9th International Conference on Speech Prosody 2018
13-16 June 2018, Poznań, Poland

VJ.PEAT: Automated measurement of prosodic features

Tillmann Pistor, Carsten Keil

Research Center Deutscher Sprachatlas
Philipps-University of Marburg, Germany
tillmann.pistor@staff.uni-marburg.de, carsten.keil2@web.de

Abstract

This paper outlines the first steps of an innovative method of phonetic measurement of prosodic features, focusing on F_0-slopes in local intonation patterns called *PEAT*. The technique presented is an algorithm aiming to measure phonetic differences in speech signals by applying machine-learning techniques. The process, operating on the basis of speech analysis and statistical computation programs such as Praat and R, successively uses robust acoustic variables processing, a smoothing process based on the physiology of natural articulation and extraction of compliant paths according to a generic cost function. This process allows a fully automated determination of the calibration parameters when conducting phonetic measurements with Praat, thus eliminating subjectivity and making the results and illustrations reliable and comparable. Furthermore, PEAT can be used to automatically detect, measure and classify prosodic units in unknown speech signals, thus bridging phonetics and machine-learning.

Index Terms: methods, measurement of prosodic parameters, phonetics of prosody, algorithm, machine-learning

1. Introduction

In order to be able to adequately examine the fundamental frequency (pitch) and intensity contours represented by an algorithm (such as the one implemented in Praat), it is not sufficient to set the cursor in an arbitrary position of the time course in the signal and to read the determined values. These values can vary, depending on the selected time interval (window size) and the settings of the upper and lower limits (floor and ceiling) of the particular measured value range (F_0 or dB), which need to be initially determined by the user. In order to obtain valid measurements, the analyst must at least have a knowledge of how the visual and measurable progression of a certain acoustic quantity is modeled by the algorithm. Because this is what most pitch trackers perform – a computer-generated, digitized modeling of natural, physical events. Digitized modeling of language signals is created with most algorithms in such a way that a point is picked at regular intervals (frames) at a certain point in the window's time course. The point's value is then determined in relation to all other measuring points in the limited window. This particular point is then connected to the other points in the same frame (cf. [3], [4], [10], [13], and [14]). A common method, suitable for intonation research, which operates in this way, is called *autocorrelation*. In this method, single measuring points are determined by means of multiple section computations of the windows (here: Gauss windows) and their measured values of (in this case) the fundamental frequency F_0 are determined. A series of relational measurement points is then formed and the points are subsequently connected to one another by interpolarization. In this way, an algorithm can represent a dynamic course of F_0, depending on the selected section.

Reading and interpreting such computer- or algorithm-based modellings of the acoustic correlates of prosodic structures can however present several challenges. Subjectivity of the results of measurements is one of these challenges that needs to be dealt with: adjusting measurement parameters for pitch or intensity manually may lead to differing results depending on the researcher's goals, biological or articulatory dispositions of the speaker or the communicative circumstances of the investigated corpus. It should also be taken into account that microprosodic deviations can influence various parameters, but do not necessarily have to. Not every microprosodic deviation contributes to the constitution of the perceptually relevant prosodic structure. Even the opposite can be the case: the interruptions of the F_0-course of for instance partially desonorized consonants represented by the algorithm in pitch trackers consistently result with no measureable values for F_0 in the corresponding intervals. Human hearing, however, perceives speech sounds selectively and categorically, depending on the expectation and, above all, the individual phoneme system of the listener (cf. [17]). Regarding intonation, listeners with German as their mother tongue only perceive holistic and smooth contours (cf. [16]). Possible interruptions occurring in the abovementioned phenomena are consequently "blanked out" (if perceptually irrelevant). Therefore, human perception of prosody requires smoothing of so-called octave jumps (cf. [8]) and closing of possible gaps in the digitized representation of prosodic features, since all of these jumps and most gaps simply constitute artifacts of measurement and distort results and their interpretations.

