Automatic Pitch Accent Annotation

Grażyna Demenko, Magdalena Oleśkowicz-Popiel

The Institute of Linguistics, Adam Mickiewicz University, Poznań
lin@amu.edu.pl, magda.jastrzebska@gmail.com

Abstract

Based on a non-expressive speech data corpus of a few hundreds of utterances from 80 speakers a description and annotation of pitch accent shape have been proposed. Phrases were transcribed automatically on segmental level and lexical accents were annotated in accordance with expert rules using self-developed software. Pitch shape annotation was based on F0 changes on accented and postaccented syllables and parameters related to the register defined on a frequency physical scale and F0 change range. The indicated accent structures were verified manually and then evaluated statistically. Statistics have shown that accent in Polish is most frequently realized through F0 level changes between accented and postaccented syllable.

Index Terms: automatic prosody annotation, intonation, pitch accent shape

1. Introduction

Basic demands made of contemporary prosodic transcription need to be clearly defined for implementation of suprasegmental information in speech technology: (a) prosodic representation should be objective and consistent, well-grounded theoretically, (b) transcription should especially consider categorization of F0 changes with reference to: position and type of F0 change, size and change rate, synchronization of pitch movement with syllables structure, (c) and transcription should be as much as possible automatic.

So far none of prosody transcriptions fully meets these demands, e.g.[1][2][3]. The problem of pitch accent shape has been addressed by many studies e.g. [4][5][6][7]. It has been noted that accent realization may indicate speaker's emotional state such as tension, stress or anger, e.g. [8][9][10]. From the point of view of intonation function analysis, modelling register width and register position on the scale of fundamental frequency in relation to the whole range of the speaker's pitch height is crucial.

In [11], the terms key and register are used to describe variation in pitch range. Key is defined as "the width of the pitch range over whole intonation groups". Key is used for discourse purposes: it signals cohesion between intonation groups and indicates "the beginning and end of a topic: high key indicates the beginning of a new topic and low key indicates the end of a topic". Register is considered as an "overall (upward and downward) shifting of the whole pitch range within which the speaker is speaking". The function of register is to signal the emotional state of a speaker, e.g. tension, stress or anger. High register may signal defense or social politeness ([11], p.130). In the literature on the subject vocal register is one of the least clear concepts, e.g.[9][12][13][14][16][17]. A commonly used technique for pitch range changes discrimination is normalization related to the distance between the lowest and the highest F0 value reached by a speaker. A different possibility for fundamental frequency height change normalization is its relative location (in relation to a preceding pitch). This approach has been partly adopted in intonation transcription system INTSINT, e.g.[18][19]. A hypothesis has been put forward, that apart from differentiating between distinctive levels, there may be a parameter describing range which indicates a shift on a scale of all the heights (or maybe some) of the distinguished levels (see [11]).

Analyzing various methods of pitch changes normalization, it should be noted, that an objective scaling of fundamental frequency can be ensured only by a physical scale of a limited range for speech, so that it allows for unambiguous indication of the individual range of changes for pitch height and location on the scale, e.g.[20][21]. The results of recent studies on pitch location within speaker's F0 range, suggests that listeners have expectations about F0 for average speakers of each sex. Absolute F0 is the most important information for deciding about both pitch level and speaker sex [22][23].

Different types of emphases have been noted in early studies such as [25], where 3 forms have been pointed out*:*

a) upward obstruction of F0 (rise/fall or jump/drop),

b) steep fall plus drop before next syllable (jump/ fall drop),

c) the secondarily emphasized syllable interrupts the more rapid fall of the surrounding of non-emphasized syllables.

It is known that pitch accent shapes may correlate with expressiveness and may convey different meanings (English prefers glides, German jumps [24]).

2. Accent modelling for Polish language

For the Polish language an accent structure model that consists of an accented (A) and postaccented (PA) syllable has been assumed. The type of F0 parameter change (no change, rising, falling, rising-falling) is determined on an accented and postaccented vowel. Measuring boundaries, determined by vowel boundaries, were extended by preceding or subsequent sonorants (irrespective of their number), not further however than a syllable or word boundary. The course of F0 parameter over plosive consonants, fricatives and affricates was not included.

