2014 ACTA UNIVERSITATIS CAROLINAE PAG. 9–19
PHILOLOGICA 1 / PHONETICA PRAGENSIA XIII

GENERAL AND SPEAKER-SPECIFIC PROPERTIES OF F0 CONTOURS IN SHORT UTTERANCES

JAN VOLÍN, TOMÁŠ BOŘIL

ABSTRACT

This study compares three major quantitative methods of contour analysis with the aim to establish their merit for intonation research. Utterances of 24 speakers (7–9 syllables long) taken from short dialogues were used to see whether general prosodic patterning determined by the intonational grammar of the language and the individual production habits of the speakers can be captured by computational means. The three methods exploited were: k-means clustering (KMC), polynomial regression analysis (PRA), and functional principal component analysis (FPCA). The numerical outputs of the methods are confronted with human perception of the contour in both auditory and visual domains. The results suggest that the observed contour properties are reflected by all three methods reasonably well: phonetically interpretable outcomes can be achieved by each of them. As to speakers' individual features, KMC seems to be least vulnerable to spurious effects.

Key words: melody of speech, cluster analysis, speaker identity, recognition, fundamental frequency

1. Introduction

Melody of speech or intonation in the narrow sense represents a very conspicuous attribute of the spoken forms of languages. The modulation of the speech signal in the domain of the fundamental frequency (F0) produced by the vocal folds of a speaking individual encodes meanings that are intended to be understood. In other words, the melodic rises and falls in our messages are not random and are produced with a purpose. They fulfil a number of important roles or functions in the sound structure of speech (for detail see, e.g., Ladd, 1996; Cruttenden, 1997; Gussenhoven, 2004).

The *lexical function* is often set aside in linguistic descriptions since, unlike the others, it is not superposed over the intellectual content of the word – together with segmental phonemes it creates the primary meaning. Another reason for the exclusive treatment of this function might be the fact that it is very uncommon in the languages of European origin and contemporary linguistics is dominated by European tradition. The *gram-*

matical function of intonation, on the other hand, has always received a great deal of attention. Consequently, the central part of the linguistic descriptions is usually occupied by nuclear accents (or melodemes in the Czech terminology). They carry the largest proportion of information about the grammatical status of the syntagmatic unit. On the other hand, there is the holistic approach to melodic contours, which stresses the relative importance of the stretches of speech outside the intonational nucleus. This approach is typical of the research that takes into consideration pragmatic meanings. Pragmatics of the speech is also closely associated with the *affective function* of intonation. Moods, attitudes, emotions, and interpersonal stances are reflected in the melodic contours as well, although not as transparently as it used to be suggested. Recent attempts to focus linguistic enterprise on realistic communication brought to life the notion of the *discourse function*, which can be studied in conversational speech. One of the areas of the discourse behaviour which had been studied long before the introduction of discourse analysis is that of the broad and narrow focus. The *accentual function* of intonation is thus a construct very firmly nested in the realm of analyses of real-life dialogues. To close the circle, the *indexical function* of intonation needs to be introduced, which like the lexical function mentioned at the outset is often disregarded in traditional descriptions, this time because of its extralinguistic nature. Whether within the space demarcated by linguists or not, some indexical information is present in every spoken message. By listening to it we can estimate with certain probability the gender, age, geographical origin and other socially relevant characteristics of the speaker.

Manifestations of the indexical function together with the individual habits (personal idiosyncrasies) of the speaker serve to establish his or her identity when the visual contact is not possible. We may recognize a familiar human being in the dark, behind a wall, on the telephone or from a speech recording. Apart from its obvious use in everyday life, this fact is exploited in one of the most common forensic tasks, which is the speaker identification or speaker discrimination.

Interestingly, though, the current forensic technology does not use linguistically informed procedures to a great extent. One of the most advanced systems that takes intonational correlates into account is SAUSI (*Semi-Automatic Speaker Identification*) by Harry Hollien and his team (e.g., Hollien, 2002). It works with geometric mean of fundamental frequency (F0), ratio of voiced speech duration to the overall duration of the sample, standard deviation of all F0 measurements within the sample, and information about the incidence of relevant frequencies in semitone bands of the overall range. Unfortunately for an intonologist, this numerical approach cannot convey any information about the shape of the melodies used by the speaker.

