3. Porovnání metod

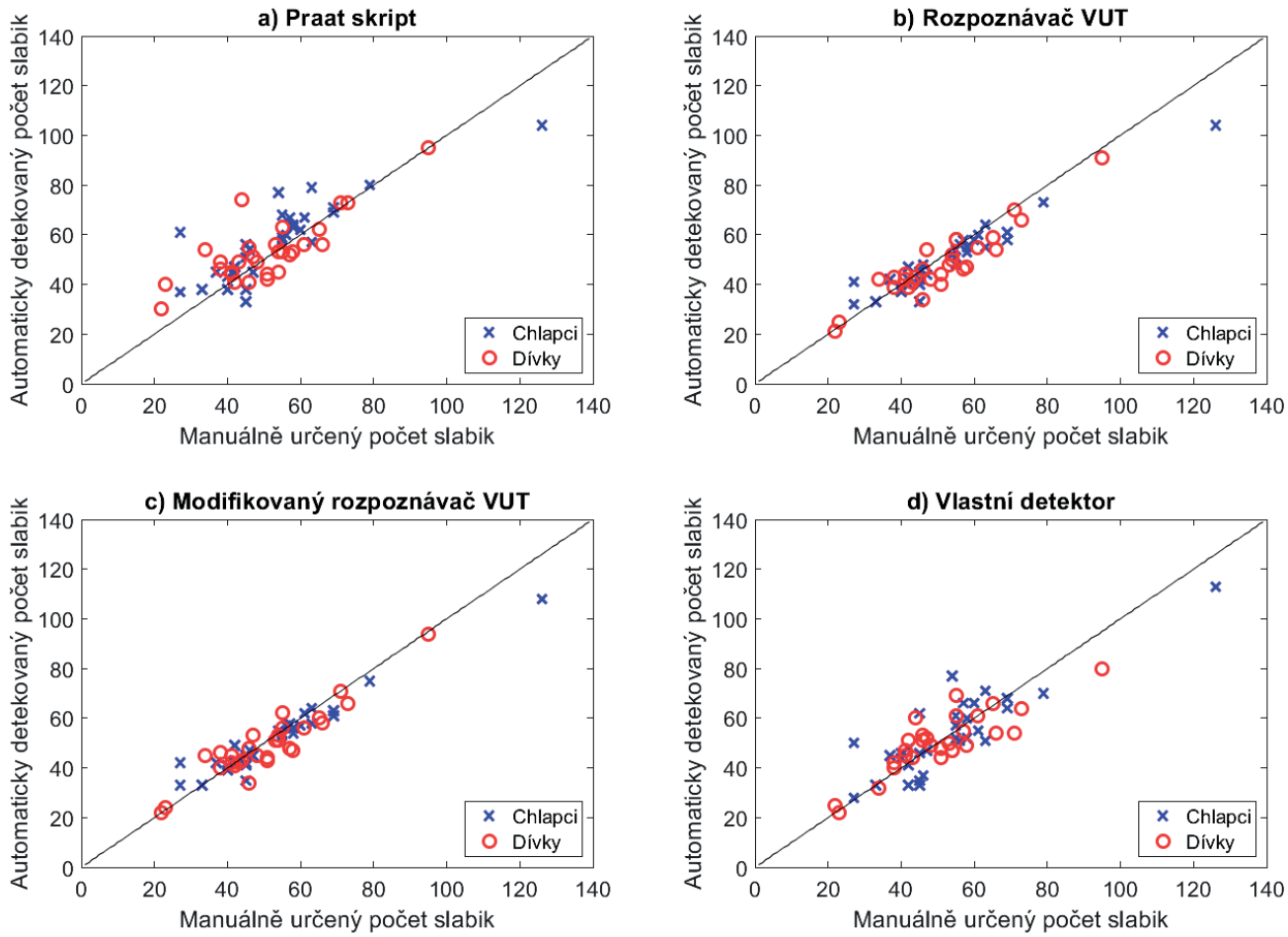
Výsledky použitých metod byly porovnány s ručně naměřenými hodnotami pomocí průměrné odchylky Δ, střední kvadratické odchylky σ a Pearsonova korelačního koeficientu r. Porovnávány byly počty slabik nalezené v jednotlivých promluvách. Výsledky viz tabulka 2. Porovnání jednotlivých metod je také vyneseno na obrázku 4.

Metoda	Δ	σ	r
Praat skript	3,57	10,13	0,82
Rozpoznávač VUT	−2,60	6,46	0,94
Modif. rozp. VUT	−1,25	5,70	0,95
Vlastní detektor	0,18	8,64	0,85

Tabulka 2: Statistické porovnání metod

Počty slabik nalezené pomocí Praat skriptu vycházejí systematicky vyšší než referenční hodnoty a mají poměrně vysoký rozptyl, viz tabulka 2. Pearsonův korelační koeficient vychází $r = 0{,}82$, což je méně než hodnota $r = 0{,}88$, kterou pro celé promluvy uvádějí autoři skriptu v [5]. Rozdílný výsledek může být zapříčiněn mnoha faktory: byly použity různé promluvy v různých jazycích, a navíc různě veliké databáze.