Previous research on microprosody (pitch change over a consonant) for Polish has shown that in most cases voiced plosive consonants and fricatives are characterized by a fall of its pitch level with respect to a pitch level of the neighboring vowels. Average pitch fall on plosive consonants (*b, d, g, J*) equals 8,5% and on fricatives (*v, z, Z*) is about 6,5% [26].

It was assumed that accented syllable initiates intonation changes on subsequent syllable/syllables and determines the pitch type (static/dynamic) and pitch configuration (falling/rising/flat). Dynamic pitch type carriers are those vowels and sonorants on which the F0 change is greater than

 doi: 10.21437/SpeechProsody.2016-16

2,5 ST (rise for L, fall for H) and extradynamic accent (L!, H!) – those where changes are greater than 4,5 ST. Static pitch type carriers are those vowels and sonorants on which there is no F0 change or the change is less than 2.5 ST (marked as L_ if the pitch on the next vowel is higher or H_ if the pitch on the next vowel is lower).
A structure, where the course of F0 parameter on the accented syllable is lower than the pitch of the next syllable is assumed to be of a rising course, and a structure where F0 on the accented syllable is higher than the pitch of the next syllable is of a falling course. First letter represents the course of F0 on an accented vowel (or a vowel and a sonorant); the next letter stands for the course of F0 on the vowel directly following the accented vowel.
Rising pitch begins with letter L (for a dynamic pitch) or L_ (for a static pitch). From a low level, F0 parameter can only rise. Falling pitch begins with letter H (for a dynamic pitch) or H_ (for a static pitch). From a high level, F0 parameter can only fall.
If the F0 course on an accented and postaccented vowel is close to level (within the range of +/-2,5 ST), then such a pitch is marked as level F. In this case F0 changes before accented syllable and after postaccented syllable determine the accent. If the F0 course within the accented syllable can be approximated by a quadratic function (with an extreme), then such a pitch change is marked with L&H. If the pitch on the accented vowel initiates F0 rise/fall on a few subsequent syllables (steepness of F0 course does not change and can be approximated with a linear function +/-2,5 ST), then such a pitch is marked as ascending L*/ descending H*.
The following types of accent structures can be distinguished for Polish:
1) pitch height change at the beginning of the accented syllable: dynamic rising pitch: LH, LL_, LH_; dynamic falling pitch: HL,HH_,HL_;
2) pitch height change after the accented syllable: static-rising pitch: L_H, L_L, L_H_, static-falling pitch: H_L_,H_H;
3) level pitch: F;
4) falling-rising/rising-falling pitch: L &H/ H&L;
5) ascending/descending pitch: L*/H*.
For the formalization of this description the following parameters should also be enclosed: position within a phrase, range of pitch changes that occur within a structure, position on the frequency scale, which can be described in form of a chain of markers:

[P][Z][N] [A {PA}][S]

where:
[P] within phrase position marker: < first accent, - subsequent accents, > last accent, the only accent.

[Z] range. Refers to the whole structure (comprising accented and not accented syllable): A – range less than or equal to 3 ST, B – range within: 4 ST <= 6 ST, C – range within: 7 ST <= 9 ST, D – range within: 10 ST <= 12 ST, E – range greater than 12 ST. Therefore, the measuring range is obtained from calculating the difference between F_{min} and F_{max} of this fragment of the utterance that contains the analyzed accent structure. The distinction of the above mentioned ranges results from perceptual research, according to which, within an octave linguistically salient are 3-4 levels. Additionally, the dynamics of F0 parameter changes should be considered only on the accented syllable. A symbol: "!" – refers to the range of F0 parameter changes on the accented syllable, if these changes are greater than 4,5 ST. These accents are marked as extradynamic.