How can the parameters be adjusted (normalized) with regard to the speaker individually but plausibly, without the need for manual re-adjustments afterwards? What representation of intensity and F_0-courses reflects the approximation of what human perception perceives most accurately? These questions are addressed using an innovative automated measurement and presentation method of prosodic features: PEAT. This paper focuses on the method and outlines the first steps of an application of this tool for automated measurement and classification of local intonation patterns, usually not extending over more than one syllable (cf. [12], [19], and [21]).

2. Process

The code name of the tool presented is in full *VokalJäger* (VJ) *2.0, prosody enhanced algorithmic toolbox* (PEAT). This process is a continuation and extension of the original VokalJäger, designed and programmed by Carsten Keil (cf. [13], and [14]).

10.21437/SpeechProsody.2018-115

The VokalJäger (literally in German: *hunter of vowels*) constitutes an algorithm to measure phonetic differences in speech signals applying machine-learning techniques. Its calculation kernel is implemented in the statistical R programming language (cf. [20]). Its original purpose was to provide an automatic method for the phonetic analyses of formants F_1 – F_3, the main acoustic correlates of vowels. Here, the scope of the algorithm was expanded to enable it to work with F_0 while still considering all other acoustic correlates of relevant prosodic features such as intensity and duration.

The process successively uses robust acoustic variables processing (sweeping), a smoothing process based on the physiology of natural articulation (DCT4) and extraction of compliant paths according to a generic cost function (best fit). These components of the process will be outlined in the following paragraphs. In subsequent steps, PEAT furthermore can be trained to detect the prevalence of so-called *binary features*, thus bridging phonetics and machine-learning.

2.1. Sweeping: F_0-floor and -ceiling

The VokalJäger algorithm significantly builds on preceding phonetic measurements performed by specialized software – most notably by the de-facto standard tool in phonetics: Praat (cf. [5], and [14]). However, such software may produce entirely different measurements depending on how certain calibration parameters have been set, which constitutes a significant challenge (cf. [7]). Optimizing those parameters thus is of high importance (cf. [23], and [24]), but usually requires experience how to tune them and what results are expected. The process of calibrating those parameters in practice can be tedious, time consuming and can, depending on the researcher's expectations, result in subjectivity. The VokalJäger employs a post-processing approach: It asks the phonetic software simply to perform measurements while sweeping numerous different parameter settings. The algorithm then picks the parameter setting, which produces the most appropriate measurement (see 2.2. and 2.3.). That allows for the building of a fully automated process and eliminates manual and arbitrary intervention when using Praat (cf. [13]).

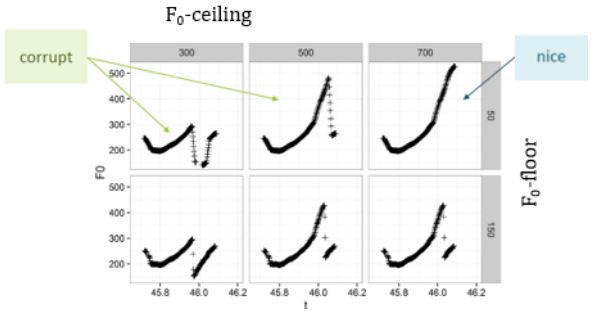

Figure 1: *Sweeping of F_0-floor and -ceiling in the measurement of a short, rising intonation pattern*

Figure 1 shows an exemplary sweeping of the measurement parameters F_0-floor and ceiling in the analysis of a short, locally rising intonation pattern. A female speaker realized this signal in a relatively high register of pitch, describing a visibly high pitch span. The x-axis (t) depicts the temporal extent of the speech signal in milliseconds. The grey columns on the right side show the settings for the F_0-floor (minimum), while the grey columns on top of the figure show the settings for the ceiling (maximum) of F_0 in the unit Hertz. The first setting (50 Hz min, 300 Hz max) is the default pitch setting in Praat. This setting obviously yields corrupt data, featuring a gap and a jump in the F_0-course, forcing the analyst to re-adjust the settings for every investigated signal or speaker. PEAT automatically calculates the optimal settings (here the odd case of 50 Hz min and 700 Hz max), as illustrated in the right upper window in figure 1.