Table 1. Typology of contours based on fundamental frequency (F0) and sound pressure level (SPL) movements after Adami, 2007 (the 5th type was a voiceless stretch).

	Type 1	*Type 2*	*Type 3*	*Type 4*
F0	rise ↗	rise ↗	fall ↘	fall ↘
SPL	rise ↗	fall ↘	rise ↗	fall ↘

Adami (2007) suggested the introduction of some reflection of the actual shapes into the forensic measurements. He observed slopes of F0 and SPL (sound pressure level) contours simultaneously and identified turning points in the contours in order to delimit stretches of speech signal with rising or falling F0 and SPL values. The two parameters (fundamental frequency and SPL) and two directions of change (rising and falling) produced four types of stretches as illustrated in Table 1. The fifth type was introduced for voiceless segments, in which no F0 could be detected.

Adami summarized the frequency of occurrence and mutual position of these five types in a computationally complex algorithm to capture individual characteristics of speakers in his sample and, indeed, this approach led to the improvement in the recognition rate relative to the previous methods. Yet again, although the properties of the contours are taken into account, we are not getting any useful information from the linguistic point of view.

With the explicit objective to find "bridges between intonational phonology and speech technology", Grabe, Kochanski and Coleman hand-labelled a number of sentences spoken in seven different British accents of English (Grabe et al., 2007: 281). They decided to use third-order polynomials to see if the resulting coefficients could be helpful in capturing the differences between the accents. Their attempt was reasonably successful in both respects: a) the regional accents could be differentiated, and b) the coefficients of polynomial functions are fairly interpretable.

A method that is computationally even more advanced was proposed by Zellers, Gubian and Post (2010). These researchers studied melodic shapes superposed over one word in different utterances produced by different speakers with the help of functional data analysis (FDA), or more specifically, with the functional principal component analysis (FPCA). The resulting principal components were curves and a set of scores quantified the similarity of the individual F0 contours to the principal curve shapes. An operative classification of contours was generated, which should encourage further research into the merit of this method.

There also exists a method computationally simpler, but conceptually more transparent. K-means clustering (KMC) uses fewer F0 values to express the fundamental frequency course in a unit: it can be one or two values per syllable depending on the duration of the sonorous peak of the syllable (Hermes, 2006). This method was exploited successfully, for instance, by Volín (2008) who tested the classification of continuation rises proposed in older literature against real speech data, and showed that some of the previous impressionistically based claims are not supported by a corpus of current recordings. Although technically the KMC takes the F0 contour as a point in an n-dimensional space, the results are relatively easy to interpret. The method, however, cannot be used reliably for longer contours. The recommended maximum of dimensions is about eight for material of tens of cases.

The aim of our present study is, therefore, to compare the existing approaches to F0 tracks and consider their performance in comparison with the human evaluation of melodic contours with a particular interest in a method that is both informative to a linguist and effective to a forensic practitioner.

2. Method

Speech samples were taken from the Mini-Dialogue part of the Prague Phonetic Corpus. The recordings of scripted dialogues that were rehearsed and performed by students of philological programmes at Charles University in Prague mimic short natural dialogues of about four to five turns. Recordings were made in a studio at 32-kHz sampling rate with 16-bit resolution. There were 24 speakers at our disposal and for the purpose of our probe we selected the middle turn of two five-turn dialogues. These dialogues were performed amidst other spoken material in a block-randomized design, separated from each other by more than 12 other dialogues. The middle turn which was chosen for analysis was almost identical in both dialogues, with the difference in grammatical number: set A was in singular, set B in plural forms. The wording of each turn was: *Řekneš jim, co si myslíš* (*You'll tell them what you think*) and *Řeknete jim, co si myslíte* (English translations are identical – *You'll tell them what you think*). The contexts of the singular and plural forms (i.e., the surrounding dialogue turns) differed.

Our 48 utterances (2 by each speaker) were manually labelled and the F0 track was extracted in Praat (Boersma and Weenink, 2010). All F0 trajectories were inspected and manually corrected where necessary (esp. octave jumps and non-modal phonation problems).