[N] accent position. A position of an accent on a frequency scale represented by number N expressed in ST (N = 3, 6, 9, 12, 15, 18, 21, 24, 27, 30, 33, 36). Semitones are related to F_{min}= 60 Hz.

[A{!} {PA}] where the first letter indicates the type of F0 course on the accented syllable (A), the second on the postaccented syllable (PA). Symbol "!" refers to the dynamics of changes on the accented syllable.

[S] steepness (DF/DT). Expressed in semitones per seconds indicates steepness of pitch height changes on the accented and postaccented syllable (from the pitch course approximation determined by its maximum and minimum).

3. Automatic accent annotation

3.1. The choice of units of measurement

For the study of intonation, pitch distances are more relevant than absolute pitch – we can recognize the same melody in different pitch ranges. For this reason it is often useful to measure pitch in semitones rather than in Hertz. One semitone roughly corresponds to a frequency difference of 6 percent. The listener is sensitive only to these pitch movements which are intended by the speaker [27].

3.2. Accent annotation procedure

Accent annotation in an utterance consists of a few stages: 1) determination of the F0 course of the utterance, 2) determination of measuring range, 3) F0 course measuring on the accented and postaccented syllable (if such exists), 4) determination of the accent type on the basis of the measurements from the previous step.

3.2.1. Determination of the F0 course

F0 course is determined using spectral comb method enriched with postprocessing, that smoothes the contour and eliminates erroneous measurement, especially those that result from frequency doubling or bisecting.

3.2.2. Determination of measuring range

An accent model consisting of an accented and postaccented syllable has been assumed. A special case is a one-syllable model e.g. when accent falls on the last syllable in an utterance or in a sequence of one-syllable words.

3.2.3. Measuring of F0 course

F0 course in both accented and postaccented syllable is approximated by a straight line and a parabola. Approximation parameters together with some related mean squared errors are stored for further stages of deduction. All the calculations are carried out on log F0 values, semitones (ST), where 1 ST= $12*\log_2 F_0$. First deduction step is a preliminary determination of accent shape within one syllable and measuring its parameters. For that, within boundaries of < left sonorant – vowel – right sonorant>, F0 course approximation by a parabola is carried out, and then a mean squared error (MSE1) is calculated. Also, an aggregated mean squared error (MSE2) of linear approximation is calculated within each of the phonemes. Within these boundaries minimum, maximum and

limit values of F0 are searched for. If MSE1<MSE2 and the extreme value calculated from parabola parameters stays in the 10% range margin of the accented vowel and arbitrarily set number of windows with reliable F0 values constitutes at least 75% of the vicinity length, then according to the shape of the parabola the real value of either minimum or maximum becomes an extreme value.

3.2.4. Determination of the accent type

The following cases have been considered:
1. The extreme lies within a vowel:
   - If F0 change is greater than the 2,5 ST on both sides of the extreme, then F0 course takes shape L&H or H&L.
   - If on one side of the extreme F0 change is greater than 2,5 ST and on the other side is smaller than the perception threshold, then F0 course takes shape L or H.
   - If F0 changes in the whole vicinity are below the 2,5 ST, then F0 course takes shape F.
2. An extreme occurs outside a vowel:
   - Depending on F0 changes in the vicinity of the vowel, F0 course takes L, H or F shape
3. In other cases the type of F0 course is not determined.

For the purpose of accent annotation software partly based on Pitchline (Fig. 1) has been designed [28].

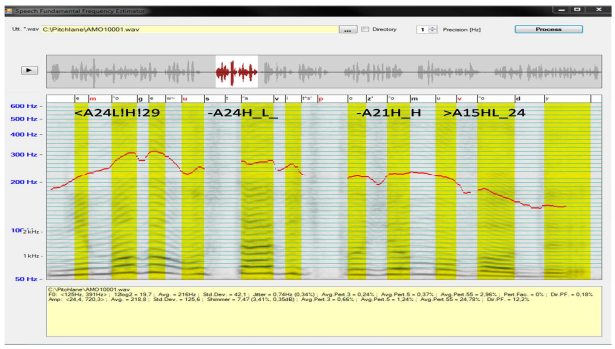

Figure 1: *A print screen of the software used in the study to generate description of accent structures. Accent structure descriptions, are shown on the spectrogram panel (4 accent structures: <A24L!H!29, -A24H_L_, -A21H_H, >A15HL_24).*

4. Evaluation

4.1. Examples of pitch accent shapes

The evaluation of the introduced annotation of pitch accent shape was run on the JURISDICT database [29]. 80 speakers were selected randomly. One part of the phrases was read and another part came from free speech (answers on a given topic).