2.2. Smoothing: DCT4

By means of curve smoothing operations, measurement artifacts derived from non-relevant microprosodic deviations (see above) can be reduced or even eliminated (cf. [9]). The following paragraph gives a short introduction to the necessity of smoothing operations. The technical implementation is then outlined below.

2.2.1. Towards the necessity of F_0-curve smoothing

A desideratum of prosodic-phonetic presentation and measurement principles concerns the modeling of intonation patterns as continuous, smooth contours without jumps and gaps. These requirements result from the physiological correlates of intonation contours, which can be found in human articulation. The following considerations are based on two premises: a) global intonation patterns are passive physiological results of decreasing subglottal pressure and b) local intonation patterns are actively controlled by muscular activity of the musculus cricothyroideus and the musculus vocalis (cf. [2], [6], and [18]). Premise a) explains the uncontroversial assumption of a continuously decreasing F_0-course over an utterance, to which the terms declination/downtrend could be applied. Premise b) is based on the results of electromyographic studies (cf. [6], and [15]): the activity/tension or passivity/relaxation of the musculus cricothyroideus is significantly responsible for short-term productions of fundamental frequency values in the human voice. Physiological processes such as muscle contractions always proceed as smooth and continuous, but never jump and rarely gap. Based on these physiological assumptions, the smoothing process of the F_0-courses described here operates via the *Discrete Cosine Transformation (DCT)*. A F_0-course smoothed by this method can only take the form of straight lines, arc and half-arc forms. The smoothing process by means of a DCT, in addition to smoothing itself, also consists of an assumption of basic (allowed) shapes and thus requires "training". Before this is discussed, the technical implementation is first outlined.

2.2.2. Technical implementation of DCT4

Within a DCT (see in detail [1]), a complicated function or data series is merged with another simple function. In this case, this concerns the sequence of fundamental frequency values F_0[t] and the cosine pattern. The curve (graph) of a cosine function is 2π-periodic and yields values from -1 to 1. The course therefore describes exactly those arc and half-arc shapes that need to be used to model the intonation of speech signals (see above). The sequence of a DCT is composed of cosine terms of increasing order (cf. [13]). Mathematically, the aim is to develop one function according to a set of other complete functions. In the original version of VJ, a DCT of order 3 was used to measure and optimize formant-courses. The formula of this basic DCT looks as follows:

$$F^3[\mathrm{t}] = \mu_F + 2G[2]\cos\left(\frac{\pi}{T}\left(\mathrm{t}-\frac{1}{2}\right)\right) + 2G[3]\cos\left(\frac{2\pi}{T}\left(\mathrm{t}-\frac{1}{2}\right)\right) \quad (1)$$

The DCT parameters G[k] can be considered as "weights" or factors, annotated to the complete functions. In the transformation, the parameters G[k] determine specific shapes or timing of shapes (i.e. moving on the x-axis). At first, three parameters influence the transformation of the original curve and thus the modeling of the smoothed curve of the F_0-slope. The first parameter with k=1 of a DCT is a constant and corresponds to the mean value μ in an interval (cf. [13]).

$$F^1[t] = G[1] = \mu_F \quad (2)$$

In this case, G[1] indicates the mean value of F_0 in the measured signal in Hertz and can thus be understood as a correlate of the pitch register. The second parameter with k=2 can represent a movement from one extremum (here: F_0-maximum) to the other (here: F_0-minimum). Accordingly, G[2] can be interpreted as a parameter which models a rise (with a preceding minus sign) or a fall (with a preceding plus sign) of the F_0-course. The third parameter with k=3 defines the curvature of the F_0-course and can additionally describe the return to the extremum's starting point. On the one hand, G[3] can be used to determine whether a rising or falling intonation pattern describes a convex or concave shape (regarding the specifics of convex or concave rising or falling intonation patterns see [11], and [19]). On the other hand, G[3] is used to model F_0-peak and -valley contours. Here, with a preceding minus sign the DCT models a peak contour. Similarly, a valley contour is modeled when the factor is preceded by a plus sign. The higher the value of G[3], the more distinct the falling-rising contour shows in a valley and vice versa in a peak.