In order to compare the singular and plural form of the same utterance, the second and third syllables in the plural were averaged (mean difference between them was 1.3 ST) and so were the fifth and sixth syllables (mean difference between them was 0.4 ST).

First of all, auditory and visual assessment of the contours was carried out by the authors of this article. Together with it, the sensations of pragmatic feel of the utterances was captured verbally. The computational analysis comprised k-means clustering (KMC), polynomial regression analysis (PRA), and one member of the family of functional data analyses – the functional principal component analysis (FPCA). The use of PRA for intonation contours was advocated, for instance, by Grabe, Kochanski and Coleman (2007), while the principles of FPCA in analysis of F0 tracks are described in Zellers, Gubian and Post (2010).

K-means cluster analysis was performed with the software STATISTICA 7.1 (Statsoft, 2005) with the initiation by maximum distances between centroids and no more than ten iterations. The same software was used for computations of polynomial functions that approximated the F0 tracks. We used third order polynomials, i.e., with two changes in the direction of the curve. FPCA was performed in MATLAB 2012a. The cluster analysis used F0 measurements taken in the middle of each of the syllables (with special treatment of plural forms – see above). The mean F0 from each sentence was subtracted to normalize for individual pitch register. Since FPCA analyzes individual curves, additional normalization in time was also necessary. This was achieved by time-axis warping with manually labelled boundaries between vowels and consonants as anchors. Similarly, polynomial functions are sensitive to the distance of the beginning of the curve from zero. It is especially their intercept that is affected by the timing of the first voiced frame, but the other coefficients can also change considerably if the utterance starts later after the time measuring starts. Therefore, the time warping was applied as well. The k-mean clustering of one value per syllable, on the other hand, is time insensitive – the flow of time is substituted for by the ordinal position of the syllables.

3. Results

Perceptual analysis had to cope with considerable variation added by the temporal properties of the utterances and the voice quality differences. However, repeated listening allowed for abstracting from these attributes and revealed two general pitch contours stretching over the seven syllables of the utterance (nine for the plural form). Contour *a* was more casual and had a stronger feel of definiteness, while contour *b* sounded more marked and implicational. Other pragmatic dimensions like strictness, degree of involvement, pledging, etc., were noted but could not be quantified as to the contribution of the contour itself and other phonetic features.

The most notable difference in the shape of the two contours rested in the highest point of the profile. Contour *a* rose to the third syllable and started gradual descent with steeper fall at the end, whereas contour *b* kept steady moderate rise towards the penultimate syllable from which it fell sharply to the base of the speaker's voice (see Figure 1).

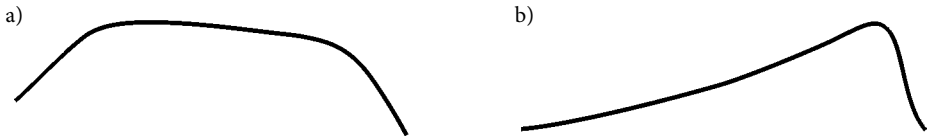

Figure 1. A diagram of two basic contours distinguished by perceptual analysis.

The computational analyses will be presented from the order of complexity, i.e., KMC, PRA, and FPCA. K-means cluster analysis suggested two groups that were similar to those proposed by the authors after visual and auditory inspection (see Figure 2). When required to produce three clusters, the algorithm did not produce any results worth