Accent structures that are characterized by a rise on the accented syllable are shown at fig.2 (LH), fig.4 (LH_) and fig.6 (LL_). Fig.8 and fig.10 illustrate cases, where there is no pitch height change on the accented syllable and F0 change occurs after the accented syllable by a change after accented syllable.

Accent structures that are characterized by a fall on the accented syllable are shown at fig.3 (LH) and fig.5. Fig.7, fig.11 and fig.12 illustrate cases, where there is no pitch height change on the accented syllable and F0 change occurs after the accented syllable by a change after accented syllable. Fig.10 and fig.11 illustrate same pattern L_L (pattern on fig. 11 shifted up on the frequency scale), very common in Polish, intonation contour of the last fragment of a question. Fig.12 shows an example where on both an accented and postaccented vowel F0 course is close to level. Fig.13 illustrates structure L&H which is frequent for an emphasis.

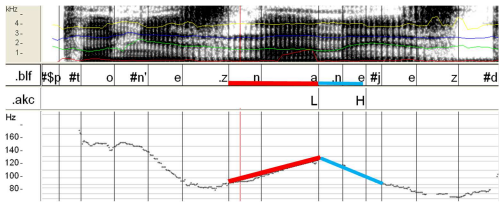

Figure 2: *Accent structure LH, on a word "unknown" (Polish "nieznane").*

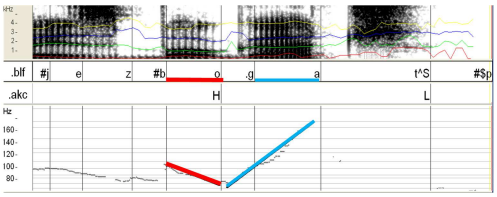

Figure 3: *Accent structure HL on a word "rich man" (Polish "bogacz").*

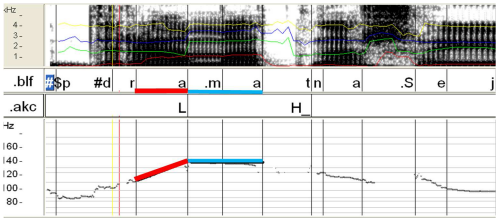

Figure 4: *Accent structure LH_ on a word "drama" (Polish "dramat").*

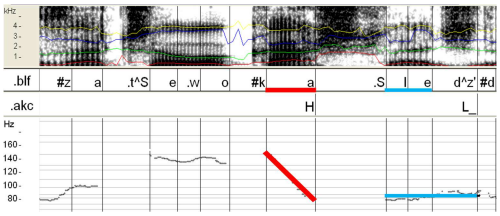

Figure 5: *Accent structure HL_ on a word "to cough" (Polish "kaszleć").*

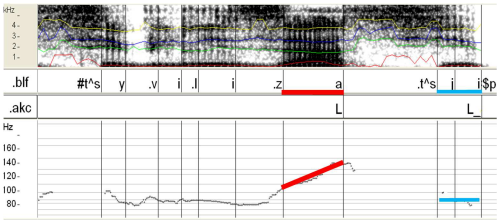

Figure 6: *Accent structure LL_ on a word "civilization" (Polish "dramat").*

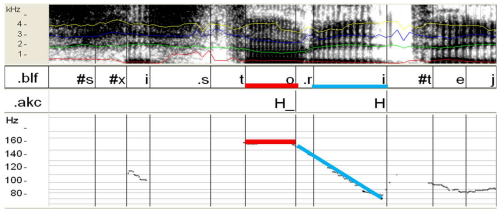

Figure 7: *Accent structure H_H on a word "history" (Polish "historii").*