A third order DCT is the default setting in the VokalJäger processing and classification of vowels. With these three parameters, one can simulate all basic articulatory movements and thus F_0-courses as well. For the analysis of more complex intonation patterns in PEAT, a fourth order DCT was applied. This means that a fourth parameter with k=4 was added to the three parameters already described above. This aimed for a finer transformation of the original curves. Adding the parameter G[4] on the one hand controls the occurrence of a further peak or valley (in addition to a possibly already present peak or valley) and, on the other hand, is to be regarded as a timing parameter. G[4] in overlay with G[3] thus shifts the peaks and valleys originally formed by G[3] on the x-axis of the signal's total temporal extent, starting from its center. Hence, G[4] allows a more accurate separation of simple rising or falling intonation patterns to other prototypes of, for example, a complex, falling-rising-falling form (cf. for instance the emotional prosodic unit called "positive evaluation" in [12], [19], and [21]).

2.3. Best fit: Identifying prototypical forms

In the process of picking the most appropriate curve, different measurements based on varying calibration parameters are considered. Selecting the best fit thus must be considered as a result of the sweeping and smoothing steps described above. In general, the selection process is carried out by assuming normatively that the paths essentially follow simple patterns (straight line, rising, falling or a combination of the three) which can be described by means of the DCT parameters G[1]–[3] and in addition for complex structures by G[4]. Thus, it is the goal to find the DCT parameters which represent the measured values most adequately. The step of picking the best fit furthermore compensates for a side effect of the two foregoing steps in the process, which might lead to inappropriate results of measurement and illustration: the transformation provides compliant results even if the F_0-curve depicted does not correspond to the actually produced signal (see first column in fig. 4) or even any naturally occurring signal. This is also a major issue of smoothing processes in other approaches. The main task in the picking of the best fit is hence an extraction of compliant paths according to a generic *cost function* (cf. [13], and [14]), including three requests, which need to be formulated as follows: 1) parameterize the curve with DCT parameters, then approximate it to DCT4, 2) close internal gaps by interpolarization, and 3) pick the curve with the least pass error.

The first two requests are described above. The pass error is defined here as the difference between the original F_0-curve and the expectation of the DCT smoothing process and its approximation. In other words: the relative deviation of a single approximated path from the original path (cf. [13]). The pass error is displayed and logarithmized in the (here) pseudo-unit decibel dB. Figure 4 shows a combination of sweeping (see fig. 1), smoothing and the picking of the best fit candidate (right upper window). In the figure, the value of the pass error in the windows is depicted in red above the F_0-curves.

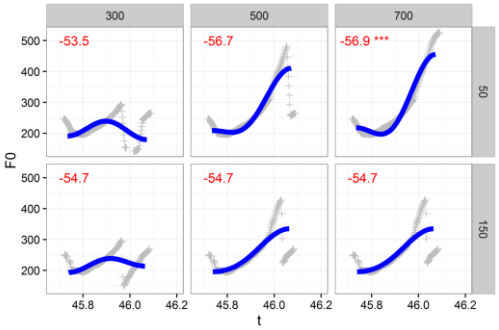

Figure 4: *Picking of the most appropriate F_0-course (right upper window) as best fit after sweeping and smoothing*

To provide the most appropriate curve, it is necessary to define what is appropriate. Concerning the structure of the F_0-curves, *appropriate* here is already defined above by criteria of natural articulatory processes (smooth and uninterrupted, see 2.2.1.). To avoid measurement artifacts and the acceptance of nonsensical values, the algorithm furthermore needs to be "trained" towards acceptable or most common values (subsequently represented by the DCT parameters and the signal's duration). This requires speech data from various speakers (male / female) in various communicative situations.