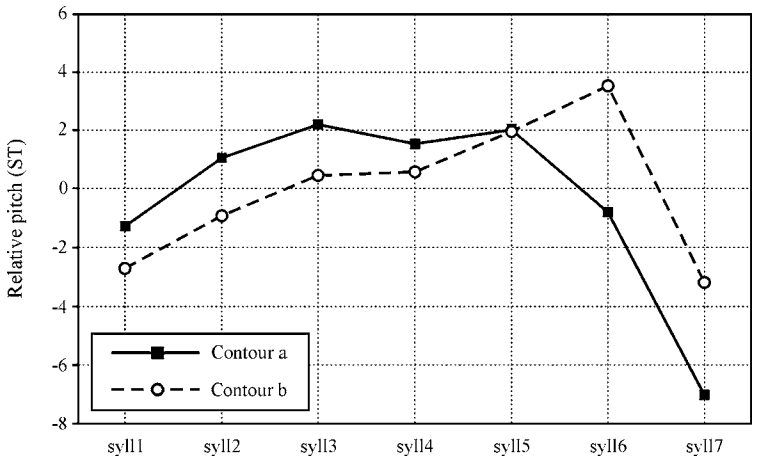

Figure 2. Two contours resulting from k-means cluster analysis. The x-axis represents individual syllables, while the y-axis refers to their relative pitch in semitones.

following – the third cluster was a mere combination of the two original shapes and represented only a few cases. Therefore, we accepted the two-contour result.

The general shape of the cubic polynomial functions exploited in the present analysis is

$$y = a + b_1x + b_2x^2 + b_3x^3.$$

The intercept *a* is related to the initial pitch of the curve (thanks to time normalization quite reliably). The variable *y* in our case represents F0, whereas *x* refers to time. Figure 3 demonstrates a typical F0 contour *b* (see above) which was approximated with the following polynomial function:

$$\text{F0} = 14.8 - 34.1t + 110.2t^2 - 69.7t^3.$$

The intercept *a* only captures the approximate F0 of the initial voiced sample, but the coefficients b_1 to b_3 already describe the shape of the curve. Therefore, we carried out a discriminant analysis (DA) to see if the coefficients could be used as reliable descriptors. DA with b_1 and b_2 recognized perceptually classified contours with the success rate of 89.6%, which is well above chance: $F(2.45) = 17.05$; $p < 0.0001$ (Wilks $\lambda = 0.57$). Grammatical category (singular vs plural), on the other hand, could not be discriminated: $F(2.45) = 1.67$; $p = 0.199$ (Wilks $\lambda = 0.93$). This is a desired outcome, since our experiment counts on considerable similarity between singular and plural utterances. Adding coefficient b_3 into the DA did not improve the success rate. It actually worsened it negligibly to 87.5%, which means one more case mistakenly assigned to a group: $F(3.44) = 15.36$; $p < 0.0001$ (Wilks $\lambda = 0.49$). Table 2 shows the resulting confusion matrix for two contours and two discriminant coefficients.

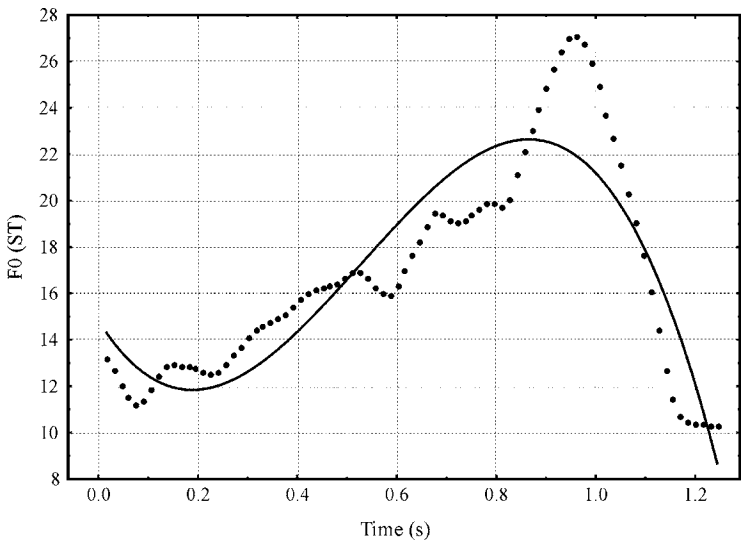

Figure 3. An example of a polynomial fit to an F0 contour of the *b* type.

Table 2. Confusion matrix for discriminant analysis of two contours with two predictors. Success rate expressed as a percentage of correctly classified cases. (For contour *a* and *b*, see above.)