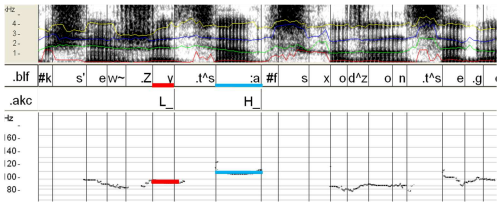

Figure 8: *Accent structure L_H_ on a word "moon" (Polish "księżyca").*

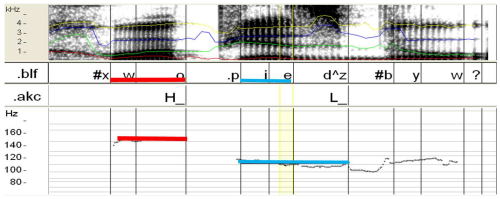

Figure 9: *Accent structure H_L_ on a word "boy" (Polish "chłopiec").*

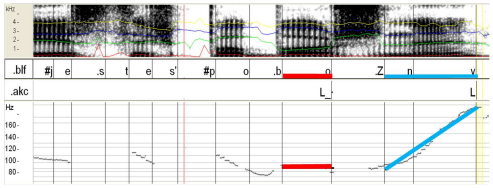

Figure 10: *Accent structure L_L on a word "devout" (Polish "pobożny").*

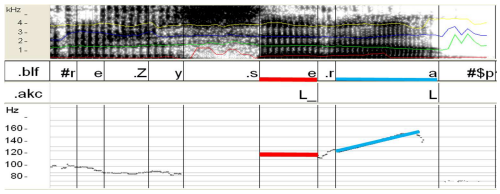

Figure 11: *Accent structure L_L on a word "(movie) director" (Polish "reżysera").*

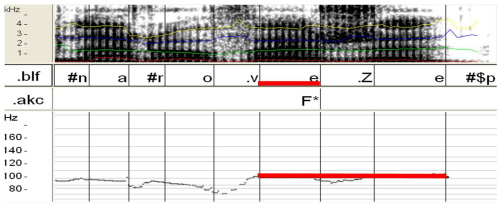

Figure 12: *Accent structure F on a word "bicycle" (Polish "rowerze").*

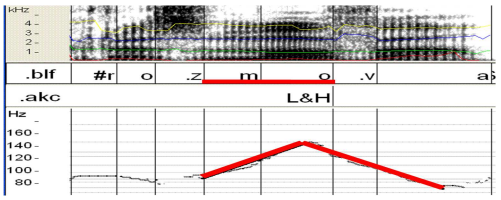

Figure 13: *Accent structure L&H on a word "conversation" (Polish "rozmowa").*

4.2. Statistical results

Annotation statistics have shown that in Polish structure F is the most frequently realized accent (fig.14). Common were also the accents for which F0 change occurred beyond the accented syllable (types: H_H: 879 and L_H_:304). Because the analysis concerned nonexpressive speech, extradynamic accents occurred rather rarely. Other parameters (N, S) were not evaluated statistically for the same reason.

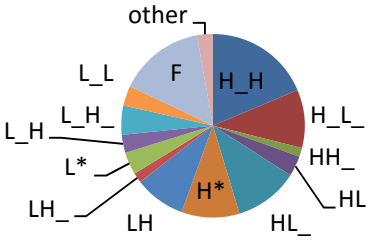

Figure 14: *Piechart of share of the most common accent structures in Polish.*

5. Discussion

Further development of the introduced method demands solving following problems: a) valuation of statistical features of accent for expressive speech, b) analysis of the influence of extralinguistic features e.g. vocal effort on parameters of modeled accent structures, c) investigation of the interaction between a type of change and phonetic factors e.g. microprosody.