From the measurement and repeated cross validation of a multitude of local intonation patterns corresponding to the target patterns, a specific range is defined within which range most values are to be assumed. This strategy of expectations has proven itself effective in dealing with formant readings (cf. [22]) and can also be applied to F_0-courses. The expectation concerns both pitch range (F_0-minima and -maxima), and the characteristic, prototypical F_0-course of a particular pattern (rising, falling, constant, etc.) including the duration of the speech signal. F_0-curves with values within the physically reasonable range are favored over unreasonable readings outside this range. Outlier values are then *folded back* (cf. [13], and [14]) into the empirically predetermined range and prototypical courses of F_0 in individual intonation patterns. This kind of expectation of the measurement results offers an automated control and assures the quality of the measurements and the reliability within the algorithm.

3. Classification

After measuring and extracting the acoustic variables, they are used to classify patterns (by feature values) in speech segments. It is hereby assumed that each intonational unit in a specific interval shows a highly characteristic pattern. The classification tools employed in VJ originate from the field of machine-learning. Machine-learning is hereby used to automatically classify the prosodic units which PEAT has been trained on. Once trained, PEAT can automatically calculate the probability of the binary feature (cf. [13]). The binary feature must be regarded as certain phonetic criteria (i.e. the timely variation of variables over a speech interval) and is either present in an unknown signal or not. This allows for the statistical testing of whether or not two different groups of speech segments (usually phonologically assembled at different points in real-time) separate significantly concerning a certain floating phonetic feature (cf. [13]) such as F_0 ± rising or ± falling, represented by DCT-coefficients.

Therefore, the variables firstly need to be named, robustly measured, and cleaned respectively standardized, as described above. Secondly, a pattern detection algorithm needs to be trained and tested to classify intonation patterns based on the variables as input. The first and most important phonetic variable to consider is the fundamental frequency trajectory over time within the speech interval: $F_0[t]$. However, the trajectory itself is assumed to contain a surplus of information. All information is assumed to exist in the very elementary movements of the trajectory pattern, describing the basic nature of the curve: is it flat, does it rise or fall and/or does it reach a peak or a valley? These basics can be modeled by annotating weights to cosines of the DCT, as described in 2.2.2. The weights can then be evaluated with the DCT. They are accordingly called *DCT-coefficients* and enable modeling, grouping and thus comparing of different intonation patterns. This is exemplified in figure 5, which shows F_0-contours of four different local intonation patterns (cf. [12], [19], and [21]).

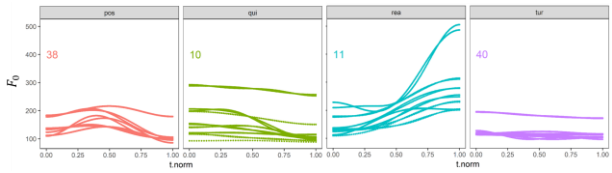

Figure 5: *Four different local intonation patterns holding the functions of "positive evaluation", "responding/closure", "reaction", and "turn holding" from left to right, as modeled by means of DCT-coefficients*

In figure 5, time (t.norm) is normalized within each interval. The value above the patterns is the mean of the interval length. The patterns significantly differ in F_0-slope and duration. Hence one can hypothesize, that if the combination of the features F_0-slope plus duration provided enough information, one might separate the patterns from each other and thus classify them.