	Predicted as a	*Predicted as b*	*Success rate*
Observed *a*	31	2	93.9
Observed *b*	3	12	80.0
Total	34	14	89.6

The outcome of FPCA comprises principal components (functions or curves) that capture the main properties of the individual F0 contours, and scores that are associated with each of the original items. Visualisation of the nature of the principal components found is displayed in Fig. 4. Only two components are shown: PC1, which explains about 68% of the variance, and PC2, which explains additional 17% of the variance. The explanatory power of these two components is 85%.

The mean function is the curve obtained by averaging all the original curves. The output scores can serve to reconstruct the original trajectories. In our case, the positive scores for PC1 point to contour *b* (see above): a steady rise towards the penultimate syllable followed by the final fall. The negative scores seem to be related to contour *a* with the longer decline from the summit on the third syllable.

The interpretation of the second principal component is slightly less straightforward. Our understanding of PC2 is that it helps to correct for higher beginnings and ends of the contours and also to adjust the F0 of the syllables in the mid part of the utterance. The corrections of edges and middle parts of the contours work in opposition, so the positive scores produce flatter curves, while the negative ones reflect greater pitch range on the contour.

Compared with PC1, the scores of the second principal component do not seem to contribute to the recognition of the perceptually established contours. This is obvious from Figure 5 where the boxes of PC2 for both contour types overlap.

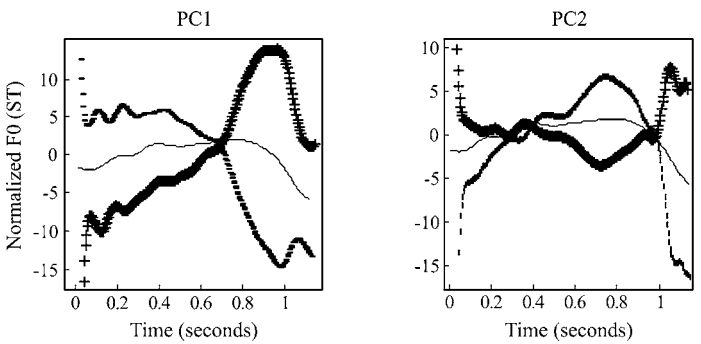

Figure 4. Principal components PC1 and PC2 as variations with respect to the mean. Normalized time on x-axis, normalized pitch (semitones) on y-axis. The mean function is the thin line in the middle, the curves from + and the − signs correspond to shapes of contours with positive or negative PC scores, respectively (±2 SD off the mean).

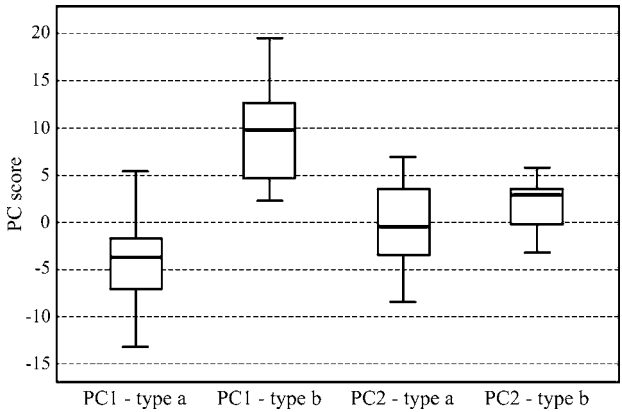

Figure 5. Boxplots of PC1 and PC2 values for perceptually established contours of type *a* and *b*. The mid line represents the median, boxes themselves quartiles and the whiskers mark the variation range.

Our final analysis concerned the speakers' consistency. We wanted to know how many subjects would produce the contour of similar shape in both utterances, i.e., the speaker who produced a melody of a certain type in singular performed likewise in plural, although in a different context and after talking about other things (i.e., not in immediate succession). To find out, the descriptors of the computational analyses were clustered by the k-means method. The results are summarized in Table 3 together with the outcome of human evaluation. It is apparent that individual methods do not produce identical results, but on average, about two thirds of the speakers incline to the same contour in both utterances. Polynomial coefficients returned the same result as human perception but closer look at individual cases reveals that it is not the same 15 (or 9) speakers that were marked as consistent (or inconsistent). This clearly leads to our final question: How many speakers were evaluated in the same way by all methods? The answer is summarized in Table 4 below.