Description of Polish intonation in terms of intonemes was given in [26] and provides a very useful framework for representation of Polish. However this system is not very useful for modelling because no interpretation of intonemes at the acoustic level is provided.

The first analysis and description of Polish intonation for the needs of speech technology applications was presented in [33]. Considering prosodic structure (following British school) an intonation tune consists of prenuclear and nuclear intonation. Perceptual and acoustic analyses of intonation contours resulted in classification of nuclear pitch accents into 9 groups and prenuclear pitch accents into 2 groups. The resulting representation of intonation does not account for possible variation of the alignment of the pitch such as segmental structure of syllables. The association of tonal targets with respect to segmental string has been investigated in a number of papers (e.g. [8] [30][31]) and belongs to the phonology of intonation. Apart from this, the alignment of pitch peaks (i.e. F0 maxima associated with ptich accents) contributes to the interpretation of the utterance meaning (e.g. [10]).

While the representation of intonation in [33] could be applied to intonation modelling, it would be reasonable to verify the inventory of pitch accents against speech material of another type. The first complete Polish intonation model for application in speech synthesis was presented in [34]. The inventory of pitch accent types [33] served as the starting point for the definition of intonation representation.

The method of stylization of intonation contours called PitchLine was presented for the first time in [28] and on this method the presented system of accent annotation has been based [32]. The main reason for developing our own stylization method, rather than using an existing one (e.g. Momel or Prosogram), was the need to establish a precise description of prosodic patterns for its potential application in extralinguistic information extraction.

6. Conclusion

The proposed description enabled statistical analysis of pitch accent shape in Polish. Accent in Polish can be induced by two different mechanisms, a jump to a new pitch level in the syllable nucleus, and a change within the syllable nucleus. The use of a jump rather than a glide or vice versa is often dependent on the make-up of the syllables over which they are spread. If it is only one syllable a glide is more likely to be used. If the pitch accent falls on a syllable with a short vowel particularly followed by voiceless consonant, and there is a following syllable, the movement from the accent is more likely to be realized as a jump. In general the use of a jump where a glide might be expected sounds abrupt, whereas the use the glide when a jump is expected sounds soothing or reproachful. Due to the universal aspects of prosody it may be assumed that the introduced method and software may prove useful for analysis of various extralinguistic intonation functions.

7. Acknowledgements

This research was supported by the National Science Center (Project ID: 2014/14/M/HS2/00631).

8. References

[1] Beckman, M. E., Hirschberg, J. 1994. The ToBI annotation conventions. Ohio State University.

[2] Hirst, D., Di Cristo, A. 1998. Intonation systems: a survey of twenty languages. Cambridge University Press.

[3] Patterson, D., Ladd, D. R. 1999. Pitch range modelling: linguistic dimensions of variation. Proc. of ICPhS, 1999, 1169-1172.

[4] Abramson, A. S. 1978. Static and dynamic acoustic cues in distinctive tones. Lang. Speech. 21(4), 319-325.

[5] Beckman, M. E. 1986. Stress and non-stress accent. Walter de Gruyter.

[6] Crystal, D. 1971. Relative and absolute in intonation analysis. J. Int. Phonetic Ass. 1(01), 17-28.

[7] Niebuhr, O., D'Imperio, M., Gili Fivela, B., Cangemi, F. 2011. Are there 'shapers' and 'aligners'? Individual differences in signalling pitch accent category. Proc. 17th ICPhS Hong-Kong, 120-123.

[8] Dilley, L. C., Ladd, D. R., Schepman, A. 2005. Alignment of L and H in bitonal pitch accents: testing two hypotheses. J Phonetics. 33(1), 115-119.

[9] Hirschberg, J., Pierrehumbert, J. 1986. The intonational structuring of discourse. Proceedings of the 24th annual meeting on Association for Computational Linguistics, 136-144.

[10] Kohler, K. 2005. Timing and communicative functions of pitch contours. PHONETICA-BASEL. 62(2-4), 88.

[11] Cruttenden, A. 1997. Intonation. Cambridge University Press.