Výsledky rozpoznávače VUT vycházejí spíš nižší než reference. Hodnoty mají proti Praat skriptu menší rozptyl



Obrázek 4: Korelace automatických měření s manuálním

a vyšší Pearsonův korelační koeficient. Modifikovaný rozpoznávač navíc dává ještě lepší výsledky, jelikož se pomocí zavedených pravidel podařilo snížit průměrnou odchylku i rozptyl a zvýšit korelační koeficient. Výsledky této metody jsou tak nejbližší hodnotám spočítaným člověkem ze všech testovaných metod.

Vlastní detektor slabikových jader má sice nejmenší průměrnou odchylku, ale co se týče rozptylu a korelace vycházejí o něco lepší výsledky než u Praat skriptu, ale horší než u rozpoznávače VUT.

4. Závěr

Za nejpřesnější metodu můžeme prohlásit Modifikovaný rozpoznávač VUT, kde korelace s ručně změřenými hodnotami vychází $r = 0{,}95$. Další v pořadí je Rozpoznávač VUT, dále vlastní detektor a nejhůře vychází Praat skript.

Z ručně měřených dat je patrná závislost tempa řeči na věku dítěte. Dalším cílem ve výzkumu této problematiky by mělo být podrobnější zkoumání této závislosti na větších databázích promluv, k čemuž by bylo vhodné použít některou automatickou metodu. Z metod zkoumaných v tomto článku by byl nejvhodnější Modifikovaný rozpoznávač VUT, jelikož jeho výsledky jsou nejbližší ručně odečteným hodnotám.

Poděkování

Výzkum je podporován z grantu GAČR „Populační normy akusticko-fonetických charakteristik dětské řeči“ (19-20887S).

Reference

[1] Pfitzinger, H. R., Burger, S., Heid, S.: Syllable detection in read and spontaneous speech, *Proceeding of Fourth International Conference on Spoken Language Processing*, ICSLP ’96, vol. 2, pp. 1261–1264, 1996

[2] Pfau, T., Ruske, G.: Estimating the speaking rate by vowel detection, *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP ’98, vol. 2, pp. 945–948, 1998

[3] Jalil, M., Butt, F., Malik, A.: Short-time energy, magnitude, zero crossing rate and autocorrelation measurement for discriminating voiced and unvoiced

segments of speech signals, *The International Conference on Technological Advances in Electrical, Electronics and Computer Engineering (TAEECE)*, pp. 208–212, 2013

[4] Boersma, P., Weenink, D.: Praat: doing phonetics by computer (Version 6.1) [Computer program], `www.praat.org`

[5] Jong. N. H. de, Wempe T.: Praat script to detect syllable nuclei and measure speech rate automatically, Behavior research methods, vol. 41, no. 2, pp. 385–390, 2009

[6] Faltlhauser, R., Pfau, T., Ruske, G.: On-line speaking rate estimation using Gaussian mixture models, *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, pp. 1355–1358, 2000

[7] Tomashenko, N., Khokhlov, Y.: Speaking Rate Estimation Based on Deep Neural Networks, *International Conference on Speech and Computer*, pp. 418–424, 2014

[8] Aharonson, V., Aharonson, E., Levi, K., Sotzianu, A., Amir, O., Zehava, O. B.: A Real-Time Phoneme Counting Algorithm and Application for Speech Rate Monitoring, *Journal of Fluency Disorders*, vol. 51, pp. 60–68, 2017

[9] Schwarz, P.: Phoneme recognition based on long temporal context, Disertační práce, Brno, Vysoké učení technické v Brně, Fakulta informačních technologií, 2008

[10] MATLAB: mathematic aplication (Version R2018b) [Computer program], `http://www.mathworks.com`

[11] Uhlíř, J., Sovka, P., Pollák, P., Hanžl, V., Čmejla, R.: Technologie hlasových komunikací, Nakladatelství ČVUT, 2007

Akustické listy: ročník 25, číslo 3–4 prosinec 2019 ISSN: 1212-4702
Vydavatel: Česká akustická společnost, z. s., Technická 2, 166 27 Praha 6
Počet stran: 16 Počet výtisků: 200
Redakční rada: M. Brothánek, O. Jiříček, R. Čmejla, J. Volín
Jazyková úprava: R. Svobodová, M. Tharp

Uzávěrka příštího čísla Akustických listů je 30. dubna 2020.