Table 3. Numbers of speakers falling in the same (*Consistent*) or different (*Inconsistent*) categories for both their utterances in each of the analyses (HUM = human perception, KMC = k-means clustering, PRA = polynomial regression analysis, FPCA = functional principal component analysis).

	HUM	*KMC*	*PRA*	*FPCA*
Consistent	15	18	15	16
Inconsistent	9	5	9	8
Percent consist.	62.5	79.2	62.5	66.7

Table 4. Agreement between all four methods of contour discrimination in numbers of speakers and percentages.

Agreement	*8 : 0*	*7 : 1*	*6 : 0*	*5 : 3*
No. of cases	11	7	5	1
Percent of cases	48.5	29.2	20.8	4.2

The results revealed complete agreement (8 : 0, i.e., 4 methods × 2 sentences per speaker) for eleven speakers, which is about a half of our sample. The agreement with one difference (7 : 1) was achieved for seven speakers, i.e., for almost one third of the sample. The results for five of the speakers produced two disagreements and one remaining speaker was evaluated with three inconsistencies.

4. Discussion

Fundamental frequency contours of spoken utterances must fulfil important communicative tasks, but they still leave some space for contextual and individual variation. Grammatically, the sentences in our experiment were imperatives and for those the Czech language requires falling contours. Other pragmatic aspects of the speaker's intention can add modifications of the prescribed contour. Our sample revealed two general shapes of which type *a* seemed to sound more conclusive and self-contained, while type *b* produced impressions of insinuations or other contextually based implications. Furthermore, each speaker contributed to the resulting melody with his/her individual habits and indexical markers. In combination with different articulation rates and phonation settings, the two contours generate a variety of phonopragmatic effects.

The three computational methods of contour quantification were able to capture both the general shapes and reflect on the speaker specific properties of the items. K-means clustering reduced the whole contour into mere seven values. Given that we took F0 measurements for the raw data every 15 ms, this means the compression in the ratio of about 1 : 11. One might argue that this drastic data reduction will eliminate important speaker specific detail. A counter argument would claim that forensic practitioners often reduce

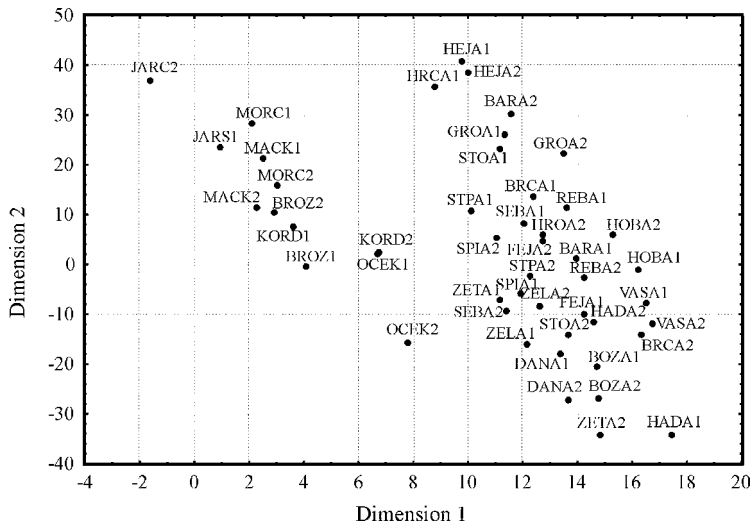

Figure 6. Mutual positions of the speakers' contours in the 2D space where Dimension 1 is the intercept and Dimension 2 the first coefficient of a cubic polynomial function.

the information about F0 into one value only – some central or baseline tendency. Our results offered another interesting fact. Although the production of the two sentences by each speaker was separated by substantial amount of other talking and the contexts for each of the two sentences differed, the cluster to which a speaker was assigned was the same for two thirds of our sample. This suggests that people have some relatively stable habits of saying certain things. In our opinion it is worth looking for ways of describing such habits.

The coefficients of polynomial functions revealed some speaker specific tendencies as well. Figure 6 shows that even mere intercept (representing the pitch of the beginning of the utterance) and the first coefficient place both utterances by the same speaker quite close to each other in a two-dimensional space.