[12] Hollien, H. 1972. On Vocal Registers. Journal of Phonetics, 2, 125-143.

[13] Ladd, D. R. 1994. Constraints on the gradient variability of pitch range (or) Pitch level 4 lives! Phonological Structure and Phonetic Form, 3, 43.

[14] Mertens, P. 2004. The prosogram: Semi-automatic transcription of prosody based on a tonal perception model. Proc. Speech Prosody 2004 Nara.

[15] Ohala, J. J., Ewan, W. G. 1973. Speed of pitch change. J. Acoust. Soc. Am. 53(1), 345-345.

[16] Shriberg, E., Ladd, D.R., Terken, J., Stolcke, A. 1996. Modeling pitch range variation within and across speakers: predicting F0 targets when "speaking up". Proc. 4th ICSLP Philadelphia,1-4.

[17] Švec, J. G., Schutte, H. K., Miller, D. G. 1999. On pitch jumps between chest and falsetto registers in voice: Data from living and excised human larynges. J. Acoust. Soc. Am. 106(3), 1523-1531.

[18] De Looze, C., Hirst, D. 2008. Detecting changes in key and range for the automatic modelling and coding of intonation. Proc. Speech Prosody 2008 Campinas.

[19] Hirst, D. 2007. A Praat plugin for Momel and INTSINT with improved algorithms for modelling and coding intonation. Proc. 16th ICPhS Saarbrucken, vol. 12331236.

[20] Hermes, D. J., Van Gestel, J. C. 1991. The frequency scale of speech intonation. J. Acoust. Soc. Am. 90(1), 97-102.

[21] Nolan, F. 2003. Intonational equivalence: an experimental evaluation of pitch scales. Proc. 15th ICPhS Barcelona, 774.

[22] Honorof, D. N., Whalen, D. 2005. Perception of pitch location within a speaker's F0 range. J. Acoust. Soc. Am. 117(4), 2193-2200

[23] Bishop, J., Keating, P. 2012. Perception of pitch location within a speaker's range: Fundamental frequency, voice quality and speaker sex. J. Acoust. Soc. Am. 132(2), 1100-1112.

[24] Cruttenden, A. (1986). Intonation Cambridge University Press. Cambridge, England.

[25] O'Shaughnessy, D., Allen, J. 1983. Linguistic modality effects on fundamental frequency in speech. J. Acoust. Soc. Am. 74(4), 1155-1171.

[26] Steffen-Batogowa, M. 2000. Struktura akcentowa języka polskiego. Wydawnictwo Naukowe PWN Warszawa.

[27] J. 't Hart, R. Collier & A. Cohen. 1990. A perceptual study of intonation: an experimental-phonetic approach to speech melody. Cambridge: Cambridge University Press.

[28] Demenko, G., Wagner, A. 2007. Prosody annotation for unit selection TTS synthesis. Arch. Acoust. 32(1), 25-40.

[29] Demenko, G., Grocholewski, S., Klessa, K., Ogorkiewicz, J., Wagner, A., Lange, M, Cylwik, N. 2008. JURISDIC – Polish Speech Database for taking dictation of legal texts, Proceedings of the Sixth International Conference on Language Resources and Evaluation, LREC, 1280–1287.

[30] Rietveld, T., & Kerkhoff, J. 2002. The temporal alignment of L* H Accents. In Speech Prosody 2002, International Conference.

[31] Atterer, M., & Ladd, D. R. 2004. On the phonetics and phonology of "segmental anchoring" of F0: evidence from German. Journal of Phonetics, 32(2), 177-197.

[32] Demenko, G. 2015. Korpusowe badania języka mówionego (Corpus based spoken language studies). Akademicka Oficyna Wydawnicza EXIT, seria: Lingwistyka komputerowa, Warszawa.

[33] G. Demenko. 1999. Analysis of Polish Suprasegmentals for Speech Technology, wyd. UAM, Poznań.

[34] Oliver, D., & Clark, R. A. 2005. Modelling Pitch Accent Types for Polish Speech Synthesis.