As in the original VJ implementation, so-called binary classifiers have been employed: a classifier is trained to accept a single intonation pattern and reject the others. Here, it is assumed that a particular intonation pattern is prevalent, if the associated binary classifier shows the highest rate of support for the patterns (i.e. the probability of the pattern being prevalent is maximal). In the approach to train the binary classifier, a *median absolute deviation (MDA)* of possible orders from 1 to 20; 50%-train/50%-test repeated cross validation with 5 repeats and Cohen's kappa (cf. [26]) as optimization target was applied. The speech material stems from three different corpora containing laboratory (cf. [19]) and spontaneous (cf. [12]) speech. Until now, a total of 525 samples was used for the training sequences. The patterns from fig. 5 were trained and tested in these data among and against each other, proving themselves separable by the DCT-coefficients, as depicted in fig. 6.

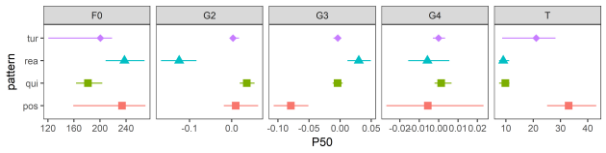

Figure 6: *Comparison of DCT-coefficients as variables in four different intonation patterns*

Clearly, G[2] singles out "rea", indicating a strong rise. G[3] singles out "pos" indicating a strong peak. The interval length T delivers a clue on "tur". It occurs that less significant information (i.e. too much overlap) is in the actual F_0 mean level, represented by G[1], and in the fourth parameter G[4]. Note that G[2], G[3] and G[4] shown are the DCT parameters normalized by average F_0, (G[1]). Hence, a classifier investigating the combinations of time and G[2] "T2"; plus G[3] "T23"; plus G[1] "T123" and finally plus G[4] "T1234" can identify and separate specific intonation patterns by means of their individual prosodic structure. When any prototypical (phonetic or phonological) and thus frequent structure is once determined (e.g. larger intonation phrases in other approaches, cf. [25]) in this way by means of the DCT-coefficients, PEAT can be trained on its measurement, classification and recognition.

4. Conclusion, discussion, and outlook

In this paper, the first steps of a fruitful new method of phonetic measurement of prosodic parameters have been presented. Automated calibrations of parameter settings, an articulation-based smoothing process and the empirically founded determination of the most appropriate candidate of an F_0-course offer an objective but precise and reliable approach to measurements, extractions and illustrations of prosodic structures.

However, the classification as well as the picking of the best fit candidate highly depend on the corpus that is used for the training of the algorithm. The algorithm can only determine an appropriate candidate when it is first trained on the factors, which define appropriation. The training sequences take into account the fact that the researchers must know about the structure and extension of the unit they want to investigate, and are thus able to segment the speech flow accordingly. This again requires having a concept for a clear delimitation of the specific unit, which can be challenging, especially concerning units that extend over more than one or two syllables and form whole utterances (such as intonation phrases or utterance phrases). For now, PEAT requires manual segmentation for the process of sweeping (see 2.1.). An approach offering an automatic detection of prosodic structures and respectively intonation patterns in unknown and unprocessed speech signals including automatic segmentation is one of the long-term objectives in the further development of PEAT and the VokalJäger in general, as well as the extension on larger prosodic units on the utterance level.

5. References

[1] Ahmed, N./Natarajan, T./Rao, R. (1974): *Discrete Cosine Transform*. In: IEEE Transaction on Computers 1: 90–93.

[2] Atkinson, J. E. (1978): *Correlation analysis of the physiological factors controlling fundamental voice frequency*. (Journal of the Acoustical Society of America 63: 211–222).

[3] Boersma, P. (1993): *Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound.* (IFA Proceedings 17: 97–110).

[4] Boersma, P. (2013): *Acoustic analysis*. In: Podesva, R. J./Devyani S. [eds.] (2013): *Research Methods in Linguistics*. New York: Cambridge University Press: 375–397.

[5] Boersma, P./Weenink, D. (2017): *Praat. Doing phonetics by computer*. Version 6.0.17.

[6] Collier, R. (1975): *Physiological correlates of intonation patterns.* (Journal of the Acoustical Society of America 58: 249–255).