Functional principal component analysis is the most modern of the methods chosen. At this stage of our research, its outcome is comparable with PRA results. However, one of the promising features of FPCA is that it does not iron out the details of the F0 contour in such opaque manner as the PRA does. Further experimenting with FPCA and smaller parts of the contours could perhaps help capture some truly speaker-specific production features.

Obviously, the methods have to be tested extensively on other types of utterances to further clarify the intonational grammar of Czech and its manifestations in a variety of contexts. Only then the reliable separation and definition of speaker-specific attributes will be possible.

ACKNOWLEDGEMENTS

This research was supported by the project GAČR 406/12/0298 and by the Programme of Scientific Areas Development at Charles University in Prague (PRVOUK), subsection 10 – Linguistics: Social Group Variation.

REFERENCES

Adami, A. G. (2007). Modelling prosodic differences for speaker recognition. Speech Communication, 49, pp. 277–291.

Boersma, P. & Weenink, D. (2010). Praat: doing phonetics by computer [Computer program]. Version 5.1.31, retrieved on April 10, 2010 from <http://www.praat.org>.

Cruttenden, A. (1997). Intonation. 2nd edition. Cambridge: Cambridge University Press.

Fant, G., Kruckenberg, A. & Nord, L. (1991). Prosodic and segmental speaker variations. Speech Communication, 10, pp. 521–531.

Grabe, E., Kochanski, G. & Coleman, J. (2007). Connecting intonation labels to mathematical descriptions of fundamental frequency. Language and Speech, 50/3, pp. 281–310.

Gussenhoven, C. (2004). The phonology of tone and intonation. Cambridge: Cambridge University Press.

Hermes, D. J. (2006). Stylization of pitch contours. In: S. Sudhoff et al. (Eds.), Methods in empirical prosody research. Berlin: Walter de Gruyter, pp. 19–61.

Hollien, H. (2002). Forensic Voice Identification. London: Academic Press.

Ladd, D. R. (1996). Intonational Phonology. Cambridge: Cambridge University Press.

Statsoft (2005). STATISTICA 7.1. Praha: StatSoft CR (Podbabská 16, Praha 6).

Volín, J. (2008). Variabilita neukončujících melodií ve světle shlukové analýzy. AUC Philologica 2/2007, Phonetica Pragensia XI, pp. 173–179.
Zellers, M., Gubian, M. & Post, B. (2011). Redescribing intonational categories with functional data analysis. In: Proceedings of Interspeech 2010. Makuhari: ISCA, pp. 1141–1144.

OBECNÉ A INDIVIDUÁLNÍ VLASTNOSTI KONTUR F0 V KRÁTKÝCH PROMLUVÁCH

Resumé

Studie porovnává tři zásadní metody kvantitativní analýzy křivek nebo konturových průběhů. Cílem je zjistit jejich užitečnost pro výzkum intonace. Materiálem byly promluvy 24 mluvčích extrahované z krátkých dialogů. Tyto promluvy byly obsahově shodné, sedm až devět slabik dlouhé, a každý mluvčí vyprodukoval dané sdělení dvakrát: jednou v singuláru, podruhé v plurálu. Obě verze byly při nahrávání odděleny řadou jiných dialogů, jedna tedy není mechanicky vyslovenou kopií druhé. Analýzy zjišťovaly, zda jsou vybrané metody schopny zachytit prozodické vzorce dané intonační gramatikou češtiny a zároveň prozradit něco o individuálních produkčních preferencích mluvčího. Zkoumanými metodami bylo: shlukování pomocí k-průměrů, nelineární regresní analýza a funkční analýza hlavních komponent. Numerický výstup metod byl porovnán s lidskou percepcí v auditivní i vizuální doméně. Výsledky ukázaly, že každá ze tří metod funguje poměrně dobře: výstupy jsou srozumitelné a lingvisticky interpretovatelné. Zdá se, že co do náhodných prvků je nejodolnější shlukování pomocí k-průměrů. Možnosti jednotlivých metod je však nutno dále prověřovat za různých podmínek a v kontextu různých hypotéz.