[7] Escudero, P./Boersma, P. (2009): *A cross-dialect acoustic description of vowels: Brazilian and European Portuguese.* (Journal of the Acoustical Society of America 126 (3): 1379–1393).

[8] Féry, C. (2017): *Intonation and Prosodic Structure*. Cambridge: Cambridge University Press (Key Topics in Phonology).

[9] Gilles, P. (2005): *Regionale Prosodie im Deutschen. Variabilität in der Intonation von Abschluss und Weiterweisung*. Berlin/New York: de Gruyter (Impulse & Tendenzen 6).

[10] Gussenhoven, C. (2004): *The Phonology of Tone and Intonation*. Cambridge: Cambridge University Press (Research Surveys in Linguistics).

[11] Kaiser, S./Baumann, S. (2013): *Satzmodus und die Diskurspartikel hm: Intonation und Interpretation*. Hamburg: Buske (Linguistische Berichte 236: 473–496).

[12] Kehrein, R. (2002): *Prosodie und Emotionen*. Tübingen: Niemeyer (Reihe Germanistische Linguistik 231).

[13] Keil, C. (2017): *Der VokalJäger. Eine phonetisch-algorithmische Methode zur Vokaluntersuchung. Exemplarisch angewendet auf historische Tondokumente der Frankfurter Stadtmundart*. Hildesheim: Olms (Deutsche Dialektgeographie 122).

[14] Keil, C. (2017a): *Der VokalJäger*. Deutsche Dialektgeographie, Vol. 122. Georg Olms Verlag, Hildesheim. *Enhanced Algorithmic Toolbox, vokaljaeger.org*.

[15] Leemann, A. (2012): *Swiss German Intonation Patterns*. Amsterdam/Philadelphia: John Benjamins Publishing (Studies in Language Variation 10).

[16] Mixdorff, H. (2012): *The application of the Fujisaki model in quantitative prosody research*. In: Niebuhr, O. [ed.] (2012): *Understanding prosody. The role of context, function and communication*. Berlin: de Gruyter: 55–57.

[17] Neppert, J. (1999): *Elemente einer akustischen Phonetik*. 4., vollständig neu bearbeitete Auflage. Hamburg: Buske.

[18] Pétursson, M./Neppert, J. (2002): *Elementarbuch der Phonetik*. 3. durchgesehene und bearbeitete Auflage. Hamburg: Buske.

[19] Pistor, T. (2016): *Prosodic universals in discourse particles*. Proceedings Speech Prosody 2016, 31. Mai – 03. Juni 2016, Boston University: 869–872.

[20] R Development Core Team (2015)**.** *R: A Language and Environment for Statistical Computing*. Software. R Foundation for Statistical Computing. Wien.

[21] Schmidt J. E. (2001): *Bausteine der Intonation?* In: Schmidt, J. E. [ed.] (2001): *Neue Wege der Intonationsforschung*. Hildesheim: Olms (Reihe Germanistische Linguistik 157–158: 9–32).

[22] Thomas, E. (2011): *Sociophonetics. An Introduction*. New York: Palgrave Macmillan.

[23] Evanini, K./Lai, C./Zechner, K. (2011): *The importance of optimal parameter setting for pitch extraction.* Proceedings of Meetings on Acoustics 2010 15. – 19. November 2010, Cancun, Mexico: 1–10. Acoustical Society of America.

[24] De Looze, C./ Rauzy, S. (2009): *Automatic detection and prediction of topic changes through automatic detection of register variations and pause duration.* Proceedings of Interspeech 2009, Brighton, England.

[25] Rosenberg, A. (2010): *AuToBi – A tool for automatic ToBi annotation*. Proceedings of Interspeech 2010, 26. – 30. September 2010, Makuhari, Chiba, Japan: 146–149.

[26] Cohen, J. (1960): *A coefficient of agreement for nominal scales.* (Educational and Psychological Measurement 20: 37–46).